Missing Data and Multiple Imputation

or, listwise deletion is evil*

Michele Claibourn, StatLab

September 25, 2013

Research Data Services in the Library

- Research Data Services: `www.library.virginia.edu/services/`
  - Data management consulting
  - GIS training and consultations
  - Locating data, acquiring data, archiving data
  - Technology for capturing data
- StatLab Services: `statlab.library.virginia.edu`
  - Individual consulting: advice, training or feedback on quantitative research
  - Workshops
- Upcoming Events

The Problem

If archaeologists threw away every piece of evidence, every tablet, every piece of pottery that was incomplete, we would have entire cultures that disappeared from the historical record. We would no longer have the Epic of Gilgamesh, or any of the writings of Sappho. It is a ridiculous proposition because we can take all the partial sources, all the information in each fragment, and build them together to reconstruct much of the complete picture without any invention (Honaker and King 2010).

Assumptions about Missingness

Given

$D =$

ID	*Female*	*Age*	*Income*	*DPid*	*BondMarket*
1	1	42	85000	7	5
2	0	32	.	5	.
3	0	27	.	2	1
4	1	72	25000	3	2
5	1	63	56000	4	.

and

$M =$

ID	*Female*	*Age*	*Income*	*DPid*	*BondMarket*
0	0	0	0	0	0
0	0	0	1	0	1
0	0	0	1	0	0
0	0	0	0	0	0
0	0	0	0	0	1

Missingness Assumptions

Assumption	Acronym	M predicted by
Missing completely at random	MCAR	–
Missing at random	MAR	D_{obs}
Not missing at random/ aka Nonignorable	NMAR/ NI	D_{obs} and D_{mis}

Listwise deletion

```
. *"True" model
. reg y x*
                                                      Number of obs =    1000
                                                      R-squared     =  0.0906
                                                      Root MSE      =  7.8544
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95 % Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.090468   .2520478     4.33   0.000      .595861    1.585075
          x2 |   1.161313   .2423481     4.79   0.000     .6857403    1.636886
          x3 |    1.43371    .253291     5.66   0.000     .9366634    1.930756
          x4 |   .4750582   .2402714     1.98   0.048     .0035608    .9465557
          x5 |   1.249664   .2505905     4.99   0.000     .7579174    1.741412
        cons |    .116961   .2486824     0.47   0.638    -.3710417    .6049638
------------------------------------------------------------------------------

. *Set 10 % of each variable to missing (MCAR)
. reg y x*
                                                      Number of obs =     591
                                                      R-squared     =  0.1182
                                                      Root MSE      =  7.8339
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .8706063   .3193665     2.73   0.007     .2433618    1.497851
          x2 |   1.471875   .3175181     4.64   0.000     .8482604    2.095489
          x3 |   1.575078    .331907     4.75   0.000     .9232032    2.226952
          x4 |   .2875535   .3202397     0.90   0.370    -.3414059     .916513
          x5 |   1.681409   .3148632     5.34   0.000     1.063009    2.299809
       _cons |   .2300728   .3237106     0.71   0.478    -.4057036    .8658492
------------------------------------------------------------------------------
```


