MISSING DATA TECHNIQUES WITH STATA

IDRE Statistical Consulting Group

2.23.16

ROAD MAP FOR TODAY

- To discuss:
  1. Commonly used techniques for handling missing data, focusing on multiple imputation
  2. Issues that could arise when these techniques are used
  3. Implementation of Stata MI Impute command
     - Assuming MVN
     - Assuming ICE/MICE
  4. Imputation Diagnostics

GOALS OF STATISTICAL ANALYSIS WITH MISSING DATA

- Minimize bias
- Maximize use of available information
- Obtain appropriate estimates of uncertainty

THE MISSING DATA MECHANISM DESCRIBES THE PROCESS THAT IS BELIEVED TO HAVE GENERATED THE MISSING VALUES.

1. Missing completely at random (MCAR)
   - Neither the unobserved values of the variable with missing nor the other variables in the dataset predict whether a value will be missing.
   - Example: Planned missingness

2. Missing at random (MAR)
   - Other variables (but not the variable with missing itself) in the dataset can be used to predict missingness.
   - Example: Men may be more likely to decline to answer some questions than women

3. Missing not at random (MNAR)
   - The value of the unobserved variable itself predicts missingness.
   - Example: Individuals with very high incomes are more likely to decline to answer questions about their own income

OUR DATA

- Subset of High School and Beyond
- Sample Size of 200 (Full and MAR)
- 13 Variables
- Student Demographics and Achievement including test scores

ANALYSIS OF FULL DATA

```
. regress read write i.female math ib3.prog

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(5, 194)       =     41.53
       Model |  10814.6553         5  2162.93105   Prob > F        =    0.0000
    Residual |  10104.7647       194  52.0864161   R-squared       =    0.5170
-------------+----------------------------------   Adj R-squared   =    0.5045
       Total |    20919.42       199  105.122714   Root MSE        =    7.2171

------------------------------------------------------------------------------
        read |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       write |   .3747415   .0746281     5.02   0.000     .2275549     .521928
             |
      female |
     female  |   -2.69884   1.095408    -2.46   0.015    -4.859277   -.5384027
        math |   .4418632   .0749972     5.89   0.000     .2939487    .5897778
             |
        prog |
    general  |   .2320562   1.512195     0.15   0.878    -2.750396    3.214509
   academic  |   1.879263   1.423068     1.32   0.188    -.9274069    4.685933
             |
       _cons |   9.623172   3.409797     2.82   0.005     2.898141     16.3482
------------------------------------------------------------------------------
```

COMMON TECHNIQUES FOR DEALING WITH MISSING DATA

1. Complete case analysis (listwise deletion)
2. Mean Imputation
3. Single Imputation
4. Stochastic Imputation

COMPLETE CASE ANALYSIS (LISTWISE DELETION)

- **Method**: Drops entire record with missing data on any variable in the analysis or model
- **Appeal**: Nothing to implement – default method
- **Drawbacks:**
  - Loss of cases/data
  - Biased estimates unless MCAR

MISSING DATA IN SAMPLE

Variable	Obs	Mean	Std. Dev.	Min	Max
read	191	52.28796	10.21072	28	76
write	183	52.95082	9.257773	31	67
female	182	.5549451	.4983428	0	1
math	185	52.8973	9.360837	33	75
prog	182	2.027473	.6927511	1	3

LISTWISE DELETION ANALYSIS DROPS OBSERVATIONS WITH MISSING VALUES

```
. regress read write i.female math ib3.prog

      Source |       SS           df       MS      Number of obs   =       130
-------------+----------------------------------   F(5, 124)       =     23.69
       Model |  5895.48143         5  1179.09629   Prob > F        =    0.0000
    Residual |  6172.12627       124  49.7752118   R-squared       =    0.4885
-------------+----------------------------------   Adj R-squared   =    0.4679
       Total |  12067.6077       129  93.5473465   Root MSE        =    7.0552

------------------------------------------------------------------------------
        read |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       write |   .4410834   .0926477     4.76   0.000     .2577076    .6244592
             |
      female |
     female  |  -2.706338   1.365195    -1.98   0.050     -5.40844   -.0042351
        math |   .3210525   .0951436     3.37   0.001     .1327367    .5093682
             |
        prog |
    general  |   .5177428   1.880833     0.28   0.784    -3.204953    4.240438
   academic  |   1.811155   1.654859     1.09   0.276    -1.464274    5.086585
             |
       _cons |    13.0265   4.123545     3.16   0.002     4.864848    21.18815
------------------------------------------------------------------------------
```

COMPLETE CASE ANALYSIS (LISTWISE DELETION)

	Full	Listwise	Full	Listwise	Full	Listwise
Parameter	β	β	SE	SE	P-value	P-value
Intercept	9.62	13.03	3.410	4.124	0.0053	0.002
Write	0.37	0.44	0.075	0.093	<.0001	<.0001
Female	-2.70	-2.71	1.095	1.365	0.0146	0.0496
Math	0.44	0.32	0.075	0.095	<.0001	0.001
PROG academic	1.88	1.81	1.423	1.655	0.1882	0.2759
PROG general	0.23	0.52	1.512	1.881	0.8782	0.7836

UNCONDITIONAL MEAN IMPUTATION

- **Method**: Replace missing values for a variable with its overall estimated mean
- **Appeal**: Simple and easily implemented
- **Drawbacks**:
  - Artificial reduction in variability b/c imputing values at the mean.
  - Changes the magnitude of correlations between the imputed variables and other variables.

