

Chapter 3

Basic Statistical Concepts: II. Data Preparation and Screening

To repeat what others have said, requires education;
to challenge it, requires brains.

—Mary Pettibone Poole

Overview

- Data preparation
- Data screening
- Score reliability and validity

Data preparation

- SEM computer programs generally accept as input either raw data files or matrix summaries of the data
- If raw data are submitted, the program will create a matrix and analyze it
- Many kinds of SEM analyses require only matrix summaries
- This allows an analysis to be replicated by those without access to the raw data (i.e., to conduct a secondary analysis)

Data preparation

- It is necessary to input a raw data file when:
 1. Nonnormal data are analyzed with a
 - a. method that assumes normality (e.g., maximum likelihood), but statistics are calculated that correct for nonnormality
 - b. special method that does not assume normality
 2. A special method is used that accommodates cases with missing observations (i.e., the raw data file is incomplete)
- For perhaps most analyses, input of either raw data or a matrix summary is fine

Data preparation

- When means are not analyzed (i.e., the model has just a covariance structure), there are two basic kinds of matrix summaries:
 1. Correlation matrix with standard deviations
 2. Covariance matrix
- Example for the same three variables (Table 3.1):

| Correlations, standard deviations | | | Covariances | | |
|-----------------------------------|--------|-------|-------------|---------|--------|
| 1.000 | | | 38.500 | | |
| .470 | 1.000 | | 42.500 | 212.500 | |
| .601 | .750 | 1.000 | 17.500 | 51.250 | 22.000 |
| 6.204 | 14.577 | 4.690 | | | |

- Most SEM computer programs can “assemble” a covariance matrix given the correlations and standard deviations

Data preparation

- In general, it may be a problem to submit just a correlation matrix for analysis:
 1. The most frequently used estimation method (i.e., maximum likelihood) assumes that the variables are unstandardized
 2. Some results, such as standard errors, can be incorrect if a correlation matrix is analyzed instead of a covariance matrix
 3. Some SEM computer programs generate error messages if the user attempts to analyze a correlation matrix
- There are special methods for correctly analyzing a correlation matrix (chap. 7)

Data preparation

- When means are analyzed (i.e., the model has both covariance and mean structures), matrix summaries must consist of the covariances and means
- Example for the same three variables (Table 3.1):

| Correlations, standard deviations, means | | | Covariances, means | | |
|--|--------|--------|--------------------|---------|--------|
| 1.000 | | | 38.500 | | |
| .470 | 1.000 | | 42.500 | 212.500 | |
| .601 | .750 | 1.000 | 17.500 | 51.250 | 22.000 |
| 6.204 | 14.577 | 4.690 | 11.000 | 60.000 | 25.000 |
| 11.000 | 60.000 | 25.000 | | | |

Data screening

- Raw data should be screened for the following general characteristics:
 1. Multivariate normality
 - a. Univariate normality
 - b. Linearity and homoscedasticity
 2. Missing data
 3. Multicollinearity
 4. Relative variances

Data screening

- Multivariate normality means that
 1. all the univariate distributions are normal
 2. the joint distribution of any pair of the variables is bivariate normal
 3. all bivariate scatterplots are linear and homoscedastic

Data screening

- However, it is often impractical to assess all aspects of multivariate normality
- Fortunately, many instances of multivariate nonnormality are detectable through inspection of univariate distributions
- It may also help to delete cases that are outliers

Data screening

- Univariate normality:
 1. Skew and kurtosis can occur either separately or together in a single variable
 2. Skew: Shape of the distribution is asymmetrical about its mean:
 - a. *Positive*: Most of the scores are below the mean
 - b. *Negative*: Most of the scores are above the mean

Data screening

- Univariate normality:

2. Skew:

- c. Standardized skew index:

$$SSI = \frac{S^3}{(S^2)^{3/2}},$$

$$S^2 = \sum (X - M)^2/N \text{ and } S^3 = \sum (X - M)^3/N$$

SSI = 0 indicates a symmetric distribution (no skew);
otherwise, sign of SSI indicates direction of skew

Data screening

- Univariate normality:

3. Kurtosis: Assuming a unimodal, symmetric distribution and relative to a normal distribution with the same variance (DeCarlo, 1997):

- a. *Positive*: Heavier tails and a higher peak (leptokurtic)

- b. *Negative*: Lighter tails and a lower peak (platykurtic)

Data screening

- Univariate normality:

3. Kurtosis:

- c. Standardized kurtosis index:

$$\text{SKI} = \frac{S^4}{(S^2)^2} - 3.00$$

$$S^2 = \sum (X - M)^2/N \text{ and } S^4 = \sum (X - M)^4/N$$

SKI = 0 in a normal distribution (i.e., no skew);
otherwise, sign of SKI indicates type of kurtosis

Data screening

- Univariate normality:
 4. There are few clear guidelines for interpreting absolute values of SSI or SKI concerning how much skew or kurtosis is too much
 5. Suggested rules-of-thumb for index absolute values:
 - a. $SSI > 3.00$ may indicate a problematic amount of skew
 - b. $SKI > 10.00$ may indicate a problematic amount of kurotosis, and $SKI > 20.00$ even more so

