CSC424 Assignment 2

Jasmine Dumas

February 4, 2016

### Problem #1 (Regression analysis - 20 points) The Housing dataset (housing.data) contains housing values in the suburbs of Boston. The detailed explanation concerning the input and output variables can be fetched from the UCI machine learning repository <http://archive.ics.uci.edu/ml/datasets/Housing>:

# change working directory  
setwd("/Users/jasminedumas/Desktop/depaul/CSC424")  
# get data  
library(readr)  
boston\_housing <- read\_table("housing.data", col\_names = FALSE)  
# add column names  
colnames(boston\_housing) <- c("CRIM", "ZN", "INDUS", "CHAS", "NOX", "RM", "AGE",   
 "DIS", "RAD", "TAX", "PTRATIO", "B", "LSTAT", "MEDV")

1. Fit a linear regression model and report goodness of fit, the utility of the model, the estimated coefficients, their standard errors, and statistical significance. Use the default method for running regression analysis in SPSS and interpret your results.

# linear model  
fit <- lm(MEDV ~., data = boston\_housing) # the dot means all variables  
summary(fit)

##   
## Call:  
## lm(formula = MEDV ~ ., data = boston\_housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.3683 -2.8462 -0.5845 1.6742 28.5372   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 28.305111 5.308863 5.332 1.49e-07 \*\*\*  
## CRIM 0.209281 0.143284 1.461 0.144762   
## ZN 0.014940 0.013501 1.107 0.269005   
## INDUS 0.012716 0.061348 0.207 0.835874   
## CHAS 3.005654 0.893097 3.365 0.000824 \*\*\*  
## NOX -15.523485 4.044154 -3.838 0.000140 \*\*\*  
## RM 4.299560 0.429753 10.005 < 2e-16 \*\*\*  
## AGE 0.002848 0.013625 0.209 0.834480   
## DIS -1.083663 0.186785 -5.802 1.18e-08 \*\*\*  
## RAD 0.193259 0.162951 1.186 0.236196   
## TAX -0.002420 0.002746 -0.882 0.378458   
## PTRATIO -0.965535 0.134533 -7.177 2.64e-12 \*\*\*  
## B 0.009435 0.002757 3.422 0.000673 \*\*\*  
## LSTAT -0.525243 0.052280 -10.047 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.932 on 492 degrees of freedom  
## Multiple R-squared: 0.7198, Adjusted R-squared: 0.7124   
## F-statistic: 97.23 on 13 and 492 DF, p-value: < 2.2e-16

***The overall p-value is low (< 0.05), High F-statistic, Moderately high R-squared and adjusted R-squared value, many of the estimate coefficients are significant to the model. The model could be improved by elimiating the un-significant variables.***

1. Perform a feature selection on this data by using the forward selection method of the regression analysis. Analyze the output in terms of the order in which the variables are included in the regression model.

min.model = lm(MEDV ~ 1, data=boston\_housing) # aka intercept only model  
biggest = formula(lm(MEDV ~ ., data=boston\_housing))   
model = step(min.model, direction='forward', scope=biggest)