Listwise deletion

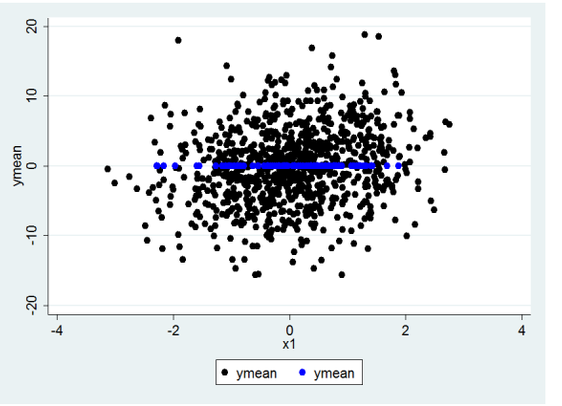
```
. *"True" model
. reg y x*
                                                      Number of obs =    1000
                                                      R-squared     =  0.0906
                                                      Root MSE      =  7.8544
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95 % Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.090468   .2520478     4.33   0.000      .595861    1.585075
          x2 |   1.161313   .2423481     4.79   0.000     .6857403    1.636886
          x3 |    1.43371    .253291     5.66   0.000     .9366634    1.930756
          x4 |   .4750582   .2402714     1.98   0.048     .0035608    .9465557
          x5 |   1.249664   .2505905     4.99   0.000     .7579174    1.741412
        cons |    .116961   .2486824     0.47   0.638    -.3710417    .6049638
------------------------------------------------------------------------------

. *Set x1 missing, dependent on y (MAR)
. reg y x*
                                                      Number of obs =     510
                                                      R-squared     =  0.0211
                                                      Root MSE      =  4.8306
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .4426475   .2148185     2.06   0.040     .0205974    .8646975
          x2 |    .445388   .2074342     2.15   0.032     .0378457    .8529303
          x3 |   .1316879    .228436     0.58   0.565    -.3171163    .5804921
          x4 |   .2224915   .2053555     1.08   0.279    -.1809667    .6259498
          x5 |   .2878971   .2221863     1.30   0.196    -.1486284    .7244226
       _cons |  -6.213315   .2258921   -27.51   0.000    -6.657121   -5.769509
------------------------------------------------------------------------------
```

Traditional Aproaches

- Listwise deletion, aka complete case analysis: exclude cases with missing data for variables used in analysis

  Induces bias, reduced power, model dependency
- Mean imputation: substituting the variable's mean for missing observations

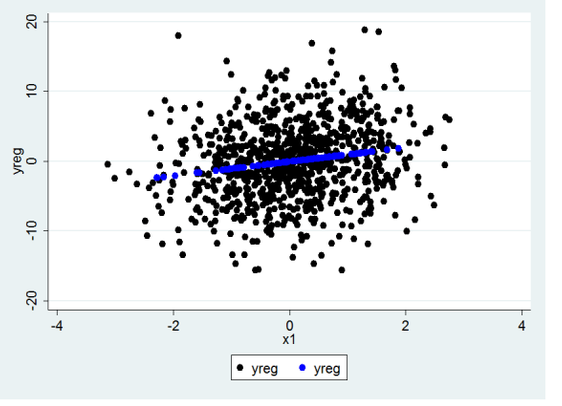


- Conditional mean imputation (regression imputation): subsitute a predicted value from a multivariate model for the observation (typically using the same variables used in themain analysis)

Traditional Aproaches

- Listwise deletion, aka complete case analysis: exclude cases with missing data for variables used in analysis

  Induces bias, reduced power, model dependency
- Mean imputation: substituting the variable's mean for missing observations
- Conditional mean imputation (regression imputation): subsitute a predicted value from a multivariate model for the observation (typically using the same variables used in the main analysis)



Traditional Aproaches, cont.

- Hot deck imputation: substitute the answer from a randomly selected similar unit for the missing value.
- Predictive mean matching: combination of regression and hot deck imputation
- Dummy indicator for missing data: add an extra dummy variable coded 1 for all missing values and 0 otherwise.
- Last value carried forward: longitudinal/repeated measures designs – substitute last observed value for missing value.