MEAN AND STANDARD DEVIATION BEFORE & AFTER MEAN IMPUTATION

```
. sum female write read math , sep(6)
```

Full

Variable	Obs	Mean	Std. Dev.
female	200	.545	.4992205
write	200	52.775	9.478586
read	200	52.23	10.25294
math	200	52.645	9.368448

Listwise

Variable	Obs	Mean	Std. Dev.
female	182	.5549451	.4983428
write	183	52.95082	9.257773
read	191	52.28796	10.21072
math	185	52.8973	9.360837

Mean Imputation

Variable	Obs	Mean	Std. Dev.
female	200	.5545	.4752727
write	200	52.95075	8.853514
read	200	52.28805	9.97715
math	200	52.8975	9.00113

CORRELATION MATRIX BEFORE & AFTER MEAN IMPUTATION

```
. corr female write read math
(obs=200)
```

Full

	female	write	read	math
female	1.0000			
write	0.2565	1.0000		
read	-0.0531	0.5968	1.0000	
math	-0.0293	0.6174	0.6623	1.0000

Listwise

	female	write	read	math
female	1.0000			
write	0.2415	1.0000		
read	-0.0262	0.6077	1.0000	
math	-0.0628	0.6324	0.6295	1.0000

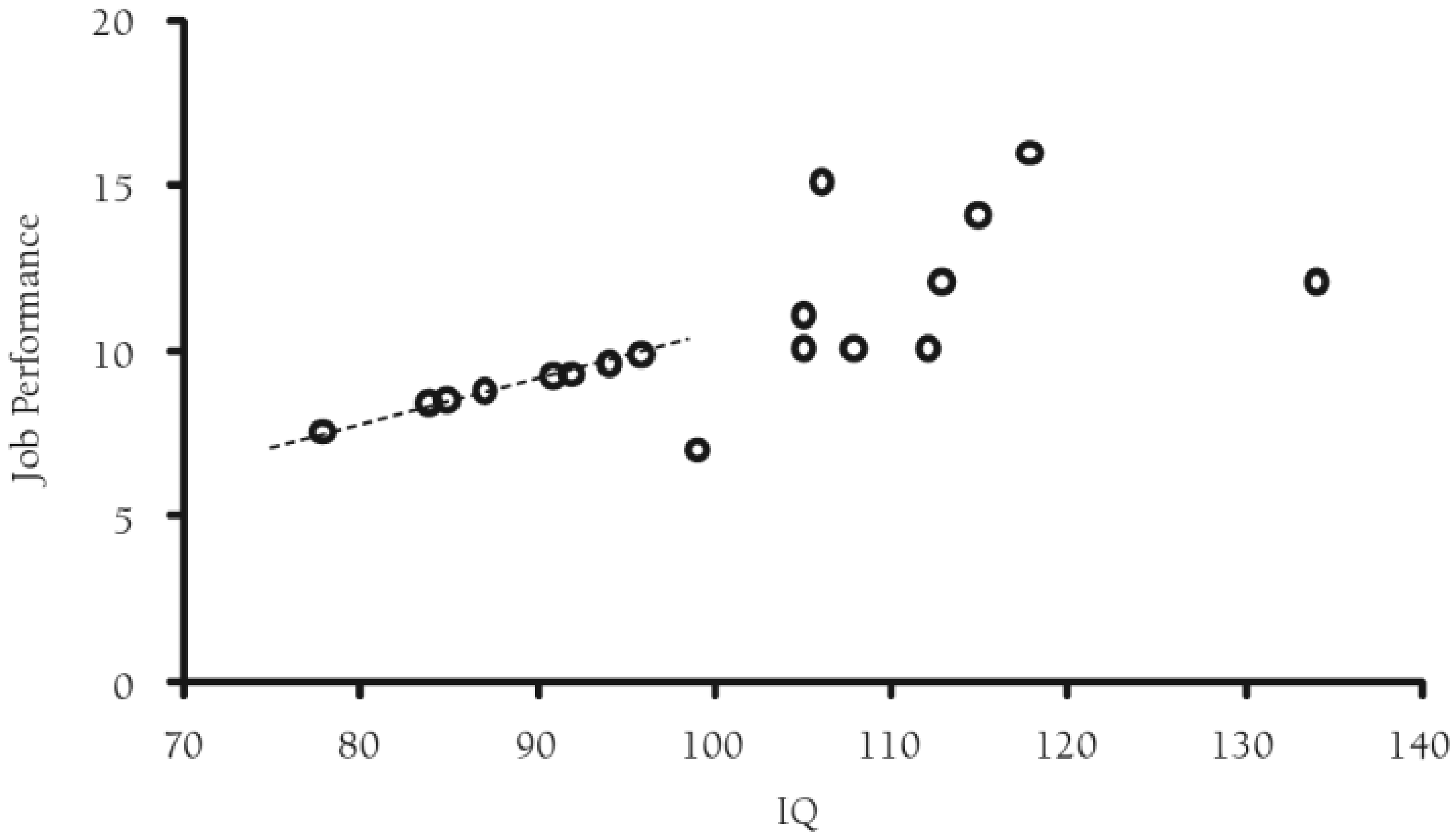
Mean Imputation

	female	write	read	math
female	1.0000			
write	0.2290	1.0000		
read	-0.0146	0.5480	1.0000	
math	-0.0204	0.5491	0.6159	1.0000

SINGLE OR DETERMINISTIC (REGRESSION) IMPUTATION

- **Method**: Replace missing values with predicted scores from a regression equation.
- **Appeal:** Uses complete information to impute values.
- **Drawback:** All predicted values fall directly on the regression line, decreasing variability.

