HW 6 STAT461

Ronald Barry

2/26/2022

Problem One

Use the following code to read in the Harvard Forest Dataset HF143.CRUI Land Use Project - Soil Properties data. It is in file datasoil.txt. It consists of soil properties along three transects in Harvard Forest, collected by Richard Bowden, Charles McClaugherty and Timothy Sipe (see http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf143).

```
f <- file.choose()  #select the text file
f
dat <- read.csv(f, header=TRUE)
class(hold)  #just checking to see if the data is a data.frame, a matrix, etc.
dim(dat)
#
#  Here you'll want to trim columns you don't want to use
# and perhaps cases you don't want to use, while keeping as much
# data as practical.  I'll give advice on this if you wish, but
# there isn't much point to keeping the first four columns.
#
out <- factanal(dat,factors=4,rotation="varimax",scores="regression")
out
```

Information on the collection and coding of this data:

https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=HF143

a. Look at the output. Is a decent amount of the variability explained by the first 4 factors? Use the test of hypothesis of sufficient numbers of factors to find a suitable number of factors to use. What proportion of the variability in the data is explained by the factor analysis?
b. Look at the factor loadings. Can you roughly interpret the loadings on the first factor? Is it always possible to do so? Why or why not?
c. Try a promax rotation. What is this and how does it differ from varimax rotation? Did the proportion of variation explained or the hypothesis test change by much? Why is this result reasonable? Why did it show factor correlations for promax and not for varimax?
d. What does the plot of scores tell you? What are scores?

```
plot(out$scores[,1],out$scores[,2],pch=19,cex=0.5)
text(out$scores[,1],out$scores[,2]-0.1,1:153,cex=0.5)
```

Problem Two

Using the same dataset as in problem one, run a principal components analysis. Use a screeplot to select the number of important PCs. Do the number of PCs seem to match what you got with factor analysis? Do the loadings look similar to those from the factors you calculated in b and c? Why do you think this happened?

Problem Three

Below is a covariance matrix based on N = 150 first year college students.

```
M <-
structure(c(0.594, 0.483, 3.993, 0.426, 0.5, 0.483, 0.754, 3.626,
1.757, 0.722, 3.993, 3.626, 47.457, 4.1, 6.394, 0.426, 1.757,
4.1, 10.267, 0.525, 0.5, 0.722, 6.394, 0.525, 2.675), .Dim = c(5L,
5L), .Dimnames = list(c("GPAreq", "GPAelec", "SAT", "IQ", "EdMot"
), c("GPAreq", "GPAelec", "SAT", "IQ", "EdMot")))
#
M
```

```
##         GPAreq GPAelec    SAT     IQ EdMot
## GPAreq   0.594   0.483  3.993  0.426 0.500
## GPAelec  0.483   0.754  3.626  1.757 0.722
## SAT      3.993   3.626 47.457  4.100 6.394
## IQ       0.426   1.757  4.100 10.267 0.525
## EdMot    0.500   0.722  6.394  0.525 2.675
```

(Finn, J. D. (1974). A general model for multivariate analysis. New York: Holt, Reinhart & Winston.)

- GPAreq is GPA in required courses
- GPAelec is GPA in elective courses
- SAT is SAT scores from high school
- IQ is IQ determined in the last year of high school
- EdMot is a measure of educational motivation

a. Try to make up two or three reasonable path analysis models for this data. Draw a structural diagram for each. Also, list all of the parameters of each model.
b. Using lavaan, run each model. Which model seems to fit the data best? How good is the fit of the model?
c. In your best fitting model, are any of the links candidates for removal? How could you determine this?
d. Why can’t we use bootstapping to get standard errors in this situation?
e. Is the model you selected as best recursive, or non-recursive? How can you tell?
f. Are there any indirect effects in your best model? What are they?
g. What variables in your best model are exogenous, and which are endogenous?