Psychology 613: Multivariate Statistics (Data Analysis III)
Midterm

The following midterm is due at **5pm** on **Thursday, May 5th**, via Canvas. **All problems must be completed individually from start to finish.** The document that you turn in should consist of a write-up, a copy of your code (numbered to correspond with the questions), and a copy of your output (also numbered). Questions that ask you to interpret results of analyses should be phrased in terms that are substantively meaningful to another human being. You should phrase interpretations in terms of the psychological constructs, e.g., "homework was significantly positively related to strength training self-efficacy."

This exam will be graded **out of 50 points**. So, **please choose your own combination of problems to add up to 50 possible points**. Extra credit **up to 60 total points** is allowed.

Matrix algebra (10 points each for Problems 1 and 2)

1. Students in 613 can be divided into three groups: those who own 1 computer, those who own 2 computers, and those who own 3 or more computers. The level of stats-related anxiety for five of those with 1 computer is: 4, 3, 4, 2, 3 (on a 5-point scale); the level of anxiety for five of those with 2 computers is: 4, 3, 3, 2, 3 (on the same scale); finally, the level of anxiety for five of those with 3 or more computers is: 2, 3, 2, 2, 3.

   Write down the GLM equation to predict stats-related anxiety based on this grouping variable. Solve for the parameter vector (β) and the error vector (ε) **by hand (showing all of your work)**. (Hint: for help on calculating the inverse of X, see Lecture 1, pp. 4.)

2. It turns out that the more statistics courses you have taken, the more computers you have destroyed out of frustration. I would like to know the equation that relates the two variables so that I can predict the number of computers future students are likely to have broken based on how many stats courses they have taken.

   These two variables are stored in a datafile called **First_day_survey.csv**: *nStatsCourses* and *nCompBreak*. Using **R code (i.e., solve the GLM using your own code – not lm())**, run a regression predicting *nCompBreak* from *nStatsCourses*. Be sure to calculate the Bs (unstandardized coefficients), the SS-regression, the SS-error, the df-regression, the df-error, the F-value for the regression model, and the p-value associated with that F-value.

Programming in R (20 points)

3. Run a simulation demonstrating the effect of sample size on correlation. Use a range of sample sizes from N=2 to N=500. For each different sample size, draw a sample from a random normal distribution (use the *rnorm()* function). Let that sample be your DV. Then create an IV that is equal to your DV plus some random noise (again using *rnorm()*). If the SD of your random noise is the same as the SD of your DV, the "true" correlation between the DV and the IV should be 0.5. Calculate what the correlation is between your IV and DV for each sample size. Plot the correlations as a function of sample size.

Interaction and moderation (20 points)

4. Using the survey data from the first day of class (**First_day_survey.csv**), I would like to test whether the type of computer you use (3 *categorical* groups: Mac (0), PC (1), or other (2); variable=*MacPC*) interacts with the number of stats classes you've taken (*continuous* variable = *nStatsCourses*) to predict your *continuous* comfort with statistics (*Comfort*): Does the effect of previous stats experience on comfort differ as a function of what kind of computer you use?

   Center the appropriate variables, create the interaction term, and run a hierarchical regression predicting *Comfort* from the two main effects (*MacPC* and *nStatsCourses*) and their interaction. *Regardless of whether or not they are significant*, plot the simple slopes of stats experience on comfort for each type of computer (using a computer) and test the significance of at least one of them. Report all of the parameters and the R^2 change test, and interpret your findings.

Problems 5 and 6 rely on the dataset called **StudentRatings.sav** on Canvas. It contains a number of variables regarding student ratings of instructors /courses (e.g., instructor preparedness, instructor confidence), as well as a number of demographic variables about the instructors (e.g., salary, years of experience) and students (e.g., GPA, student rank). This dataset also has some information about students' political traditionalism and policy preferences. All of the relevant variables have descriptive labels in the SPSS file. This dataset was graciously loaned to me by Jim Sidanius, and was part of research he conducted at the University of Texas at Austin.

Components analysis (20 points)

5. To make life as easy as possible for my beloved students, I want to try to reduce the end-of-quarter survey to as few items as possible.

   A) Help me find a way to capture as much of the variance in the survey as possible by running a components analysis on it (items 13 through 24 in the data file). How many components emerge using the traditional threshold? What would you label them?

   B) How much variance do each of the components explain? Does an orthogonal or oblique solution provide the best interpretability? If oblique, what is the correlation between the components? Plot the items according to your final component solution.

   C) If I were to only ask only one question to measure each component, which question(s) would I ask? How much variance of each of those items is explained by the components together (i.e., the items' total communalities)?

Structural equation modeling (20 points)

6. I would like to further understand the relationship between academic success and traditionalist attitudes. In my model, academic success is a latent factor indicated by overall GPA and expected grade in the class. Traditionalism is a latent factor that is explained by two other latent factors, gender preference and social dominance preference. Gender preference is indicated by items 26, 29, 31, 33, 35, 37, 39, 41, 42, 44, 45, 47, 49, 51, and 52. Social dominance preference is indicated by items 25, 27, 28, 30, 32, 34, 36, 38, 40, 43, 46, 48, and 50.

   A) Draw this factor model (by hand or using your favorite computer tool). It should have 4 latent factors: academic success, traditionalism, gender preference, and social dominance preference. In addition to the structural model (i.e., variances of and covariances among the latent factors), include all details relevant to the measurement model (i.e., the relationship between the indicators and the latent variables and the error terms of the indicators).

   B) Calculate the information necessary to compute this model: the covariance matrix of the indicators. How many unique observations are there? How many parameters need to be estimated by the model? How many degrees of freedom will this model have? Based on the 50:1 *N:q* rule, roughly how many subjects will be required to adequately measure this model?