## Start: AIC=2246.51  
## MEDV ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + LSTAT 1 23243.9 19472 1851.0  
## + RM 1 20654.4 22062 1914.2  
## + PTRATIO 1 11014.3 31702 2097.6  
## + INDUS 1 9995.2 32721 2113.6  
## + TAX 1 9377.3 33339 2123.1  
## + NOX 1 7800.1 34916 2146.5  
## + AGE 1 6069.8 36647 2171.0  
## + CRIM 1 5600.1 37116 2177.4  
## + ZN 1 5549.7 37167 2178.1  
## + B 1 4749.9 37966 2188.9  
## + DIS 1 2984.5 39732 2211.9  
## + CHAS 1 1312.1 41404 2232.7  
## + RAD 1 550.5 42166 2241.9  
## <none> 42716 2246.5  
##   
## Step: AIC=1851.01  
## MEDV ~ LSTAT  
##   
## Df Sum of Sq RSS AIC  
## + RM 1 4033.1 15439 1735.6  
## + PTRATIO 1 2670.1 16802 1778.4  
## + CHAS 1 786.3 18686 1832.2  
## + DIS 1 440.5 19032 1841.4  
## + AGE 1 304.3 19168 1845.0  
## + TAX 1 274.4 19198 1845.8  
## + B 1 198.3 19274 1847.8  
## + RAD 1 165.1 19307 1848.7  
## + ZN 1 160.3 19312 1848.8  
## + INDUS 1 98.7 19374 1850.4  
## <none> 19472 1851.0  
## + NOX 1 4.8 19468 1852.9  
## + CRIM 1 2.4 19470 1853.0  
##   
## Step: AIC=1735.58  
## MEDV ~ LSTAT + RM  
##   
## Df Sum of Sq RSS AIC  
## + PTRATIO 1 1711.32 13728 1678.1  
## + CHAS 1 548.53 14891 1719.3  
## + B 1 512.31 14927 1720.5  
## + TAX 1 425.16 15014 1723.5  
## + DIS 1 217.03 15222 1730.4  
## + CRIM 1 129.58 15310 1733.3  
## + RAD 1 83.95 15355 1734.8  
## + INDUS 1 61.09 15378 1735.6  
## <none> 15439 1735.6  
## + ZN 1 56.56 15383 1735.7  
## + AGE 1 20.18 15419 1736.9  
## + NOX 1 14.90 15424 1737.1  
##   
## Step: AIC=1678.13  
## MEDV ~ LSTAT + RM + PTRATIO  
##   
## Df Sum of Sq RSS AIC  
## + DIS 1 411.67 13316 1664.7  
## + B 1 389.68 13338 1665.6  
## + CHAS 1 377.96 13350 1666.0  
## + AGE 1 66.24 13662 1677.7  
## <none> 13728 1678.1  
## + TAX 1 44.36 13684 1678.5  
## + RAD 1 29.98 13698 1679.0  
## + NOX 1 24.81 13703 1679.2  
## + ZN 1 14.96 13713 1679.6  
## + INDUS 1 0.83 13727 1680.1  
## + CRIM 1 0.57 13727 1680.1  
##   
## Step: AIC=1664.73  
## MEDV ~ LSTAT + RM + PTRATIO + DIS  
##   
## Df Sum of Sq RSS AIC  
## + NOX 1 608.74 12708 1643.0  
## + B 1 504.70 12812 1647.2  
## + CHAS 1 280.13 13036 1656.0  
## + TAX 1 221.85 13094 1658.2  
## + INDUS 1 152.54 13164 1660.9  
## + CRIM 1 56.21 13260 1664.6  
## <none> 13316 1664.7  
## + ZN 1 37.56 13279 1665.3  
## + AGE 1 24.67 13292 1665.8  
## + RAD 1 21.53 13295 1665.9  
##   
## Step: AIC=1643.05  
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX  
##   
## Df Sum of Sq RSS AIC  
## + B 1 342.24 12365 1631.2  
## + CHAS 1 341.43 12366 1631.3  
## <none> 12708 1643.0  
## + RAD 1 30.98 12677 1643.8  
## + TAX 1 17.47 12690 1644.3  
## + CRIM 1 11.70 12696 1644.6  
## + AGE 1 8.18 12699 1644.7  
## + ZN 1 6.02 12702 1644.8  
## + INDUS 1 0.72 12707 1645.0  
##   
## Step: AIC=1631.23  
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + B  
##   
## Df Sum of Sq RSS AIC  
## + CHAS 1 298.877 12066 1620.8  
## <none> 12365 1631.2  
## + CRIM 1 35.180 12330 1631.8  
## + RAD 1 25.089 12340 1632.2  
## + ZN 1 13.941 12351 1632.7  
## + AGE 1 0.867 12364 1633.2  
## + INDUS 1 0.063 12365 1633.2  
## + TAX 1 0.000 12365 1633.2  
##   
## Step: AIC=1620.85  
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + B + CHAS  
##   
## Df Sum of Sq RSS AIC  
## <none> 12066 1620.8  
## + CRIM 1 35.684 12031 1621.4  
## + ZN 1 19.035 12047 1622.0  
## + RAD 1 16.670 12050 1622.2  
## + TAX 1 0.866 12066 1622.8  
## + INDUS 1 0.704 12066 1622.8  
## + AGE 1 0.000 12066 1622.8

summary(model)

##   
## Call:  
## lm(formula = MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + B + CHAS,   
## data = boston\_housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.1619 -2.7574 -0.6029 1.6037 28.9387   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 27.842510 4.869270 5.718 1.86e-08 \*\*\*  
## LSTAT -0.515932 0.048103 -10.725 < 2e-16 \*\*\*  
## RM 4.463809 0.407610 10.951 < 2e-16 \*\*\*  
## PTRATIO -1.000843 0.113917 -8.786 < 2e-16 \*\*\*  
## DIS -1.035913 0.169491 -6.112 1.99e-09 \*\*\*  
## NOX -14.608996 3.206151 -4.557 6.55e-06 \*\*\*  
## B 0.009404 0.002674 3.517 0.000477 \*\*\*  
## CHAS 3.101675 0.883132 3.512 0.000485 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.922 on 498 degrees of freedom  
## Multiple R-squared: 0.7175, Adjusted R-squared: 0.7136   
## F-statistic: 180.7 on 7 and 498 DF, p-value: < 2.2e-16