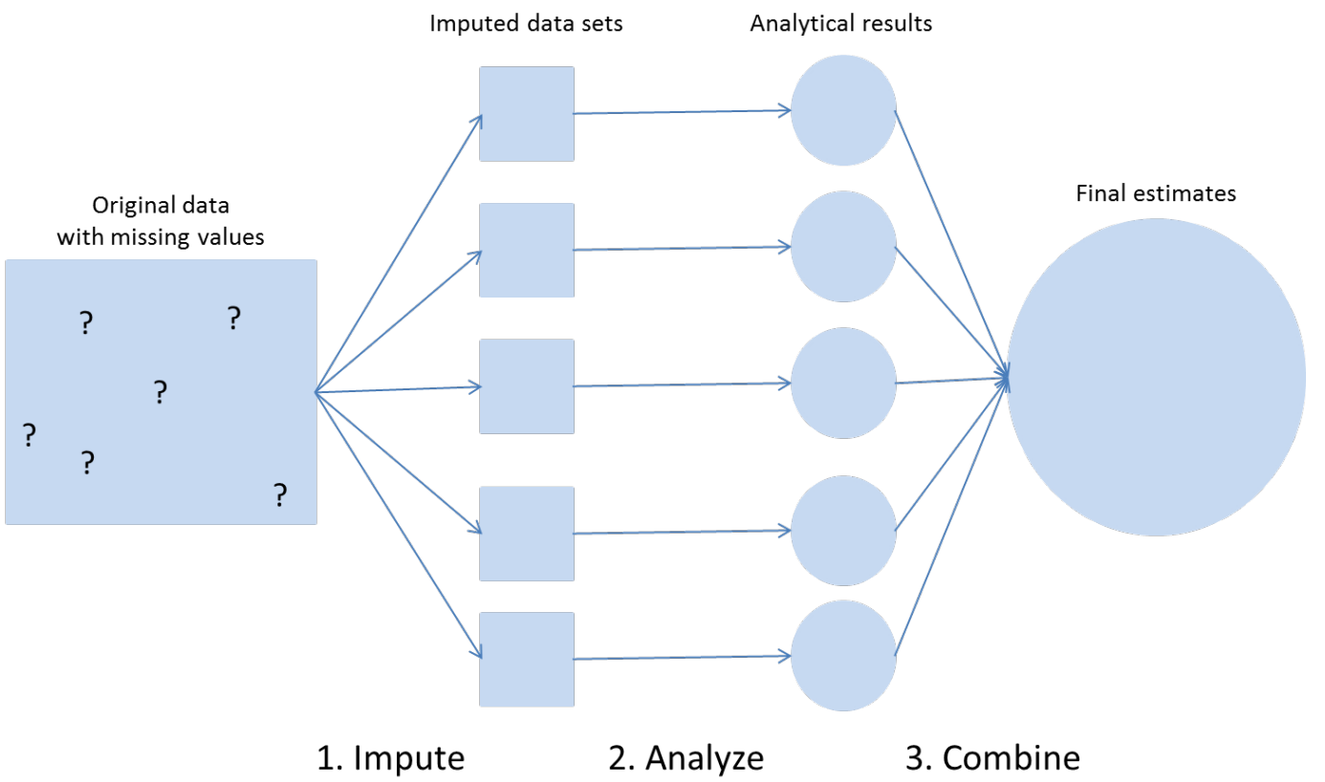
Regression, hot deck, predictive mean matching have some theoretical logic; but all single imputation approaches produce overly precise estimates by ignoring the uncertainty in the imputations.

Patterns of Missingness

- Number of missing – which variables have a lot of missing values?
- Fraction of incomplete cases
- Patterns of missingness – are there groups of subjects with little information available?
- Is Missingness MCAR? Model missingness as function of other variables

Can demonstrate data are not MCAR, but it's generally impossible to distinguish between MAR and NMAR. MI methods assume MAR. Happily, MI is often unbaised with NMAR data (Schafer & Graham 2002).

