PSYC 601
Data screening 2024
Dr. Kline

Essential Lessons

This lab assignment is intended to reinforce the following principles:

1. The 80–20 rule for analysis: Preparing the data for analysis takes at least 80% of the time and effort, actually performing the analysis takes no more than 20%.

2. Archival data sets are often in terrible shape due to inadequate data cleaning. Information about archival data may also be incomplete.

3. When, not if, a particular computer tool for statistical analysis does not print exactly what you want, then you must calculate the desired result by hand working with selected quantities in the output (or use a different computer program).

4. When, not if, the graphical user interface of a statistics computer tool does not provide desired output, then you may need to write syntax that generates that output.

Notice

The raw data file for this exercise is new. That is, it is a different data file than used in any other previous lab section of this class. This means that the problems with these data for this year's section are unique.

Data Screening Exercise

This exercise (20% of course grade) involves screening the raw data contained in the comma-separated values (CSV) format file

screen2024.csv

that can be downloaded from Moodle. For the unscreened raw data file, N = 2,589. After downloading the data, check for duplicate cases; next, then generate a unique ID for each case. Finally, determine whether the final sample size differs from 2,589. There is no information about minimum or maximum scores, so you must use judgment about outliers, or extremely high or low scores.

Students can use any computer tool to detect potential problems, including improper coding of missing data, data loss patterns, outliers, extreme collinearity, linearity, and distribution shape, among other issues. There are five substantive questions. For each question, describe all detected problems and recommend remedial steps. If changes are made, then demonstrate that these modifications address the original problem(s). If using significance tests as part of your answer, then state assumptions or limitations.

The length limit for the text of your report is 10 pages with 1.5 line spacing, 1" margins, and

12-point Times Roman font. The title page, tables, figures, or references do not count toward the length limit. **Use a minimum number of tables or figures;** that is, respect the need for conciseness and brevity in scientific communication.

The **four areas of evaluation and their relative contributions** to the overall grade are listed next:

1. Clarity of rationale and actions taken, 25%
2. Detection of problems, 25%
3. Writing quality, 25%
4. Effective and efficient use of tables or graphics, 25%

Questions

1. Multivariate normality. The variables

   mvn1–mvn5

   are five different continuous outcome variables. Evaluate the assumption of multivariate normality and suggest corrective measures, if needed. Also comment on any other striking patterns of distrubutional characteristics or descriptive statistics for these outcome measures.

2. Group comparison. The variable

   group

   is coded as "1" for the control condition and "2" for the treatment condition in a between-subject design. Higher scores on the variable

   dv

   indicate a better result. The scores on the dependent variable should not be transformed. Evaluate whether an independent-samples *t* test is feasible. If not, then suggest an alternative. Describe effect size at the both the group level and the case level (distribution overlap). Create only a single data graphic that shows detail about the information just mentioned. Be creative yet informative in generating your data graphic that illustrates the distinctiveness of the group difference.

3. Missing data. Evaluate missing observations for the variables

   wealth, *alcohol*, *healthprobs*, *career*, *family*, and *support*

   which are total scores from a self-report questionnaire with the scales Family Wealth, Alcohol Use, Health Problems, Career Satisfaction, Family Relationships, and Support. Higher scores indicate, respectively, more family financial assets, more alcohol use, more health problems, greater satisfaction with career, better family relationships, and higher levels of perceived support. Distributions of scores on all five scales are expected to be normal. Report on the extent of missing data and describe patterns of data loss.

For example, what is the extent of missing data? Does loss of information on one variable predict the same on other variables? Are missing data patterns related to the variables

age or *gender* ("1" = men, "2" = women)?

If so, then describe the pattern(s). The goal is to summarize how missing data could affect possibily the generalizability of results that concern these variables.

4. Regression screening. In a planned multiple regression analysis, the outcome variable

   criterion

   will be regressed on the predictors

   predictor1–predictor5

   Screen the data for possible problems, including extreme collinearity, outliers, nonlinearity, or distribution shape. Evaluate assumptions about the residuals. Offer recommendations about how to proceed with the regresssion analysis, given your findings.

5. Repeated-measures screening. The variables

   trial1–trial4

   are repeated-measures variables for the same cases. They will be analyzed in a repeated-measures ANOVA for a single group. Evaluate the assumptions for the analysis. Comment on any other problems apparent in the data from these trials.