EPSY 8265: Take-home midterm exam
Due 5PM, November 8 (Wed), 2023

1. [20pt] The datafile called “world.csv” that contains data collected from a number of different countries. You will see that one variable identifies the country name (‘country’), and another the region in which the country is located (‘region’). The other variables are:
    - lifeexpf: Average female life expectancy
    - lifeexpm: Average male life expectancy
    - literacy: People who read (%)
    - popincr: Population increase (% per year))
    - babymort: Infant mortality (deaths per 1000 live births)
    - birthr: Birth rate per 1000 people
    - deathr: Death rate per 1000 people
    - gdp: Log (base 10) of GDP
    - aidsr: Log (base 10) of AIDS rate
    - bdratio: Birth to death ratio
    - fertility: average number of kids
    - literacym: Males who read (%)
    - literacyf: Females who read (%)

[Note that some of the variables have been log-transformed. A log transformation preserves the original direction of the variable, so should not influence your interpretation in terms of the sign of the loadings/weights.]

You may import this data into R by using the following code:

```
world <- read.csv("world.csv", header = T)
```

Perform a principal component analysis using all but the country name and region variables. This dataset has some missing values. For the exam, you may ignore the missing values by putting `na.omit()` around the data object name.

(a) Perform two separate analyses using the covariance and correlation matrices. Do the loadings differ across the two solutions? If so, why? Which type of matrix do you think is more appropriate to analyze for this data? [5pt]

Based on the solution that you think is more appropriate, answer the following questions.

(b) How many principal components appear to be necessary to describe adequately the variability in these variables? [3pt]

(c) Provide an interpretation to the principal components that you determined to extract. What types of countries would receive high and low scores on each component? [4pt]

(d) Obtain component scores for each of the countries on the components you extracted. Request a scatter plot of the first two component scores, using the region variable to label the points in the plot. Does there appear to be an association between either of the component scores and the region a country belongs to? [5pt]

(e) Show by hand calculation that the weight vectors for the first two components are orthogonal. [3pt]

2. [20pt] The correlation matrix (shown below) was obtained from responses of 230 individuals on 11 subtests of the Wechsler Adult Intelligence Scale (WAIS). The 11 tests were as follows: (1) Information, (2) Comprehension, (3) Arithmetic, (4) Similarities, (5) Digit Span, (6) Vocabulary, (7) Digit Symbol, (8) Picture Completion, (9) Block Design, (10) Picture Arrangement, and (11) Object Assembly.

You may copy and paste the below code to read in the correlation matrix in R.

```
wais <- matrix(c(1.00, .37, .34, .40, .27, .59, .09, .25, .27, .22, .26,
                 .37, 1.00, .27, .25, .38, .46, .10, .26, .29, .22, .24,
                 .34, .27, 1.00, .36, .28, .33, .18, .32, .38, .29, .30,
                 .40, .25, .36, 1.00, .22, .35, .08, .31, .26, .25, .20,
                 .27, .38, .28, .22, 1.00, .29, .16, .14, .18, .15, .22,
                 .59, .46, .33, .35, .29, 1.00, .08, .27, .24, .28, .26,
                 .09, .10, .18, .08, .16, .08, 1.00, .19, .13, .22, .17,
                 .25, .26, .32, .31, .14, .27, .19, 1.00, .36, .36, .40,
                 .27, .29, .38, .26, .18, .24, .13, .36, 1.00, .30, .60,
                 .22, .22, .29, .25, .15, .28, .22, .36, .30, 1.00, .25,
                 .26, .24, .30, .20, .22, .26, .17, .40, .60, .25, 1.00),
ncol=11, byrow=T)

rownames(wais) <- colnames(wais) <- c("Information", "Comprehension",
"Arithmetic", "Similarities", "Digit Span", "Vocabulary", "Digit Symbol",
"Picture Completion", "Block Design", "Picture Arrangement", "Object
Assembly")
```

Using a principal component analysis of this correlation matrix:

(a) Identify an appropriate number of principal components. [3pt]

(b) Compute the *component loadings* (i.e., correlations between the variables and PC scores in this case) for the principal components you determined to extract, and based on these values, give an interpretation to the components. [5pt]

Using a maximum likelihood factor analysis solution:

(c) Perform a statistical test to determine the appropriate number of factors. Give an interpretation to each of the factors. Contrast the results with what was obtained in part (a) using components. [5pt]

(d) Verify by hand calculation that the estimated communalities for the Information and Vocabulary subtests as reported in the R output can be computed from their estimated factor loadings. [3pt]

(e) Based on the estimated factor solution, compute the "fitted" correlation between the Information and Comprehension subtests. How does that fitted correlation compare to the sample correlation observed between there two tests (i.e., what is residual)? [4pt]

3. [10pt] Using the following 3×3 sample covariance matrix, answer to the following questions.

$$\mathbf{S} = \begin{bmatrix} 25 & & \\ 8 & 15 & \\ 7 & 6 & 10 \end{bmatrix}$$

(a) Determine estimates of the parameters for a one factor model. (You should compute estimates for a total of six parameters: $\lambda_1, \lambda_2, \lambda_3, \psi_1, \psi_2, \psi_3$) [6pt]

(b) Suppose you wanted to fit a two-factor model to this covariance matrix. What problem do you encounter? [4pt]