SINGLE OR DETERMINISTIC (REGRESSION) IMPUTATION



p.46, Applied Missing Data Analysis, Craig Enders (2010)

SINGLE OR DETERMINISTIC (REGRESSION) IMPUTATION

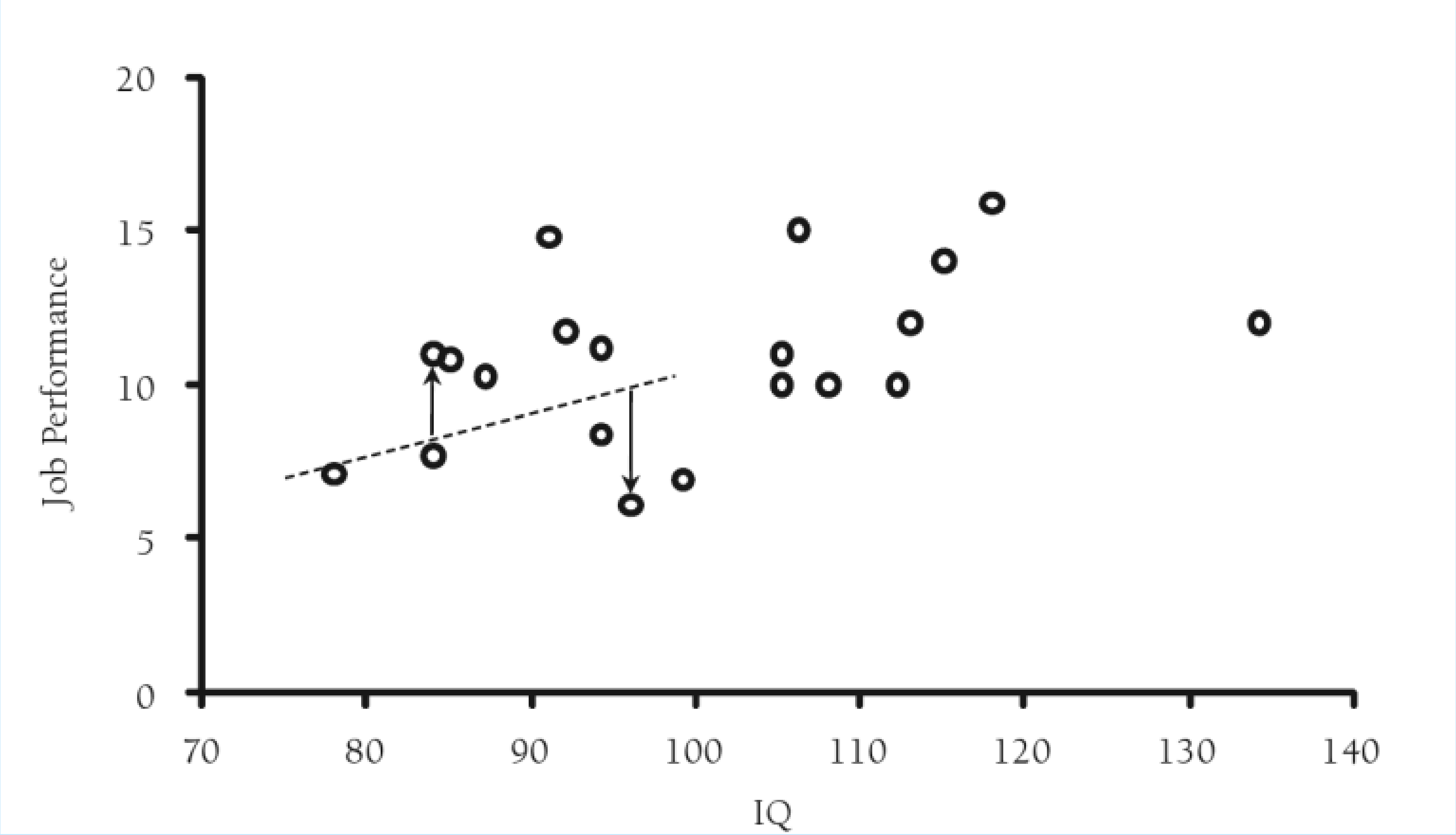
- Imputing values directly on the regression line:
  - Underestimates uncertainty (undeserved precision)
  - Inflates associations between variables because it imputes perfectly correlated values
  - Upwardly biases R-squared statistics, even under the assumption of MCAR

STOCHASTIC IMPUTATION

- Stochastic imputation addresses these problems with regression imputation by incorporating or "adding back" lost variability.

- **Method:** Add randomly drawn residual to imputed value from regression imputation. Distribution of residuals based on residual variance from regression model.

