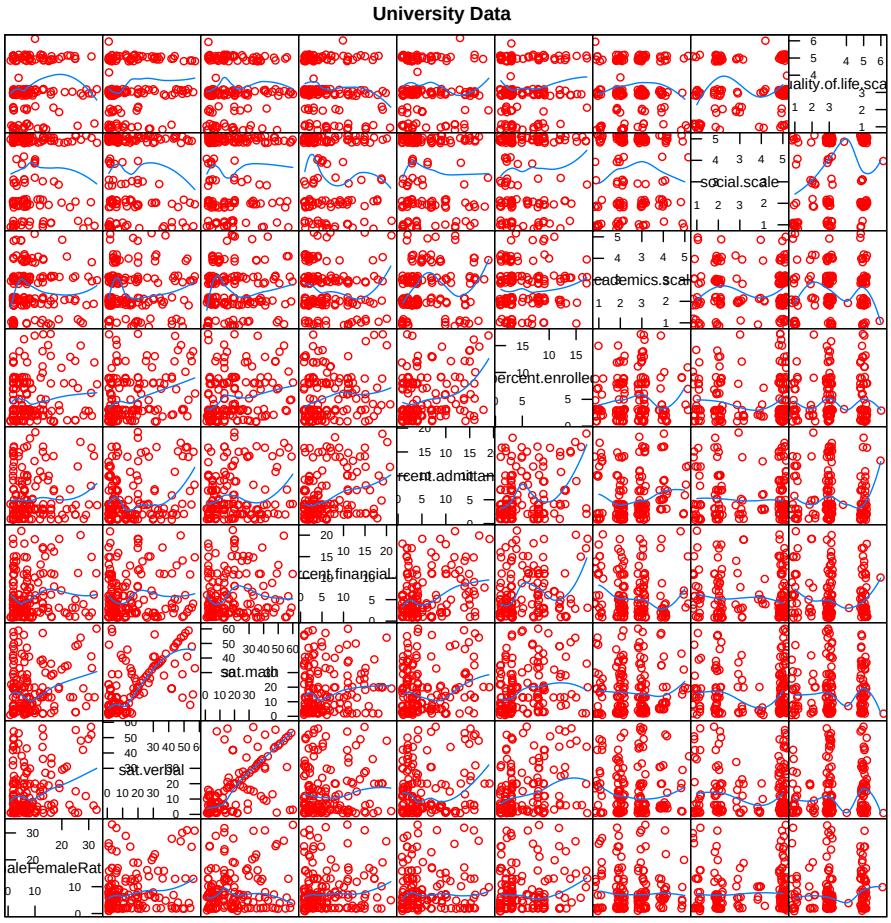
3/4/2010 36-402/608 ADA-II H. Seltman
Breakout #15 Results

Summary: To perform a canonical correlation analysis in R, use `cancor(X, Y)` which produces a list with components "cor" (correlations of the canonical variable pairs), "xcoef" and "ycoef" (coefficients transforming data to canonical variables), and "xcenter" and "ycenter" (original X and Y column means). There is no special `print()` or other methods, but running `cancor()` does a special printout. The built in function does not compute p-values; use `p.asym()` (or `p.perm()` in package "CCP" to get the p-values.

The "University Data Set" at the UCI Machine Learning Repository has records on college and university characteristics, presumably collected in the 1980s. We will look at data on 242 institutions, considering several objective measures of who is admitted as the explanatory variables. Our outcomes are three (subjective) quality measures. We will use CCA to find dimensions of university characteristics that predict (correlate with) the measures of quality.

```
names(univ)[xvar]
# [1] "maleFemaleRatio"    "sat.verbal"       "sat.math"  "percent.financial.aid"
# [5] "percent.admittance" "percent.enrolled"
names(univ)[yvar]
# [1] "academics.scale"       "social.scale"          "quality.of.life.scale"

round(cor(univ[,xvar],univ[yvar]),2)
#                       academics.scale social.scale quality.of.life.scale
# maleFemaleRatio                 -0.04        -0.01                   0.11
# sat.verbal                      -0.03        -0.04                   0.16
# sat.math                         0.02         0.03                   0.08
# percent.financial.aid           -0.12        -0.13                  -0.14
# percent.admittance               0.11        -0.05                  -0.11
# percent.enrolled                 0.07         0.07                   0.06
```



Question 1: What do you observe on the original scales?

The correlations are all fairly weak (and remember these are "r" values, so the highest r^2 here is $0.16^2 = 0.026$. Academics is slightly positively correlated with percent admittance and negatively with percent receiving aid, while QOL is positively associated with sat verbal and negatively with percent aid and percent admittance. Social life is slightly negatively correlated with % financial aid. We hope higher correlation will come from CCA.

```
cc=cancor(univ[,xvar],univ[,yvar])

print(round(cc$cor,3))
#  0.271 0.257 0.111

print(round(cc$xcoef,4))
#                      [,1]    [,2]    [,3]    [,4]    [,5]    [,6]
# maleFemaleRatio    0.0022  0.0031 -0.0001 -0.0021 -0.0098  0.0005
# sat.verbal         0.0010  0.0043 -0.0043  0.0022  0.0022 -0.0002
```

```
# sat.math                0.0007 -0.0017  0.0026 -0.0057  0.0011  0.0014
# percent.financial.aid -0.0128  0.0070  0.0034  0.0002  0.0005  0.0069
# percent.admittance    -0.0034 -0.0107 -0.0138 -0.0025 -0.0036  0.0003
# percent.enrolled       0.0096 -0.0062  0.0063  0.0104 -0.0009  0.0134

print(round(cc$ycoef,3))
#                        [,1]   [,2]   [,3]
#academics.scale       0.025 -0.059 -0.047
#social.scale          0.021 -0.020  0.041
#quality.of.life.scale 0.046  0.035 -0.015

require(CCP)
p.asym(cc$cor, nrow(univ), length(xvar), length(yvar))
#Wilks' Lambda, using F-approximation (Rao's F):
#              stat    approx df1      df2   p.value
#1 to 3:  0.8551364 1.2977101  18 410.6072 0.1846153
#2 to 3:  0.9226508 1.1993408  10 292.0000 0.2908825
#3 to 3:  0.9877572 0.4554981   4 147.0000 0.7682613
```

Question 2: What tells you that we haven't found any interesting new scales? If you pretend that the first p-value is 0.00185, what conclusions would you reach?

Even though we have found some higher correlation values, the "1 to 3" p-value retains the null hypothesis of no correlations at all in the data. If the first CVs were correlated we would explain it as a smaller than average financial aid percent coupled with a larger than average percent enrolled is associated with a high mean rating (with extra emphasis on QOL).

The Greek air pollution example show how CCA can pull out meaningful variables such as the set of pollutants associated with two types of combustion, the separate ozone pollution and the weather patterns that are correlated with these.