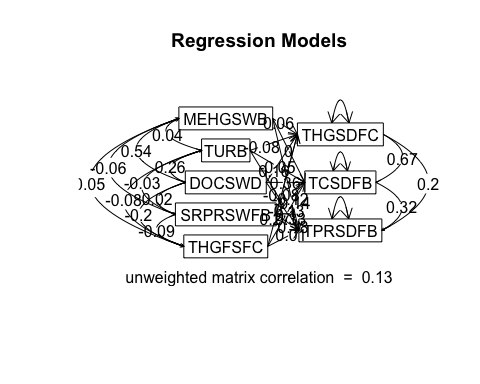
**With the forward selection, the output of terms are only included if they decrease the AIC score therefore when starting with MEDV ~ 1 at the intercept the covariate with the lowest AIC is added first into the model and the process continues until the row is at the top (lowest AIC compared to the other variables). The last step confirms that the inclusion of anymore variables beyond CHAS would not prove to be significant to be included in the model. In comparing the full model fit with the forward selection, the full model deemed the same significant variables that were included in the forward selection and also the estimated coefficients were similar in value for both the full model and forward model.**

### Problem #2 (Canonical Correlation Analysis – 20 points): Water, soil, and mosquito fish samples were collected at n = 165 sites/stations in the marshes of southern Florida.

Answer the following questions regarding the canonical correlations.

Additional Source: SPSS\_MANOVA\_output.doc

# get the data  
library(readr)  
# removed a row with label units and transformed to csv file format  
marsh <- read\_csv("data\_marsh\_cleaned\_hw2.csv")  
# create two groups for water & soil variables  
water <- marsh[, 2:6]  
soil <- marsh[, 7:9]  
# CCA summary  
library(psych)  
a = set.cor(x=2:6, y=7:9, data = marsh)



print(a) # print produces the t and p-values

## Call: setCor(y = y, x = x, data = data, z = z, n.obs = n.obs, use = use,   
## std = std, square = square, main = main)  
##   
## Multiple Regression from raw data   
##   
## Beta weights   
## THGSDFC TCSDFB TPRSDFB  
## MEHGSWB 0.06 0.00 -0.14  
## TURB -0.08 -0.05 -0.13  
## DOCSWD 0.16 0.36 0.33  
## SRPRSWFB -0.08 -0.12 0.18  
## THGFSFC 0.27 0.10 0.01  
##   
## Multiple R   
## THGSDFC TCSDFB TPRSDFB   
## 0.33 0.36 0.33   
## multiple R2   
## THGSDFC TCSDFB TPRSDFB   
## 0.11 0.13 0.11   
##   
## Unweighted multiple R   
## THGSDFC TCSDFB TPRSDFB   
## 0.25 0.27 0.21   
## Unweighted multiple R2   
## THGSDFC TCSDFB TPRSDFB   
## 0.06 0.07 0.04   
##   
## SE of Beta weights   
## THGSDFC TCSDFB TPRSDFB  
## MEHGSWB 13.01 2.45 37.61  
## TURB 0.47 0.09 1.35  
## DOCSWD 0.72 0.14 2.09  
## SRPRSWFB 730.75 137.52 2112.49  
## THGFSFC 0.08 0.01 0.22  
##   
## t of Beta Weights   
## THGSDFC TCSDFB TPRSDFB  
## MEHGSWB 0.00 0.00 0.00  
## TURB -0.16 -0.52 -0.10  
## DOCSWD 0.23 2.65 0.16  
## SRPRSWFB 0.00 0.00 0.00  
## THGFSFC 3.48 6.65 0.03  
##   
## Probability of t <   
## THGSDFC TCSDFB TPRSDFB  
## MEHGSWB 1.00000 1.0e+00 1.00  
## TURB 0.87000 6.0e-01 0.92  
## DOCSWD 0.82000 8.9e-03 0.87  
## SRPRSWFB 1.00000 1.0e+00 1.00  
## THGFSFC 0.00064 4.4e-10 0.98  
##   
## Shrunken R2   
## THGSDFC TCSDFB TPRSDFB   
## 0.081 0.104 0.084   
##   
## Standard Error of R2   
## THGSDFC TCSDFB TPRSDFB   
## 0.044 0.047 0.044   
##   
## F   
## THGSDFC TCSDFB TPRSDFB   
## 3.88 4.80 4.01   
##   
## Probability of F <   
## THGSDFC TCSDFB TPRSDFB   
## 0.002430 0.000412 0.001880   
##   
## degrees of freedom of regression   
## [1] 5 159  
##   
## Various estimates of between set correlations  
## Squared Canonical Correlations   
## [1] 0.149 0.119 0.072  
## Chisq of canonical correlations   
## [1] 26 20 12  
##   
## Average squared canonical correlation = 0.11  
## Cohen's Set Correlation R2 = 0.3  
## Shrunken Set Correlation R2 = 0.23  
## F and df of Cohen's Set Correlation 3.92 15 420.01  
## Unweighted correlation between the two sets = 0.13

1. Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value. ***The hypothesis test of at least one canonical correlation being not equal to zero is equivalent to testing whether the first canonical correlation is significantly different from zero.***

# the test statistic (Wilks)  
(1 - a$cancor2[1]) \* (1 - a$cancor2[2]) \* (1 - a$cancor2[3])

## [1] 0.6963021

# df (n - 1) where n = 165  
a$df[1] + a$df[2]

## [1] 164

print("p-value from SPSS, Sig. F = 0.000")

## [1] "p-value from SPSS, Sig. F = 0.000"

# ct = corr.test(water, soil)  
# ct$p

1. Test the null hypothesis that the second and third canonical correlations equal zero. Give your test statistic, d.f., and p-value. ***This question follows the first in assessing wether the 2nd or 3rd canonical correlations are significantly different from zero.***

# the test statistic   
(1 - a$cancor2[2]) \* (1 - a$cancor2[3])

## [1] 0.8179043

# df (n - 1) where n = 165  
a$df[1] + a$df[2]

## [1] 164

print("p-value from SPSS, Sig. F = 0.000")

## [1] "p-value from SPSS, Sig. F = 0.000"

1. Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic, d.f., and p-value. ***A very similar analysis as above***

# the test statistic   
(1 - a$cancor2[3])

## [1] 0.9284064

# df (n - 1) where n = 165  
a$df[1] + a$df[2]

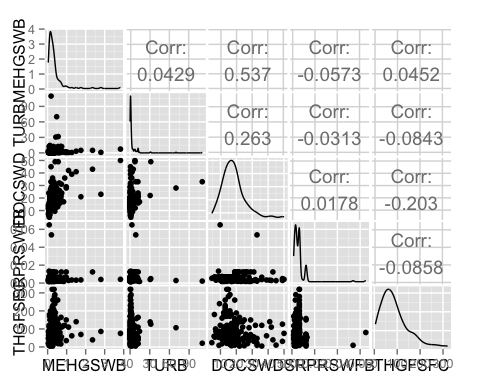
## [1] 164

print("p-value from SPSS, Sig. F = 0.000")

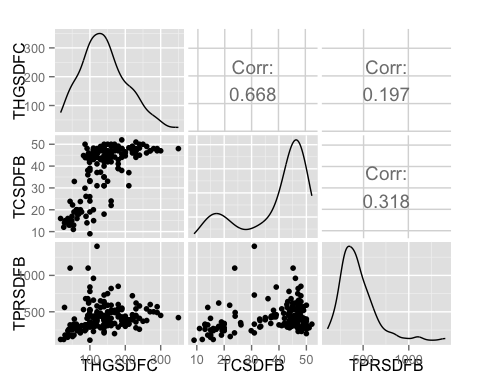
## [1] "p-value from SPSS, Sig. F = 0.000"

1. Present the three canonical correlations, together with their standard errors. (Report the standard errors only if you are using SAS; SPSS will not output the standard errors)

library(CCA)  
library(ggplot2)  
library(GGally)  
ggpairs(water)



ggpairs(soil)