STOCHASTIC IMPUTATION



p.48, Applied Missing Data Analysis, Craig Enders (2010)

STOCHASTIC IMPUTATION

- **Appeals:**
  - Restores some lost variability.
  - Superior to the previous methods as it will produce unbiased coefficient estimates under MAR.

- **Drawback:** SE's produced during stochastic estimation, while less biased, will still be attenuated.

WHAT IS MULTIPLE IMPUTATION?

- Iterative form of stochastic imputation.
- Multiple values are imputed rather than a single value to reflect the uncertainty.
- Each imputed value includes a random component whose magnitude reflects the extent to which other variables in the model cannot predict it's “true “value

- **Common misconception:** imputed values should represent "real" values.

- **Purpose:** To correctly reproduce the variation and associations among the variable that would have present in the full dataset

ISN'T MULTIPLE IMPUTATION JUST MAKING UP DATA?

- No.
- This argument applies to single imputation methods
- MI analysis methods account for the uncertainty/error associated with the imputed values.
- Estimated parameters never depend on a single value.
- Remember imputed values are NOT equivalent to observed values and serve only to help estimate the variances of each variable and covariances/correlations between variables needed for inference

THREE PHASES

- 1. Imputation or Fill-in Phase: Missing values are imputed, forming a complete data set. This process is repeated *m* times.
- 2. Analysis Phase: Each of the *m* complete data sets is then analyzed using a statistical model (e.g. linear regression).
- 3. Pooling Phase: The parameter estimates (e.g. coefficients and standard errors) obtained from each analyzed data set are then combined for inference.

THE IMPORTANCE OF BEING COMPATIBLE

- The imputation model should be "**congenial**“ to or consistent with your analytic model:
  - Includes, at the very least, the same variables as the analytic model.
  - Includes any transformations to variables in the analytic model
    - E.g. logarithmic and squaring transformations, interaction terms
- Why?
  - All relationships between variables should be represented and estimated simultaneously.
- Otherwise, you are imputing values assuming they are uncorrelated with the variables you did not include.

PREPARING FOR MULTIPLE IMPUTATION

1. Examine the number and proportion of missing values among your variables of interest.
2. Examine Missing Data Patterns among your variables of interest.
3. If necessary, identify potential auxiliary variables
4. Determine imputation method

EXAMINE MISSING VALUES: NOTE VARIABLE(S) WITH HIGH PROPORTION OF MISSING – THEY WILL IMPACT MODEL CONVERGENCE THE MOST

mdesc female write read math prog

Variable	Missing	Total	Percent Missing
female	18	200	9.00
write	17	200	8.50
read	9	200	4.50
math	15	200	7.50
prog	18	200	9.00

MI SET

- Stata has a suite of multiple imputation (mi) commands to help user not only impute their data but also explore the missingness in the data.
- To se the entire suite of mi command as well as all the compatible estimation procedures type “help mi”
- In order to use these commands the dataset in memory must be declared or **mi set** as "mi" dataset.
- **mi set mlong**
  - Creates three new **mi** variables including _mi_m (imputation number indicator that ranges from 0 to *m*)

MI STYLES

- A dataset that is **mi set** is given an **mi style**. This tells Stata how the multiply imputed data is to be stored once the imputation has been completed.

- Styles (help mi_styles)
  - Flong
    - Imputed datasets are stacked or appended under original data
    - Includes observations with missing data and those without
  - Mlong
    - Imputed datasets are stacked or appended under original data
    - Includes observations with missing data ONLY
  - Wide
    - Stores imputed value in wide format in stead of long
    - write read write_1 read_1 write_2 read_2
  - Flongsep
    - Stores imputed datasets in different files

MI MISSTABLE PATTERNS

- **mi misstable *patterns* female write read math prog**

```
      Missing-value patterns
        (1 means complete)

              |   Pattern
  Percent     |  1  2  3  4    5
--------------+-------------------
      65%     |  1  1  1  1    1
              |
       8      |  1  1  1  0    1
       8      |  1  1  1  1    0
       7      |  1  1  0  1    1
       6      |  1  0  1  1    1
       5      |  0  1  1  1    1
       1      |  1  0  0  1    1
      <1      |  1  0  1  0    1
      <1      |  1  0  1  1    0
      <1      |  1  1  0  0    1
      <1      |  1  1  0  1    0
      <1      |  1  1  1  0    0
--------------+-------------------
     100%     |

Variables are  (1) read  (2) math  (3) write  (4) female  (5) prog
```

IDENTIFY POTENTIAL AUXILIARY VARIABLES

- Characteristics:
  - Correlated with missing variable (rule of thumb: $r \geq 0.4$)
  - Predictor of missingness
  - Not of analytic interest, so only used in imputation model

- Why? Including auxiliary variables in the imputation model can:
  - Improve the quality of imputed values
  - Increase power, especially with high fraction of missing information (FMI >25%)
  - Be especially important when imputing DV
  - Increase plausibility of MAR

HOW DO YOU IDENTIFY AUXILIARY VARIABLES?

