METHODS OF QUANTITATIVE DATA ANALYIS
MSR Course, 2011-2012

Harry B.G. Ganzeboom

Lecture 4b: Missing values

May 3 2012

Missing values

- In practical data analysis, missing values are a major concern. They cause many problems / errors and it takes much time to address those.

The kinds of MV

- MCAR: Missing completely at random: the missings are assumed to occur by a simple random mechanism. The valid data are not *biased*, we have just fewer cases.
- MAR: Missing at random: missings arise at random, but given values of other variables in your data.
- MNAR: Missing not at random: missings arise due to some systematic, but unknown mechanism. The valid data are biased.

Design missings

- Before you start any missing values analysis, it is always good to think about the reasons why some cases are not observed.
- E.g. it is quite common to ask conditional or filtered questions. For these “design missings” a ‘logical value can be substituted.
  - E.g. persons who do not work have no (=0) earnings.
- If you have filled in the design missings, you are left with missings that are missing for seemingly no reason at all. This is actually the condition you want.

An aside: Filter questions

- My private but very strongly held opinions:
  - Never ask a filter question, if it can be avoided
  - It is almost always possible to avoid filtering
  - If you still need to filter, think about changing your research aims.
  - Even if you would need to filter, it is best to simply avoid it and leave it to the respondent to say “not applicable”.

Listwise deletion / complete cases

- Most statistical procedures work best / only on complete cases.
- MCAR / MAR: but there may be considerably loss of statistical power. In MAR situations, the data become biased by using only complete cases.
- MNAR: the results are biased if you cannot address the missing values mechanism. But this is a matter of design, not of an analysis.
- Despite its computational attractions, complete case analysis is not generally better.

Strategies

- Simple imputation
- Multiple imputation
- Available case analysis, pairwise deletion of missing data.
- Full information maximum likelihood (FIML).

Simple imputation

- Mean substitution
- Plausible value / regression based substitution.
- Nearest neighbour (‘hotdeck’) substitution

Mean substitution

- The simplest method is to replace all missings by the mean of the respective variables.
- (This can be used by SPSS Factor, when generating factor scores.)
- This strategy is often combined with the use of a 0/1 control variable, that indicates whether substitution was used.

Plausible value substitution

- We can also use more nuanced expected values as substitutes.
- Regression based substitutes regress a variable on a number of (more completely observed) criteria variables, and then use the expected value from the regression.
- This can again be combined with using a control variable that indicates substitution.

Problems with substitution

- One way of the other, you decrease the variance and uncertainty in your data.
- While the explained variance goes up, you are actually cheating.
- Both your coefficients may become biased and you underestimate your standard errors.

Hot-deck / nearest neighbor

- An even more nuanced variety of substitution is not to use some predicted value, but sort the file by the predicted values, and then borrow the value of the nearest neighbor with a valid value.
- This technique was already popular in the time of punch-card sorters.
- Its decisive advantage is that it does not use a value that perfectly fits the model (and reduces variance), but brings in natural (residual) variance in the substitution.

Multiple Imputation

- See Treiman 185-..
- The idea is to to conduct the imputation step several times and then draw (nearest neighbor) substitutes to bring in natural variation.
- Typical number of imputation steps is 5 times (standard in spss).
- You can then use “Rubin’s rule” to generate coefficients and standard errors that are unbiased. SPSS does this automatically.
- However, working with multiple imputation results is rather time consuming and tedious.

My simulation

- The effects of missing values manipulations can best be studied in simulated data with a known missing values mechanism. Applying the various procedures on real data (see Treiman) does not inform which result is right or wrong.
- I generate an elementary causal model X → M → Y, with 30%, 60% and 10% random missings in the three variables. These data are MCAR.

Pairwise deletion

- Some important statistical techniques (linear regression, factor analysis) can be conducted on a correlation / covariance analysis only.
- Hence the SPSS option of *pairwise deletion of missing values*. Each correlation is calculated on the available cases.
- Treiman warns against this, but for reasons of computational difficulties. This is the wrong reason:
  - If so, what would happen with listwise data?
  - If so, what is the nature of multiply imputed data?

FIML / MLMV

- I think, the ideas behing pairwise treatments are actually quite sound: you want to use as much information as you (locally) have.
- You would then expect that your models would have more uncertainty (larger SE) at parts of the models that are estimated from fewer cases.
- SPSS does not do this correctly. However, there are newer estimation techniques around that take different number of observations into account:
  - LISREL – FIML estimation
  - Stata 12 - MLMV estimation.
- I think this is the way to go, certainly when you are willing to assume MCAR.

How does FIML/MLMV work

- Think about you data in missing values patterns
  - 111
  - 101
  - 011
  - 110
- For each of these patterns we have a correlation matrix with 3, 1, 1, 1 informative correlation.
- You think about your model as SEM:
  - For 1 1 1, three equations with three unknowns
  - For the other patterns, three equation with three unknowns.
  - Solve for the unknowns, with some minimization procedure that takes into account the relative N (weight) of each correlation matrix.