# correlations between the two groups of variables  
matcor(water, soil)

## $Xcor  
## MEHGSWB TURB DOCSWD SRPRSWFB THGFSFC  
## MEHGSWB 1.00000000 0.04286195 0.53653344 -0.05729504 0.04523356  
## TURB 0.04286195 1.00000000 0.26262016 -0.03127880 -0.08426556  
## DOCSWD 0.53653344 0.26262016 1.00000000 0.01784706 -0.20284406  
## SRPRSWFB -0.05729504 -0.03127880 0.01784706 1.00000000 -0.08581679  
## THGFSFC 0.04523356 -0.08426556 -0.20284406 -0.08581679 1.00000000  
##   
## $Ycor  
## THGSDFC TCSDFB TPRSDFB  
## THGSDFC 1.0000000 0.6677804 0.1966074  
## TCSDFB 0.6677804 1.0000000 0.3178176  
## TPRSDFB 0.1966074 0.3178176 1.0000000  
##   
## $XYcor  
## MEHGSWB TURB DOCSWD SRPRSWFB THGFSFC  
## MEHGSWB 1.00000000 0.04286195 0.53653344 -0.05729504 0.04523356  
## TURB 0.04286195 1.00000000 0.26262016 -0.03127880 -0.08426556  
## DOCSWD 0.53653344 0.26262016 1.00000000 0.01784706 -0.20284406  
## SRPRSWFB -0.05729504 -0.03127880 0.01784706 1.00000000 -0.08581679  
## THGFSFC 0.04523356 -0.08426556 -0.20284406 -0.08581679 1.00000000  
## THGSDFC 0.15971021 -0.05151880 0.11909492 -0.09647552 0.25310209  
## TCSDFB 0.19749008 0.04374098 0.32344092 -0.11800127 0.03809560  
## TPRSDFB 0.02092839 -0.05980083 0.22121653 0.19411633 -0.07060351  
## THGSDFC TCSDFB TPRSDFB  
## MEHGSWB 0.15971021 0.19749008 0.02092839  
## TURB -0.05151880 0.04374098 -0.05980083  
## DOCSWD 0.11909492 0.32344092 0.22121653  
## SRPRSWFB -0.09647552 -0.11800127 0.19411633  
## THGFSFC 0.25310209 0.03809560 -0.07060351  
## THGSDFC 1.00000000 0.66778043 0.19660738  
## TCSDFB 0.66778043 1.00000000 0.31781764  
## TPRSDFB 0.19660738 0.31781764 1.00000000

# display the canonical correlations  
cc1 <- cc(water, soil)  
cc1$cor

## [1] 0.3855843 0.3449978 0.2675698

# raw canonical coefficients  
cc1[3:4]

## $xcoef  
## [,1] [,2] [,3]  
## MEHGSWB 0.720571333 -0.613310304 0.442819677  
## TURB 0.014902006 0.003947628 0.046585662  
## DOCSWD -0.122898091 -0.045649299 -0.038307498  
## SRPRSWFB -15.972715690 77.864165952 -98.959103678  
## THGFSFC 0.004124619 -0.009849176 -0.009493841  
##   
## $ycoef  
## [,1] [,2] [,3]  
## THGSDFC 0.011415578 -0.010169482 -0.014106076  
## TCSDFB -0.077556675 -0.037720634 0.072787341  
## TPRSDFB -0.002969355 0.002268621 -0.004222605

1. What can you conclude from the above analyses?

***The analysis above shows the canonical correlation coefficients test for the existence of overall relationships between two sets of variables. The coeffcients are low and which means that the water variables and the soil variables are not positively correlated with each other. The R-squared values are very low.***

Answer the following questions regarding the canonical variates.

1. Give the formulae for the significant canonical variates for the soil and water variables.

***The linear combination of the sets of variables (predictor and DV). Significant canonical variates will have a low p-value (< 0.05).***

1. Give the correlations between the significant canonical variates for soils and the soil variables, and the correlations between the significant canonical variates for water and the water variables.

# compute canonical loadings  
cc2 <- comput(water, soil, cc1)  
  
# display canonical loadings/latent variables  
cc2[3:6]

## $corr.X.xscores  
## [,1] [,2] [,3]  
## MEHGSWB -0.2138288 -0.54424426 0.05580913  
## TURB -0.1207027 -0.03435814 0.49853147  
## DOCSWD -0.8920181 -0.39006177 0.02464817  
## SRPRSWFB -0.1719363 0.58138401 -0.63983875  
## THGFSFC 0.4914315 -0.62009828 -0.52589688  
##   
## $corr.Y.xscores  
## [,1] [,2] [,3]  
## THGSDFC -0.003665011 -0.30485575 -0.12523874  
## TCSDFB -0.246423901 -0.26504660 0.00980968  
## TPRSDFB -0.275332457 0.05094524 -0.18310544  
##   
## $corr.X.yscores  
## [,1] [,2] [,3]  
## MEHGSWB -0.08244902 -0.18776307 0.014932836  
## TURB -0.04654108 -0.01185348 0.133391950  
## DOCSWD -0.34394820 -0.13457045 0.006595106  
## SRPRSWFB -0.06629592 0.20057620 -0.171201505  
## THGFSFC 0.18948827 -0.21393254 -0.140714106  
##   
## $corr.Y.yscores  
## [,1] [,2] [,3]  
## THGSDFC -0.009505083 -0.8836455 -0.46806012  
## TCSDFB -0.639092107 -0.7682559 0.03666214  
## TPRSDFB -0.714065477 0.1476683 -0.68432782