- *A priori* knowledge
- Previous literature
- Identify associations in data

AUXILIARY VARIABLES ARE CORRELATED WITH MISSING VARIABLE

	female	write	read	math	progcat1	progcat2	socst
female	1.0000 182						
write	0.2508 166	1.0000 183					
read	-0.0174 173	0.5872 174	1.0000 191				
math	-0.0241 168	0.6182 170	0.6589 176	1.0000 185			
progcat1	-0.0317 165	-0.0604 166	-0.1058 173	-0.1651 168	1.0000 182		
progcat2	0.0500 165	0.3439 166	0.3902 173	0.4457 168	-0.5635 182	1.0000 182	
socst	0.0889 182	0.5975 183	0.6160 191	0.5451 185	-0.0768 182	0.4096 182	1.0000 200
science	-0.0918 166	0.5498 168	0.6329 176	0.6296 169	0.0567 167	0.2038 167	0.4512 184

AUXILIARY VARIABLES ARE PREDICTORS OF MISSINGNESS

*generate missing data indicator for math

```
generate math_flag=1
replace math_flag=0 if math==.
```

*t-test to determine if mean of science is different between those missing math value and non-missing

```
ttest socst, by(math_flag)
```

AUXILIARY VARIABLES ARE PREDICTORS OF MISSINGNESS

ttest socst, by(math_flag)

```
math_flag

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      15    45.33333    3.080919    11.93235    38.72542    51.94125
       1 |     185    52.97838    .7690379    10.46005    51.46111    54.49564
---------+--------------------------------------------------------------------
combined |     200      52.405    .7591352    10.73579    50.90802    53.90198
---------+--------------------------------------------------------------------
    diff |           -7.645045    2.837886               -13.24141   -2.048684
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.6939
Ho: diff = 0                                     degrees of freedom =      198

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0038         Pr(|T| > |t|) = 0.0077          Pr(T > t) = 0.9962
```

IMPUTATION MODEL EXAMPLE 1:
MI USING MULTIVARIATE NORMAL DISTRIBUTION (MVN)

ASSUMING A JOINT MULTIVARIATE NORMAL DISTRIBUTION

- Probably the most common approach.

- Assumes variables are individually and jointly normally distributed
  - Note: Categorical variables have to be dummied

- Assuming a MVN distribution is robust to violations of normality given a large enough sample size.

- Biased estimates may result when the same size is relatively small and the proportion of missing information is high.

MVN IMPUTATION SYNTAX

- mi set mlong
- mi register imputed female write read math progcat1 progcat2 science
- mi impute mvn female write read math progcat1 progcat2 science = socst, add(10) rseed (53421)
- mi estimate: regress read write female math progcat1 progcat2

IMPUTATION PHASE

- **2 Commands:**
- **Register**
  - **mi register imputed** female write read math progcat1 progcat2 science
  - Identifies which variables in the imputation model have missing information
- **MVN Imputation**
  - **mi impute mvn** female write read math progcat1 progcat2 science = socst, add(10) rseed (53421)
    - The number of imputations is for example only, in practice you may need many more

INCLUDE PICTURE OF STACKED DATA

	id	read	write	math	science	_mi_m
198	198	47	61	51	63	0
199	199	.	59	50	61	0
200	200	.	54	75	.	0
201	1	29.57419	44	40	39	1
202	3	63	65	64.95034	63	1
203	5	47	40	44.36683	45	1

MI IMPUTE OUTPUT

```
Performing EM optimization:
  observed log likelihood = -1601.2096 at iteration 12

Performing MCMC data augmentation ...

Multivariate imputation                      Imputations =         10
Multivariate normal regression                     added =         10
Imputed: m=1 through m=10                        updated =          0

Prior: uniform                                Iterations =       1000
                                                 burn-in =        100
                                                 between =        100
```

Variable	Complete	Incomplete	Imputed	Total
	Observations per *m*			
female	182	18	18	200
write	183	17	17	200
read	191	9	9	200
math	185	15	15	200
progcat1	182	18	18	200
progcat2	182	18	18	200
science	184	16	16	200

(complete + incomplete = total; imputed is the minimum across *m* of the number of filled-in observations.)

ANALYSIS PHASE/POOLING PHASE

- **mi estimate: regress** read write female math science progcat1 progcat2

```
Multiple-imputation estimates                     Imputations       =         10
Linear regression                                 Number of obs     =        200
                                                  Average RVI       =     0.1503
                                                  Largest FMI       =     0.2468
                                                  Complete DF       =        194
DF adjustment:   Small sample                     DF:     min       =      77.11
                                                          avg       =     114.70
                                                          max       =     173.43
Model F test:       Equal FMI                     F(   5,   174.4)  =      35.61
Within VCE type:          OLS                     Prob > F          =     0.0000
```

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
write	.38937	.081702	4.77	0.000	.2278283	.5509116
female	-2.747438	1.143912	-2.40	0.017	-5.005218	-.4896576
math	.4019564	.086768	4.63	0.000	.2294949	.5744179
progcat1	.5163397	1.684931	0.31	0.760	-2.827111	3.85979
progcat2	2.812393	1.602013	1.76	0.083	-.3775475	6.002334
_cons	10.35629	3.686673	2.81	0.006	3.052564	17.66003

