STAT 740 – Spring 2019 – Exam 1

*Directions:*

* *This exam consists of five (5) questions that are not equally weighted; the number of points a problem is worth is indicated next to the item stem.*
* *Sub-parts may also be worth different point values, but these will be determined during grading.*
* *Each problem is on a different page for convenience.*
* *You should provide your answers in this file, in the space below each sub-part. It would be helpful if you would change the font color of the questions or answers to something else (e.g. red) to help with scoring, but this is optional.*
* *You may submit your answers in LaTeX form instead of using this file if you so choose.*
* *You should submit your answer file (e.g. this file) as well as a separate file containing your R code.*
* *If necessary, you may insert photographs of handwritten work in lieu of a typed answer. This should be rare.*
* *You have 72 hours from when you were emailed this exam to submit it.*
* *Submissions should be made to the Dropbox on D2L.*
* *It is okay to alter data files (e.g. to change the file format, add headers, or other data management/processing in another program such as Excel), but you must submit them on D2L also. (If you alter the data files and don’t submit them, your code won’t run for me!)*
* *You may reference our course materials (e.g. the book, slides, notes, etc.), but you may not use outside resources (e.g. Google). An exception to this is you may seek out R help about specific functions (e.g. for Question 5) in any source, but you still may not seek out specific outside help about the exam.*
* *If you have any questions, email me (*[whitakerdo@uwstout.edu](mailto:whitakerdo@uwstout.edu)*).*
* *The total exam is graded out of* ***50 points****.*

1. **(20 points)** The dataset PSYCHPROFILE.DAT consists of 130 observations generated by scores on a psychological test administered to Peruvian teenagers (ages 15, 16, ad 17). The gender (male = 1, female = 2) and socioeconomic status (low = 1, medium = 2) were also recorded. The scores were accumulated into five subscale scores labeled *independence* (indep), *support* (supp), *benevolence* (benev), *conformity* (conform), and *leadership* (leader).
   1. Calculate  , , and using matrices.
   2. Calculate . What is the difference between and ? Why might you prefer one over the other?
   3. What would the value of be if you calculated it from instead of ? Why?
   4. Construct an appropriate multivariate graphical display for these data.
   5. Examine each of the five subscale scores for marginal normality. Describe your findings.
   6. For any subscales in part (e) that you identify as non-normal, suggest an appropriate transformation for the data.
   7. Assess the multivariate normality of the five subscale scores.
   8. Suppose an overall psychological profile score is to be calculated for each participant averaging their five subscale scores. Considering only the new overall profile score, what are the and matrices? (Answer this without actually creating a profile score for each participant.)
2. **(5 points)** A necessary and sufficient condition for the positive definiteness of a symmetric matrix is given below.

**Theorem.** Let represent a symmetric matrix, where , and is partitioned as

where is a symmetric matrix with dimensions , called the leading principle submatrix. is positive definite if and only if is positive definite and .

1. Use this theorem to show that the following matrix is positive definite.
2. Describe and use another approach for demonstrating that is positive definite.
3. **(5 points)** Given the matrix answer the following questions:
   1. What is the inverse of ?
   2. What are the eigenvalues and eigenvectors of?
   3. What is the spectral decomposition of ?
4. **(5 points)** Consider the matrices below.
   1. Compute and . Describe what you found.
   2. Propose an explanation for what you found in part (a).
5. **(15 points)** The dataset for this problem is available online at:

<https://edg.epa.gov/data/Public/ORD/NHEERL/EQI/EQI_RESULTS_2013JULY22.csv>

The official abstract of the dataset is as follows:

The US Environmental Protection Agency's (EPA) National Health and Environmental Effects Research Laboratory (NHEERL) in the Environmental Public Health Division (EPHD) is currently engaged in research aimed at developing a measure that estimates overall environmental quality at the county level for the United States. This work is being conducted as an effort to learn more about how various environmental factors simultaneously contribute to health disparities in low-income and minority populations, and to better estimate the total environmental and social context to which humans are exposed. This dataset contains the finalized Environmental Quality Index (EQI), and an index for each of the associated domains (air, water, land, built environment, and sociodemographic environment). Indices are at the county level for all counties in the United States.

The variables of interest for this problem are:

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Measurement Type |
| stfips | FIPS (Federal Information Processing Standard) code that identifies the state and county associated with the variables | Nominal |
| county\_name | The county name | Nominal |
| state | The state name | Nominal |
| cat\_rucc | Rural-Urban Continuum Code:   * 1 represents metropolitan-urbanized, such as:   + Hennepin County, the location of Milwaukee, WI   + Eau Claire County, the location of Eau Claire, WI * 2 represents nonmetropolitan-urbanized, such as:   + Manitowoc County, the location of Manitowoc, WI * 3 less urbanized, such as   + Dunn County, the location of Menomonie, WI * 4 thinly populated, such as   + Pepin County, the location of Durand, WI | Ordinal |
| air\_EQI\_22July2013 | EQI for data sources in the Air Domain   1. E.g. National-Scale Air Toxics Assessment | Scale |
| water\_EQI\_22July2013 | EQI for data sources in the Water Domain   1. E.g. National Contaminant Occurrence Database | Scale |
| land\_EQI\_22July2013 | EQI for data sources in the Land Domain   1. E.g. National Geochemical Survey | Scale |
| sociod\_EQI\_22July2013 | EQI for data sources in the Sociodemographic Domain   1. E.g. Uniform Crime Reports | Scale |
| built\_EQI\_22July2013 | EQI for data sources in the Built-Environment Domain   1. E.g. Housing and Urban Development Data | Scale |
| EQI\_22July2013 | Environmental Quality Index (EQI) for the county   1. Note that HIGHER values suggest WORSE environmental quality | Scale |

A full report on the creation of the EQI variables from the EPA is available here: <https://edg.epa.gov/data/PUBLIC/ORD/NHEERL/EQI/EQI%20Overview%20Report__Final.pdf> (and is included with the exam in the supplements folder). You do NOT need to download or read this report, but it is available to you if you want more information. (For example, the report contains several maps displaying EQI data at the county-level which may aid in understanding the dataset.) Note that this is an Overview Report for a non-technical audience; a separate Technical Report is also available that describes the technical details of the analysis in detail.

1. Obtain , , and for the EQI variables.
2. Using Generalized Variance, compare the environmental quality of Alabama (AL), California (CA), Connecticut (CT), and Wisconsin (WI) using all six (6) EQI variables; interpret these Generalized Variance values. Also note any potential problems with the use of the Generalized Variance and make adjustments as-needed to compare these four states.
3. Nevada has 17 counties. What is the correlation between each of these counties? Do you notice any patterns?   
   *Hint: Consider installing and using the corrplot package to create a graph of the correlations. It also may help to order the rows of the Nevada data subset by cat\_rucc before calculating the correlation matrix. The order function in R might help with this. A small example of using the function for you to run in R:*

fruit <- structure(list(Name = structure(c(5L, 2L, 4L, 3L, 7L, 1L, 6L),

.Label = c("Avocado", "Banana", "Grapefruit", "Orange", "Pineapple", "Starfruit", "Tangerine"),

class = "factor"), Tier = c(2L, 3L, 1L, 4L, 2L, 2L, 3L)), class = "data.frame",

row.names = c(NA, -7L))

print(fruit) # Notice the row order

print(fruit[order(fruit$Tier),]) # Notice the row order has changed

*You may need to consult R documentation about corrplot or order if you use them because we have not yet used them in this class.*

1. Are the EQI variables normally distributed (marginally)?