Data screening

- Transformations:

1. One way to deal with univariate nonnormality
2. Original scores are converted with a mathematical operation to new ones that may be more normally distributed
3. Examples:
 - a. Square root ($X^{1/2}$) and inverse functions ($1/X$) for positive skew
 - b. Odd-root functions (e.g., $X^{1/3}$) may bring in outliers from both tails
 - c. Odd-power polynomials (e.g., X^3) may help for negative kurtosis

Data screening

- Transformations:

4. However, some distributions can be so nonnormal that no transformation helps
5. Transformation means that the original scale is lost, which can be a sacrifice if it was meaningful (e.g., survival in years)

Data screening

- Outliers:

1. Dealing with outliers can also address multivariate normality
2. Two types:
 - a. *Univariate*: Extreme score (e.g., > 3 standard deviations away from the mean) on a single variable
 - b. *Multivariate*: A pattern of scores across multiple variables that is extreme, but no individual score may be an univariate outlier
3. Some SEM computer programs (e.g., EQS) identify individual cases that contribute the most to multivariate nonnormality

Data screening

- Outliers:

4. Another method is based the Mahalanobis distance (D) statistic, which measures the distance between a case and the set of group means (*centroids*)
5. D^2 in large samples follows a Pearson chi-square (χ^2) distribution where $df =$ number of variables
6. Cases with D^2 that are statistically significant at $p < .001$ may be multivariate outliers—see Stevens (2002)

Data screening

- Missing data:

1. Recommended works about missing data include

- a. books by Allison (2001) and Little and Rubin (2002)
- b. a special issue of *Psychological Methods* (West, 2001)
- c. articles by Allison (2003) and Peters and Enders (2002) about missing data techniques for SEM

Data screening

- Missing data:
 2. There are two types of *ignorable* data loss patterns:
 - a. *Missing at random (MAR)*: Missing observations on X differ from observed scores only by chance
 - b. *Missing completely at random (MCAR)*: MAR and data loss pattern on X is unrelated to any other variable
 3. MCAR is just a stronger assumption about the randomness of data loss than MAR, but it is doubtful whether MCAR holds in actual data sets

Data screening

- Missing data:

4. Categories of methods (Vriens & Melton, 2002):

- a. *Available case methods*: Take little advantage of structure in the data (i.e., they are post hoc), generally assume MCAR, deletes cases with incomplete records
- b. *Single imputation methods*: Use more information about the data, also generally assume MCAR, replace a missing observation with a single imputed value

Data screening

- Missing data:

4. Categories of methods (Vriens & Melton, 2002):

- c. *Model-based imputation methods*: Replace a missing score with one or more imputed values from a predictive distribution that models the underlying data loss mechanism

- d. There are also some special multivariate estimation methods that do not delete cases or impute scores, and they generally assume MAR instead of MCAR

5. The latter methods (4d) in SEM are usually special forms of maximum likelihood estimation for incomplete data sets

Data screening

- Missing data:

6. Available case methods for missing data:

- a. *Listwise deletion*: Cases with missing scores on any variable in the analysis are deleted, so the effective sample size is the same for all analyses
- b. *Pairwise deletion*: Cases are deleted only if they have missing data on variables involved in a particular computation, so the effective sample size can vary

Data screening

- Missing data:

7. The latter property of pairwise deletion (6b) can result in a *nonpositive definite (singular)* correlation or covariance matrix
8. Such a matrix has at least one *out-of-bounds element*, which means that its value would be impossible in a data set with no missing data
9. A singular data matrix cannot generally be analyzed by the computer—thus, pairwise deletion is not recommended

Data screening

- Missing data:

10. Single imputation methods for missing data:

- a. *Regression-based*: A missing observation is replaced with a predicted score generated by using multiple regression based on nonmissing scores on other variables—simple, but may distort the underlying distribution
- b. *Pattern-matching*: Replaces a missing observation with a score from another case with a similar profile of scores across other variables—available in PRELIS of LISREL
- c. *Random hot-deck*: Separately sorts complete and incomplete cases based on background variables, randomly interleaves the records, and then replaces a missing score with one from the nearest complete record

Data screening

- Missing data:

11. A drawback to all these single-imputation methods is that if the proportion of missing observations is relatively high, then error variance may be substantially underestimated due to imputing just a single value (Vriens & Melton, 2002)

Data screening

- Multicollinearity:

1. Another cause of singular correlation or covariance matrices
2. Occurs when intercorrelations among some variables are so high (e.g., $> .85$) that certain mathematical operations are either impossible or unstable because some denominators are close to zero
3. Researchers can also inadvertently cause multicollinearity when both a composite and its constituent variables (e.g., a total score and subscores) are included in the same analysis

Data screening

- Multicollinearity:

4. Some measures of multicollinearity at a multivariate level:

a. Squared multiple correlation (R_{smc}^2) between each variable and all the rest

b. Tolerance, or $1 - R_{\text{smc}}^2$

c. Variance inflation factor (VIF), or $1/(1 - R_{\text{smc}}^2)$

5. $R_{\text{smc}}^2 > .90$, tolerance $< .10$, or VIF > 10.00 suggests multicollinearity

Data screening

- Relative variances:

1. Covariance matrices where the ratio of the largest to the smallest variance is greater than, say, 10.00, are known as *ill-scaled*
2. An ill-scaled covariance matrix can cause problems in the analysis
3. The most widely used estimation methods in SEM are *iterative*, which means that initial estimates are derived and then modified through subsequent cycles of calculations
4. The goal of iterative is to find the optimal solution for a particular sample

Data screening

- Relative variances:

5. Iterative estimation stops when the improvement in the solution from step-to-step becomes very small (i.e., estimation has converged)
6. When the computer adjusts the estimates from step-to-step, the sizes of these changes may be huge for some variables but trivial for others in an ill-scaled covariance matrix, so the whole analysis may fail
7. To prevent this problem, variables with extremely low or high variances can be rescaled by multiplying their scores by a constant
8. This changes the variance by a factor that equals the squared constant

Score reliability and validity

- Scores from variables analyzed in SEM should be both reliable and valid
- Reliability and validity are attributes of scores in a particular sample, not of measures (B. Thompson, 2003)
- Measurement theory has not been emphasized as strongly as it should be in psychology curricula (e.g., Frederich, Buday, & Kerr, 2000), but SEM requires strong knowledge in this area
- See Bryant (2000), Strube (2000), and Nunnally and Bernstein (1994) for more information about measurement theory

Score reliability and validity

- Score reliability:

1. Reliability concerns the degree to which the scores are free from random measurement error
2. A reliability coefficient r_{XX} estimates the proportion of total variance not due to random error
3. Cronbach's coefficient alpha (α) is the most commonly reported estimate of reliability
4. It measures *internal consistency reliability*, the degree to which responses are consistent across the items within a single measure

Score reliability and validity

- Score reliability:

5. Other types of reliability coefficients:

- a. *Test-retest*: degree to which scores are stable over time

- b. *Alternate-forms*: degree to which scores are stable across different versions of the same test

- c. *Interrater*: degree to which scores are subject to examiner-specific factors

6. In general, reliability coefficients about .90 are “excellent,” about .80 are “very good,” and about .70 are “adequate” for research purposes

Score reliability and validity

- Score validity:

1. Concerns whether the scores measure what are they supposed to measure, but also not measure what they are not supposed to measure (B. Thompson, 2003)
2. Reliability is a requirement for validity (i.e., a necessary condition)
3. Most general form is *construct validity*, which concerns whether the scores measure a particular hypothetical construct

Score reliability and validity

- Score validity:
 4. Construct validity is not established in a single study
 5. Usually requires the application of a set of different research methods such as both true experiments and nonexperimental studies
 6. There are also standards for establishing construct validity (e.g., American Psychological Association, 1999)

Score reliability and validity

- Score validity:

7. Some facets of construct validity:

- a. *Content validity*: whether test items are representative of the domain it is supposed to measure
- b. *Criterion-related validity*: whether a measure X relates to an external standard (criterion) Y against which the measure can be evaluated

Score reliability and validity

- Score validity:

7. Some facets of construct validity:

- c. *Convergent validity*: whether correlations among variables believed to measure the same construct are at least moderate in magnitude

- d. *Discriminant validity*: whether correlations among variables believed to measure different constructs are sufficiently low

8. The technique of confirmatory factor analysis (CFA) is often used to evaluate convergent and discriminant validity

Score reliability and validity

- Score validity:

9. Validity coefficients are often designated as r_{XY}

10. Unreliability in the scores of either X or Y attenuates their correlation:

Theoretical maximum absolute value of $r_{XY} = \sqrt{r_{XX} r_{YY}}$

Score reliability and validity

- Score validity:

11. The *correction for attenuation* generates an estimated validity coefficient assuming that the scores on both X and Y are perfectly reliable:

$$\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX} r_{YY}}}$$

12. Perhaps a better way to take reliability into account is through the use of multiple measures of each construct

References

- Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*, 545-557.
- American Psychological Association. (1999). *Standards for educational and psychological testing* (revised ed.). Washington, D.C.: Author.
- Bryant, F. B. (2000). Assessing the validity of measurement. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 99-146). Washington, DC: American Psychological Association.
- DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292-307.
- Frederich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology, 27*, 248-257.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Peters, C. L. O., & Enders, C. (2002). A primer for the estimation of structural equation models in the presence of missing data. *Journal of Targeting, Measurement and Analysis for Marketing, 11*, 81-95.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.
- Vriens, M., & Melton, E. (2002). Managing missing data. *Marketing Research, 14*, 12-17.
- Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.
- West, S. G. (2001). New approaches to missing data in psychological research [Special section]. *Psychological Methods, 6*(4).