COMPARE MIANALYZE ESTIMATES TO ANALYSIS WITH FULL DATA

	Full	Listwise	MVN	Full	Listwise	MVN	Full	Listwise	MVN
Parameter	β	β	β	SE	SE	SE	P-value	P-value	P-value
Intercept	9.62	13.03	10.35	3.410	4.124	3.687	0.0053	0.002	0.006
Write	0.37	0.44	0.39	0.075	0.093	0.082	<.0001	<.0001	<.0001
Female	-2.70	-2.71	-2.74	1.095	1.365	1.144	0.0146	0.0496	0.017
Math	0.44	0.32	0.40	0.075	0.095	0.087	<.0001	0.001	<.0001
PROG academic	1.88	1.81	2.81	1.423	1.655	1.602	0.1882	0.2759	0.083
PROG general	0.23	0.52	0.52	1.512	1.881	1.685	0.8782	0.7836	0.76

DIAGNOSTICS: HOW DO I KNOW IF IT WORKED?

- Compare means and frequencies of observed and imputed values.
  - Use boxplots to compare distributions
  - Note choice of mi set style
- **Look at "Variance Information" table**
- Plots - Assess convergence of imputation algorithm

MI ESTIMATE OUTPUT

```
Multiple-imputation estimates                  Imputations       =         10
Linear regression                              Number of obs     =        200
                                               Average RVI       =     0.1503
                                               Largest FMI       =     0.2468
                                               Complete DF       =        194
DF adjustment:   Small sample                  DF:     min       =      77.11
                                                       avg       =     114.70
                                                       max       =     173.43
Model F test:       Equal FMI                  F(   5,  174.4)   =      35.61
Within VCE type:          OLS                  Prob > F          =     0.0000
```

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
write	.38937	.081702	4.77	0.000	.2278283	.5509116
female	-2.747438	1.143912	-2.40	0.017	-5.005218	-.4896576
math	.4019564	.086768	4.63	0.000	.2294949	.5744179
progcat1	.5163397	1.684931	0.31	0.760	-2.827111	3.85979
progcat2	2.812393	1.602013	1.76	0.083	-.3775475	6.002334
_cons	10.35629	3.686673	2.81	0.006	3.052564	17.66003

MI ESTIMATE OUTPUT

```
Imputations          =        10
Number of obs        =       200
Average RVI          =    0.1503
Largest FMI          =    0.2468
Complete DF          =       194
DF:       min        =     77.11
          avg        =    114.71
          max        =    173.44
F(   5,   174.4)     =     35.62
Prob > F             =    0.0000
```

VARIANCE INFORMATION

- **mi estimate, vartable: regress read write female math progcat1 progcat2**

Variance information

	Imputation variance				
	Within	Between	Total	RVI	FMI
write	.005939	.000669	.006675	.123977	.113855
female	1.24261	.059921	1.30852	.053044	.051507
math	.005947	.001438	.007529	.265958	.219719
progcat1	2.31652	.474974	2.83899	.225541	.191897
progcat2	1.9623	.549235	2.56646	.307883	.246847
_cons	11.4877	1.91258	13.5915	.183139	.160802

VARIANCE: WITHIN (V_W)

- Variability expected with **no missing data**.
- Average of variability of coefficients within an imputation
- Reflects our uncertainty in knowing the “true” coefficient
- This is equivalent to summing the SE^2 for **write** from each of the 10 imputations and then dividing by 10

VARIANCE INFORMATION

- **mi estimate, vartable: regress read write female math progcat1 progcat2**

Variance information

	Imputation variance				
	Within	Between	Total	RVI	FMI
write	.005939	.000669	.006675	.123977	.113855
female	1.24261	.059921	1.30852	.053044	.051507
math	.005947	.001438	.007529	.265958	.219719
progcat1	2.31652	.474974	2.83899	.225541	.191897
progcat2	1.9623	.549235	2.56646	.307883	.246847
_cons	11.4877	1.91258	13.5915	.183139	.160802

VARIANCE: BETWEEN (V_B)

- Variability in estimates across imputations
- Estimates the **additional** variation (uncertainty) that results from missing data.
- Example: Take all 10 of the parameter estimates (β) for **write** and calculate the variance

VARIANCE INFORMATION

- **mi estimate, vartable: regress read write female math progcat1 progcat2**

Variance information

	Imputation variance Within	Between	Total	RVI	FMI
write	.005939	.000669	.006675	.123977	.113855
female	1.24261	.059921	1.30852	.053044	.051507
math	.005947	.001438	.007529	.265958	.219719
progcat1	2.31652	.474974	2.83899	.225541	.191897
progcat2	1.9623	.549235	2.56646	.307883	.246847
_cons	11.4877	1.91258	13.5915	.183139	.160802

TOTAL VARIANCE

- The total variance is sum of 3 sources of variance.
  - Within (V_W)
  - Between (V_B)
  - Additional source of sampling variance.
- $V_T = V_W + V_B + V_B/m$
- Estimated SE = $\sqrt{V_T}$

- What is the sampling variance?
  - V_B/m
  - Sampling error associated with the overall coefficient estimates.
  - Correction factor for using a specific *m*.