# tests of canonical dimensions  
ev <- (1 - cc1$cor^2)  
  
n <- dim(water)[1]  
p <- length(water)  
q <- length(soil)  
k <- min(p, q)  
m <- n - 3/2 - (p + q)/2  
  
w <- rev(cumprod(rev(ev)))  
  
# initialize  
d1 <- d2 <- f <- vector("numeric", k)  
  
for (i in 1:k) {  
 s <- sqrt((p^2 \* q^2 - 4)/(p^2 + q^2 - 5))  
 si <- 1/s  
 d1[i] <- p \* q  
 d2[i] <- m \* s - p \* q/2 + 1  
 r <- (1 - w[i]^si)/w[i]^si  
 f[i] <- r \* d2[i]/d1[i]  
 p <- p - 1  
 q <- q - 1  
}  
  
pv <- pf(f, d1, d2, lower.tail = FALSE)  
(dmat <- cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv))

## WilksL F df1 df2 p  
## [1,] 0.6963021 4.051995 15 433.8093 6.185853e-07  
## [2,] 0.8179043 4.176302 8 316.0000 9.094796e-05  
## [3,] 0.9284064 4.087068 3 159.0000 7.921523e-03

1. What can you conclude from the above analyses?

***The analysis above that the soil variable group is more related to each other than to the water variable group and the same logic for the water being correlated with the water variables but has very little correlation to the soil group. The canonical variates are interesting enough to use to represent the relationship.***

# Problem 3 (Principal Component Analysis - 20 points): The data given in the file ‘problem3.txt’2 is the percentage of people employed in different industries in European countries during 1979. Techniques such as Principal Component Analysis (PCA) can be used to examine which countries have similar employment patterns. There are 26 countries in the file and 10 variables as follows:

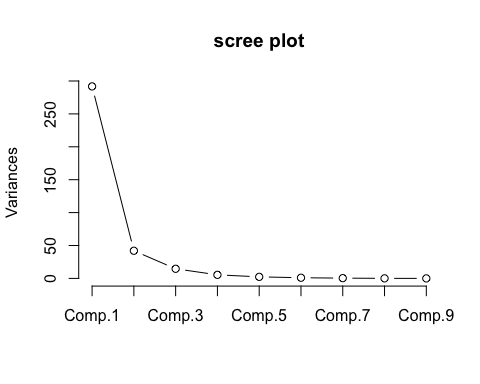
# get the data  
europe <- read\_tsv("problem3.txt")  
# source: http://www.statmethods.net/advstats/factor.html  
# from the correlation matrix   
fit2 <- princomp(europe[,-1], cor=F, n.obs=.) # -1 to get a numeric matrix, cor=F is a covariance matrix  
summary(fit2) # print variance accounted for

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4  
## Standard deviation 17.0817636 6.4823470 3.82393204 2.32861792  
## Proportion of Variance 0.8157836 0.1174827 0.04088179 0.01516024  
## Cumulative Proportion 0.8157836 0.9332663 0.97414811 0.98930835  
## Comp.5 Comp.6 Comp.7 Comp.8  
## Standard deviation 1.532782553 1.002896265 0.63612956 0.2498589145  
## Proportion of Variance 0.006568567 0.002812041 0.00113136 0.0001745417  
## Cumulative Proportion 0.995876918 0.998688959 0.99982032 0.9999948600  
## Comp.9  
## Standard deviation 4.287707e-02  
## Proportion of Variance 5.139960e-06  
## Cumulative Proportion 1.000000e+00

loadings(fit2) # pc loadings

##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
## Agr 0.892 0.118 0.180 -0.153 0.335  
## Min 0.456 -0.766 0.290 0.324  
## Man -0.271 0.770 0.185 0.336 -0.201 0.162 0.337  
## PS 0.231 -0.909 0.340  
## Con -0.724 -0.558 -0.194 0.325  
## SI -0.192 -0.234 -0.580 -0.608 0.266 0.104 0.337  
## Fin -0.130 -0.470 0.781 0.121 0.123 0.334  
## SPS -0.298 -0.567 0.598 0.236 -0.248 0.332  
## TC 0.159 -0.435 0.546 0.567 0.224 0.334  
##   
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8  
## SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000  
## Proportion Var 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111  
## Cumulative Var 0.111 0.222 0.333 0.444 0.556 0.667 0.778 0.889  
## Comp.9  
## SS loadings 1.000  
## Proportion Var 0.111  
## Cumulative Var 1.000

