Excel Transformations

1. Things from audit sheet

Master

* Row 5: Remove quotations from Row 5 to change from character and fix equation chain.
* BA45: Changed to 0

1. Converted all sheets to values only to preserve figures while cleaning.

Master

* Delete A and B
* Delete 1
* Delete names in A
* Fill in Category names where blank
* Delete BG

Raw Data

* Delete A
* Delete 1, 2
* C2 = “Category”
* D1 = “Indicator”
* D (“Indicator”) moved to A
* D (“Category”) moved to B
* C = “Total”
* C (“Total”) Deleted
* Row 101 moved to 54
* Remove space in Safety, Health, Employment, Education, Public space, Diversity, Film
* Remove extra i in affordability
* Fill in Category Names where blank
* Delete 102:105
* Delete C
* Rename Indicator to be row.names friendly

Attribute Scores

* Delete 1, 2
* Insert to column D
* D = “Region”
* C = “Country”
* Region populated according to legend in A, B
* Delete A, B
* Delete X: AC

Weighted Attribute Scores

* Fill in all Category names where missing
* Delete 1
* Delete all blank columns
* Switch “Indicators” to “Category” (1 to 2)
* A1 = “Country

Safety

* Delete 1, 2

Indicators Attributes Ranking

* Delete Rows 1 and 2

Attribute Scores

* Delete Rows 1 and 2

Imputations

* Provide a complete dataset for analysis
* Hopefully reduce nonresponse bias

<http://www.stat.columbia.edu/~gelman/arm/missing.pdf>

Four “missingness mechanisms”

1. Missingness completely at random – prob of missingness is same for all units (FALSE)
2. Missingness at random (ignorablity assumption)
   1. probability a variable is missing depends only on available information (TRUE)
   2. Regression must include variables that could affect this. Requires that sufficient information has been collected that we can “ignore” the assignment mechanism (assignment to treatment, assignment to response)
3. Missingness that depend on unobserved predictors
4. 4. Missingness that depends on the missing value itself

Methods that discard data – not extremely useful here.

* Complete case analysis
* Available case analysis

General impossibility of proving that data are missing at random

Mean imputation

* Can distort distribution of the variable
  + Leads to underestimates of the standard deviation
  + Distort relationship between variable by “pulling estimates of the correlation toward zero.
* Using information from related observation
  + Useful, and what is currently being done. May overexaggerate similarity to city within the same region and downplay similarities to others. Perhaps similar population is a better predictor than similar region, for example.

<http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf>

In multiple imputation, missing values for any variable are predicted using existing values from other variables. The predicted values, called “imputes”, are substituted for the missing values, resulting in a full data set called an “imputed data set.”

Standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. These analysis results are then combined toproduce one overall analysis.

Multiple imputation accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data.

Maintaining the original variability of the missing data is done by creating imputed values which are based on variables correlated with the missing data and causes of missingness. Uncertainty is accounted for by creating different versions of the missing data and observing the variability between imputed data sets.

It is important to note that imputed values produced from an imputation model are not intended to be “guesses” as to what a particular missing value might be, rather, this modeling is intended to create an imputed data set which maintains the overall variability in the population while preserving relationships with other variables. Thus, in performing multiple imputation, a researcher is interested in preserving important characteristics of the data set as a whole (e.g., means, variances, regression parameters). Creating imputes is merely a mechanism to deliver an analysis which makes use of all possible information.

The performance of multiple imputation in a variety of missing data situations has been well-studied and it has been shown to perform favorably (Graham et al., 1997; Graham & Schafer, 1999; Schafer & Graham, 2002).

Further, multiple imputation has been shown to be robust to departures from normality assumptions and provides adequate results in the presence of low sample size or high rates of missing data.

**Multiple Imputation For Missing Data: What Is It And How Can I Use It?**

Our regression lines will need predictor variables to help preserve relationships in the data.

These variables should be chosen either because they are correlated with the missing variable, the reason for missingness, or both. For example, if the missing variable of interest is a high school achievement test score, variables such as the student’s previous test scorescould be included since they are likely correlated with achievement test scores