

Normality assumptions

If observed variables $\mathbf{x} = (x_1, x_2, \dots, x_p)$ are modeled as continuous variables (i.e. not categorical), the assumption used for SEM parameter estimation via maximum likelihood is that \mathbf{x} is multivariate normally distributed.

If a set of variables is distributed as multivariate normal, then each variable must be normally distributed. However, when all individual variables are normally distributed, the set of variables may not be distributed as multivariate normal. Hence, testing each variable only for univariate normality is not sufficient. Mardia (1974) proposed tests of multivariate normality based on sample measures of multivariate skewness and kurtosis.

AMOS will test the individual variables for normality, it will also provide a test for Mardia's multivariate kurtosis (there's a SAS Macro that does the multivariate test too).

Even though it is not sufficient to conclude multivariate normality, often researchers report that each of the univariate distributions have skew and kurtosis within reasonable ranges. Different rules of thumb have been given (based on various simulation studies for factor analysis type models)...Kline states that skew > 3 and kurtosis > 10 are "extreme", otherwise probably ok.

Normality assumption

- Note if an observed variable is purely exogenous, it does not need to be normally distributed.
- Departures from normality cause the chi-square test to be larger than it should be and standard errors to be smaller than they should be
- If n is very large (i.e. rule of thumb $n > 1000$), do not need to worry about non-normality. Amemiya and Anderson (1988, 1990) show asymptotic robustness of the normality assumption for linear factor analysis model.

Sample size

If someone is talking about sample size they should also be talking about power. The larger the sample, the more the power there is for testing certain parameters and thus the smaller the type 2 errors, i.e. not rejecting when you should. As said by David Howell (<http://www.uvm.edu/~dhowell/gradstat/psych341/lectures/MultipleRegression/multreg3.html>) “While you can’t talk about power without talking about sample size, on the other hand, much has been said about sample size without putting it in the context of power. For example, there is a long-standing, and most likely incorrect, rule of thumb that says that you need at least 10 observations per variable in a multiple regression. This appears to be saying that your multiple regression will be invalid if you don’t meet that rule, but really it is a rule about power. It is really saying that you don’t have much of a chance of finding a significant relationship unless your n is that large, which is quite different from saying that your regression won’t be legitimate.”

Sample size

Nevertheless many other rules of thumb exist in SEM analysis that go from as low as five cases per parameter estimate in the SEM analyses up to 15 per parameter estimate. Notice that this is cases **per parameter estimate** rather than per measured variable.

Many simulation studies done, comparing different rules of thumb for different kinds of confirmatory factor models, some recommendations end up saying investigator should plan on collecting at least 100 cases, with 200 being better.

Consequences of using smaller samples include more convergence failures, improper solutions (including negative error variance estimates for measured variables), and lowered accuracy of parameter estimates, in particular, standard errors SEM program standard errors are based on asymptotic results.

In reality there is no rule of thumb that applies to all situations... Muthen and Muthen (2002) How to Use a Monte Carlo Study to Decide on Sample size and determine power, *Structural Equation Modeling*, 9(4), 599-620.

Missing data

EXAMPLE for thinking about different kinds of missing data.

Survey for a large medical organization to investigate violence in the workplace.

Employees were sent a survey at their homes and were asked many things, including whether they had been the victim of physical violence within the last year.

Of the 4000 surveys sent, the response rate was 55%

Different kinds of missing - are they reasonable for this data?

MCAR - Missing completely at random

MAR - Missing at random

Non-ignorable

Types of missing data

- MCAR - There is no relation between the reason the data is missing and what the persons response to the question about violence would be.
 - could be reasonable if data missing because of incorrect mailing addresses.
- MAR - There might be a relation between the reason the data is missing and what the persons response would have been but this relation goes away if we adjust for observed covariates.
 - Assume there is a covariate related to the probability of responding, e.g. nurses are more likely to respond to questionnaires than doctors. If nurses have higher rates of violence then the data are not MCAR because there is some relation between responding and experience violence. BUT, if this relation exists only because nurses tend to respond more, then since we can adjust for whether the person is a nurse or not, the data are MAR. In other words, within observed covariate subgroups, the data are MCAR.
- Non-ignorable - The reason the data is missing is directly related to what the persons response would have been.
 - persons who have experienced violence are more likely to respond to the questionnaire because they have some feeling of wanting to get the message out about violence.

Ways of dealing with missing data

- *listwise deletion* - if a record is missing on any one variable then throw it out
- *pairwise deletion*, for bivariate correlations, compute statistics based upon the available pairwise data
- *mean substitution* - plug in the mean value for the variable in each case where it is missing
- *regression methods* - develop a regression equation based on complete cases for each variable and then predict the missing values using the regressions
- *Hot deck imputation* - identify the cases with the most similar matching of observed variables to the case which has a missing value and then substitute the non-missing values from the case with none missing
- *FIML* (Full information maximum likelihood) - Using all available data to generate maximum likelihood based sufficient statistics (i.e. under multivariate normality assumption this is the vector of means and the covariance matrix)
- *Multiple imputation* - similar to FIML except the that actual raw data values are simulated. Typically multiple data sets are created, then analyzed using usual statistical methods treating the data as if they were complete case data. Then the results are combined into a single summary finding.
- NOTE: both the FIML and multiple imputation can be implemented using the *EM*

Techniques only valid for MAR (or MCAR)

There is no test for MAR. If one suspects ways in which MAR is violated, non-ignorable missing data modeling can be attempted to see if results differ.

Pattern mixture modeling - put the type of missing pattern in as a categorical covariate.

FIML implemented in AMOS and Mplus

- In Mplus when you have missing data you must add the line `missing are .;` in the Variables command, then you also need to add the command `Type = Missing` in the Analysis command. (SEE Mplus manual handout)
- In AMOS when you have missing data you must check the box that says “Estimate means and intercepts” and the “maximum likelihood” box under the Analysis properties Estimation tab.