plot(fit2,type="lines", main="scree plot") # scree plot



fit2$scores # the principal components

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## [1,] -17.516687 -4.92622849 -2.35528094 -0.1940007 0.49072745  
## [2,] -11.496688 -11.66176637 3.00202830 2.5328564 -0.26205034  
## [3,] -9.128686 -2.16828207 -2.75030568 0.1289183 0.37355869  
## [4,] -14.393424 5.04749385 0.20568951 1.2143729 2.57857191  
## [5,] 4.458174 -6.13156498 -1.92400082 -3.4809060 0.34805906  
## [6,] -4.026684 -0.38889529 -2.40586194 -4.8100625 -0.24693944  
## [7,] -12.089752 2.33236877 -4.62806669 -1.9653471 -0.03246954  
## [8,] -13.900455 -9.72359023 -1.70981367 0.6817669 -1.51095027  
## [9,] -18.728675 -3.33178946 0.58938403 1.0073981 2.29959070  
## [10,] -6.471418 3.35662962 -4.75660272 -1.2741444 -0.51798957  
## [11,] -6.837047 -3.97634061 0.06757235 0.9640481 0.27383594  
## [12,] 25.427083 -1.80467718 -2.91613130 -2.8567757 -1.87953959  
## [13,] -10.972019 -8.85877780 0.22621023 -0.5923626 -2.00656646  
## [14,] 9.403865 -0.08570061 -1.23656256 -2.1803190 0.06780428  
## [15,] 5.774973 6.15867547 -4.87904446 4.4809274 -3.04038851  
## [16,] -15.311975 -8.52674423 3.92210148 2.6339347 1.41721975  
## [17,] -12.683839 9.77920054 -5.68921238 -1.0438972 1.24534771  
## [18,] 52.115644 -8.64165980 2.96515501 -1.7960412 3.17620617  
## [19,] 4.156791 6.70685051 4.93995679 -0.1372004 0.56054055  
## [20,] -3.246127 9.23467980 3.78225558 -0.1477919 -0.08257234  
## [21,] -17.415527 10.73233092 4.89564722 0.2453801 1.36457552  
## [22,] 3.135737 4.98695108 2.98354179 -0.7306784 -1.31541699  
## [23,] 13.315709 2.94482700 3.58894681 -0.4854554 -1.36139215  
## [24,] 17.011336 9.12523022 2.58152423 0.2380157 -0.03962236  
## [25,] 4.587043 -0.87197041 8.44875566 0.9788693 -2.93255455  
## [26,] 34.832648 0.69274975 -6.94788580 6.5884944 1.03241438  
## Comp.6 Comp.7 Comp.8 Comp.9  
## [1,] 0.58725632 0.02747862 0.261512512 -0.008040795  
## [2,] -1.00014269 -0.16185051 0.113221471 0.098177832  
## [3,] -0.74062040 -0.44855421 -0.072003249 -0.002822468  
## [4,] -0.26124048 0.38250268 0.114442694 -0.043957458  
## [5,] 0.53989717 -0.35750610 -0.420871836 0.072290598  
## [6,] -1.50206770 -0.24928213 0.051645913 0.034174567  
## [7,] 1.08490154 -1.70709926 0.657656083 0.021025084  
## [8,] -0.83583339 -0.51429138 -0.273012947 -0.066889344  
## [9,] 0.44287447 -0.44816057 -0.272105619 -0.033049026  
## [10,] 0.83150513 0.50543393 -0.255764025 -0.015749373  
## [11,] 0.68129276 0.92797085 -0.279485018 0.039399018  
## [12,] 0.59887931 0.83983051 0.137603535 -0.016322341  
## [13,] 1.41674308 0.88213892 0.371151398 -0.058993757  
## [14,] -1.03041185 0.52381804 -0.075302827 0.017535977  
## [15,] -1.70355376 -0.25749789 -0.039086940 -0.039716393  
## [16,] -0.73494415 -0.03745853 -0.024572541 -0.018992003  
## [17,] -1.16079249 1.25830466 0.004260537 0.008689042  
## [18,] -0.36536971 -0.05968267 0.116399862 -0.058269133  
## [19,] -0.54824453 0.22155473 0.184340613 -0.023526405  
## [20,] 0.24180861 -0.57988066 -0.123332326 -0.065489414  
## [21,] 1.27464580 0.43732622 0.146397327 0.046067998  
## [22,] 1.88083050 -0.57747180 -0.571402883 -0.003437050  
## [23,] 0.32653024 -0.66030520 0.017151170 -0.008766194  
## [24,] -1.46062121 -0.55746312 -0.026902957 0.047287948  
## [25,] -0.07846592 0.62537794 0.198893703 0.034657402  
## [26,] 1.51514335 -0.01523307 0.059166352 0.044715689