VARIANCE INFORMATION

- **mi estimate, vartable: regress read write female math progcat1 progcat2**

Variance information

	Imputation variance Within	Between	Total	RVI	FMI
write	.005939	.000669	.006675	.123977	.113855
female	1.24261	.059921	1.30852	.053044	.051507
math	.005947	.001438	.007529	.265958	.219719
progcat1	2.31652	.474974	2.83899	.225541	.191897
progcat2	1.9623	.549235	2.56646	.307883	.246847
_cons	11.4877	1.91258	13.5915	.183139	.160802

RELATIVE INCREASES IN VARIANCE (RVI)

- Proportional increase in total variance (V_T or SE^2) due to missing information

$$\frac{[V_B + V_B/m]}{V_W}$$

- *Write* RVI = 0.1239
- Variance (V_T or SE^2) is 12.4% larger than it would have been with complete data.

VARIANCE INFORMATION

- **mi estimate, vartable: regress read write female math progcat1 progcat2**

Variance information

	Imputation variance				
	Within	Between	Total	RVI	FMI
write	.005939	.000669	.006675	.123977	.113855
female	1.24261	.059921	1.30852	.053044	.051507
math	.005947	.001438	.007529	.265958	.219719
progcat1	2.31652	.474974	2.83899	.225541	.191897
progcat2	1.9623	.549235	2.56646	.307883	.246847
_cons	11.4877	1.91258	13.5915	.183139	.160802

FRACTION OF MISSING INFORMATION (FMI)

- Directly related to RVI.
- Proportion of total variance (V_T or SE^2) that is due to missing data

$$\frac{[V_B + V_B/m]}{V_T}$$

- *Write* FMI=.1138
- 11.4% of total variance (V_T or SE^2) is attributable to missing data.

VARIANCE INFORMATION

- **mi estimate, vartable: regress read write female math progcat1 progcat2**

Variance information

	Imputation variance				
	Within	Between	Total	RVI	FMI
write	.005939	.000669	.006675	.123977	.113855
female	1.24261	.059921	1.30852	.053044	.051507
math	.005947	.001438	.007529	.265958	.219719
progcat1	2.31652	.474974	2.83899	.225541	.191897
progcat2	1.9623	.549235	2.56646	.307883	.246847
_cons	11.4877	1.91258	13.5915	.183139	.160802

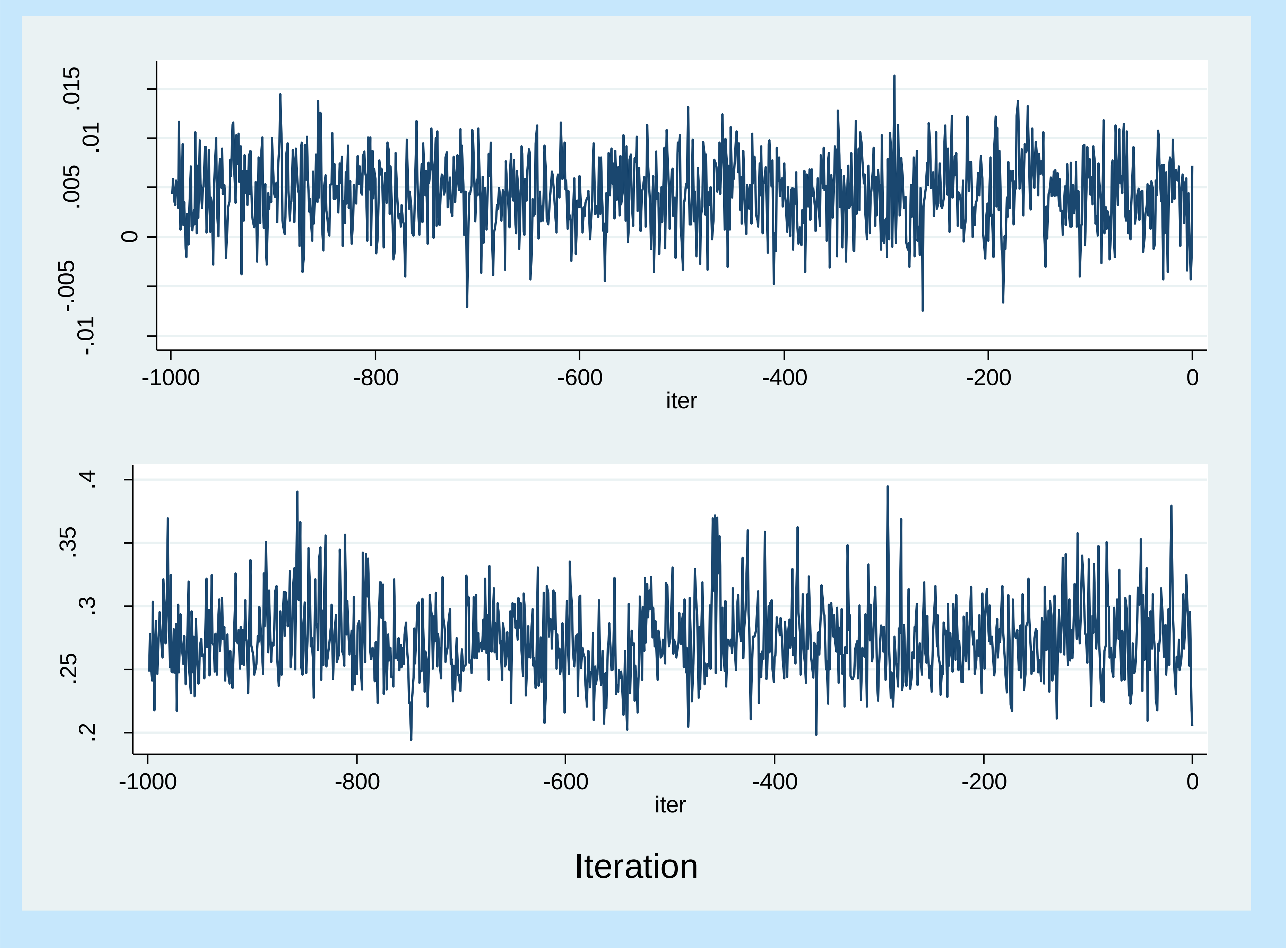
DIAGNOSTICS: HOW DO I KNOW IF IT WORKED?