MI Conceptually



MI Models

MI models rely on iterative methods. Two main approaches:

1. Multivariate Normal (MVN)/Joint Modeling:
   - Stronger theoretical justification, procedures have known statistical properties
   - Assumes all variables have multivariate normal distribution
   - Lacks flexibility
   - Implemented in Stata (`mi impute/mi estimate`), R (`Amelia`), SAS (`proc mi/proc mianalyze`)
2. Imputation by Chained Equations (MICE)/Fully Conditional Specification:
   - Semi-parametric, sepcifies multivariat model by a series of conditional models
   - Flexible, easily applied to a variety of data types
   - Statistical properties are difficult to establish
   - Implemented in Stata (`mi impute chained/mi estimate`), R (`mi, mice`), SAS (`IVEWare`), SPSS (`impute method=fcs`)

MICE

Given 3 variables, X_1 (binary), X_2 (continuous), X_3 (ordinal)

1. Do simple imputations to fill in missing values for X_1, X_2, X_3
2. Using cases with observed X_1, fit logistic regression model of $X_1 \sim X_2 + X_3$; predict missing values of X_1 with expected value pluse draw from the posterior distribution fo the residuals
3. Using cases with observed X_2, fit normal regression model of $X_2 \sim X_1 + X_3$; predict missing values of X_2 with expected value pluse draw from the posterior distribution fo the residuals
4. Using cases with observed X_3, fit proportional odds regression model of $X_3 \sim X_1 + X_2$; predict missing values of X_3 with expected value pluse draw from the posterior distribution fo the residuals
5. Iterate Steps 2-4, cycling through the routine many times so that model converges (hopefully)
6. Repeat to get multiple imputations

Implementing MICE

In general

- How many imputations? 5-10, or approximately equal to fraction of incomplete cases
- What does in the imputation model? Eveything you intend to use in the analysis model, including dependent variable, including interactions, plus auxilliary variables associated with mechanism of missingness.

In Stata

- Set data as `mi` and select shape for imputed data (wide, long)
- Register variables, identify those to be imputed, thoe to be left alone (regular), and those that are determined by other variables (passive)
- Consider running imputation models outside of `mi estimate chained` to ensure they converge, uncover issues[1]

[1] e.g., mlogit can have trouble converging with many categories; mlogit, ologit, logit ocassionally have a covariate that predicts perfectly, and would require the `augment` option in Stata

Checking Imputations

Imputatoin diagnostics are a matter of active research, but at a minimum, you should

1. Check the imputed values:
   - Does the distribution of imputed values seem reasonable given the distribution of observed values?
   - Frequency tables for binary, categorical, ordinal variables
   - Means, standard deviations, shape (histogram, density plot) for continuous variables
   - Check imputations separately, as well
2. Check for convergence
   - The first iteration is often atypical, and because iterations are correlated, subsequent iterations may be atypical as well
   - Stata iterates 10 times before saving an imputed data set; is that enough?
   - Examine the tracefile

Analyzing MI data sets, Pooling estimates

After impuation, peform the analyses you would have performed on the original data on each of the imputed data sets. Then combine the results.[2]

Call the estimated quantity of interest q, in each data set j. To combine estimates across m data sets (Rubin 1987):

- Overall point estimate:

$$\bar{q} = \frac{(1)}{m} \sum_{j=1}^{m} q_j$$

- Overall standard error: If SE_{q_j} is the estimated standard error of the quantity, q_j from data set j, and S_q^2 is the sample varaince across the m point estimates (e.g., $S_q^2 = \sum_{j=1}^{m}(q_j - \bar{q}^2/(m-1))$, then the variance of the MI point estimate is

$$(SE_q)^2 = \frac{1}{m} \sum_{j=1}^{m} (SE_{q_j})^2 + S_q^2(1 + 1/m)$$

[2]Note: Some post-estimation procedures (e.g., goodness-of-fit) are not directly applicable to MI results.

References

Cited works

- Honaker, J. and G. King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data."
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Schafer, J. L., J. W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7: 141-177.

Resources

- Azur, M. J., E. A. Stuart, C. Frangakis, P. J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research* 20: 40-49.
- Lee, K. J., and J. B. Carlin. 2010. "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." *American Journal of Epidemiology* 171: 624-632.
- Little, R. J., D. Rubin. 2002. *Statistical Analysis with Missing Data.* New York: Wiley.
- Rubin, D. B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91: 473-489.
- White, I. R., P. Royston, and A. M. Wood. 2011. "Multiple Imputation using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30: 377-399.