Perform a principal component analysis using the covariance matrix:

1. How many principal components are required to explain 90% of the total variation for this data?

***At the 2nd principal component, the cumulative proportion of variance is at 0.9332663 (~93%) meeting the explained threshold of 90% of the total variation for this data.***

1. For the number of components in part a, give the formula for each component and a brief interpretation.

***The greatest variance by some projection of the data comes to lie on the first principal component - Agr, the second greatest variance on the second coordinate - Min. The agriculture & mining industry is highly specialized and it appears from this dataset that skill-set is limited to certain countries and not widespread in others. Those industries also heavily depend on environmental factors (warm countries farm more and mining is very costly to the environment so government regulation has decreased this industry.) and import/export trends.***

PC1 = b11 X1 + b21 X2 + ... bk1 Xk where X = Agriculture (Agr)

PC2 = b12 X1 + b22 X2 + ... bk2 Xk where X = Mining (Min)

1. What countries have the highest and lowest values for each principal component (only include the number of components specified in part a). For each of those countries, give the principal component scores (again only for the number of components specified in part a).

***The countries with the highest value for principal component 1 are: Turkey, Yugoslavia, & Greece. The lowest values are: United Kingdom, Belgium, E. Germany. The countries with the highest value for principal component 2 are: E. Germany, Switzerland, & Czechoslovakia. The lowest are Denmark, Netherlands, Norway.***

# this just identifies the top and bottom PC values and can be match to the row number pertaining to that country  
d = fit2$scores   
d[order(d[, 1],decreasing=T)[1:3], 1] # top 3 countries for PC1

## [1] 52.11564 34.83265 25.42708

d[order(d[, 1],decreasing=F)[1:3], 1] # lowest 3 PC countires for PC1

## [1] -18.72867 -17.51669 -17.41553

d[order(d[, 2],decreasing=T)[1:3], 2] # top 3 countries for PC2

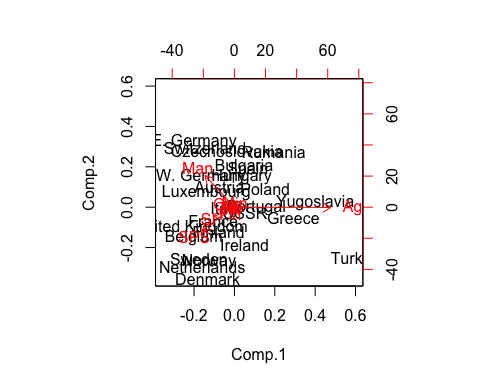
## [1] 10.732331 9.779201 9.234680

d[order(d[, 2],decreasing=F)[1:3], 2] # lowest 3 PC countires for PC2

## [1] -11.661766 -9.723590 -8.858778

1. Include and interpret the scatter plot of the data using the first two principal components.

***The biplot uses points to represent the scores of the observations on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components. In this example, the points represent Countries, and the vectors represents employment industries. The countries that a close to together in value have a similar employment patterns. A vector arrow points in the direction which is most like the variable represented by the vector. This is the direction which has the highest squared multiple correlation with the principal components. The length of the vector is proportional to the squared multiple correlation between the fitted values for the variable and the variable itself. Vectors that point in the same direction correspond to variables that have similar response profiles, and can be interpreted as having similar meaning in the context set by the data.*** *Source:* <http://forrest.psych.unc.edu/research/vista-frames/help/lecturenotes/lecture13/biplot.html>

biplot(fit2, xlabs = europe[,1]) # biplot

# Problem 4 (overview – 5 points): Briefly describe the similarities and differences between:

1. Linear regression and canonical correlation

***Linear regression is an approach for modeling the relationship between a dependent variable y and one or more explanatory variables (or independent variables), X. CCA is a method to find relationships between two sets of variables (independent and dependent). CCA is not as common as Linear Regression. CCA focueses on how the best linear predictors relate ot the best DVs. In Linear regression the goal is also to identify covariates that the effectively model (or account for the most error) for the target variable.***

1. Canonical correlation and principal component analysis

***PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. PCA finds weights to maximize variance, and finds an ideal projection (orthogonal transformation) to convert a set of observations of possibly correlated variables into a set of values linearly correlated variables. PCA removes redundancy and multicolinearity from one set of variables which differs from CCA which utilizes two sets of variables. PCA and CCA both seek to optimize the linear combinations therefore reducing the amount of variables.***