- Compare means and frequencies of observed and imputed values.
  - Use boxplots to compare distributions
  - Note choice of mi set style
- Look at “Variance Information” table
- **Plots - Assess convergence of imputation algorithm**

TRACE PLOTS: DID MY IMPUTATION MODEL CONVERGE?

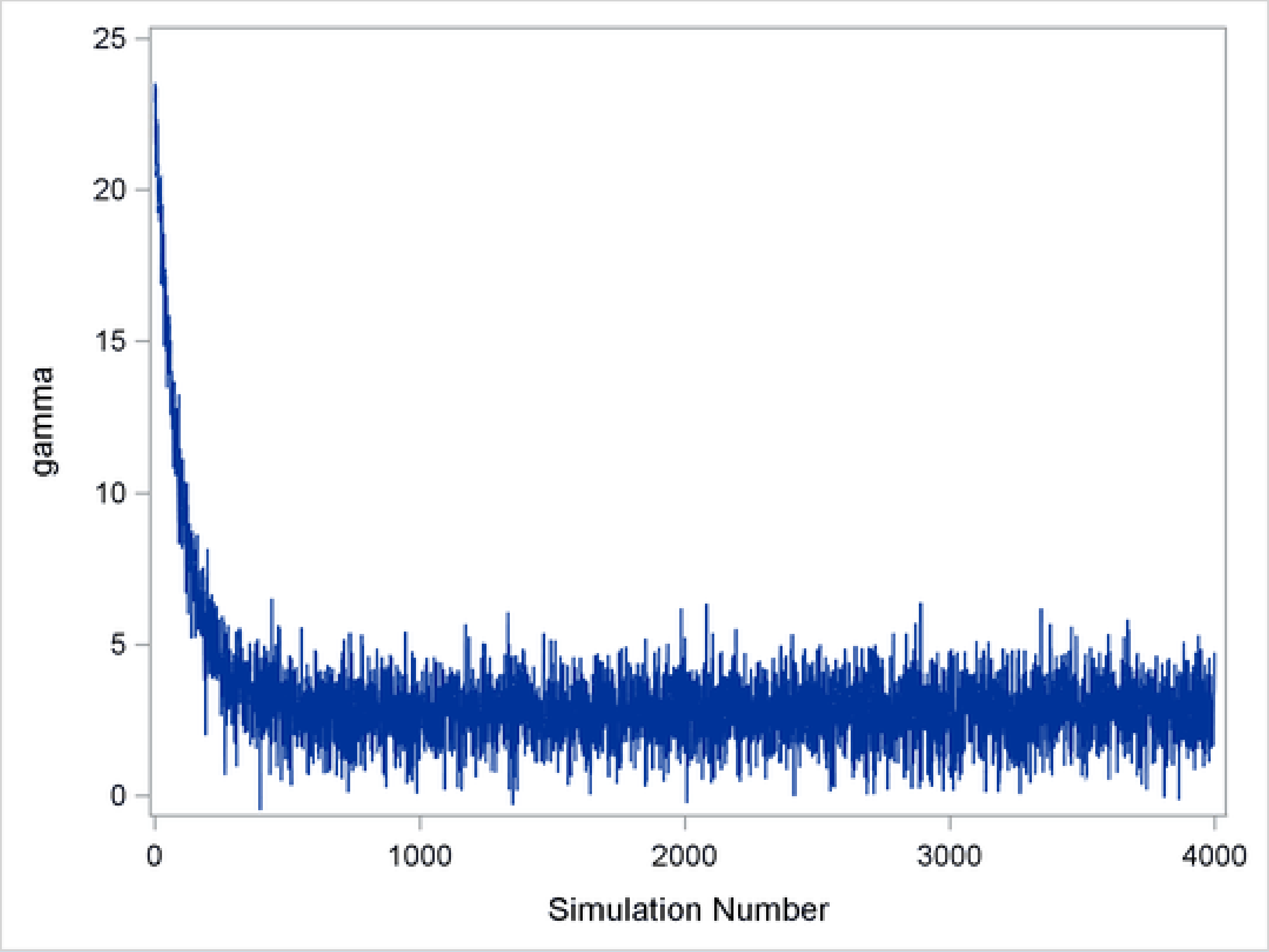
- Convergence for each imputed variable can also be assessed using trace plots.
- Examine plot for each imputed variables
- Special attention to variables with a high FMI

- Option after mi impute mvn
  - **saveptrace(trace, replace)**



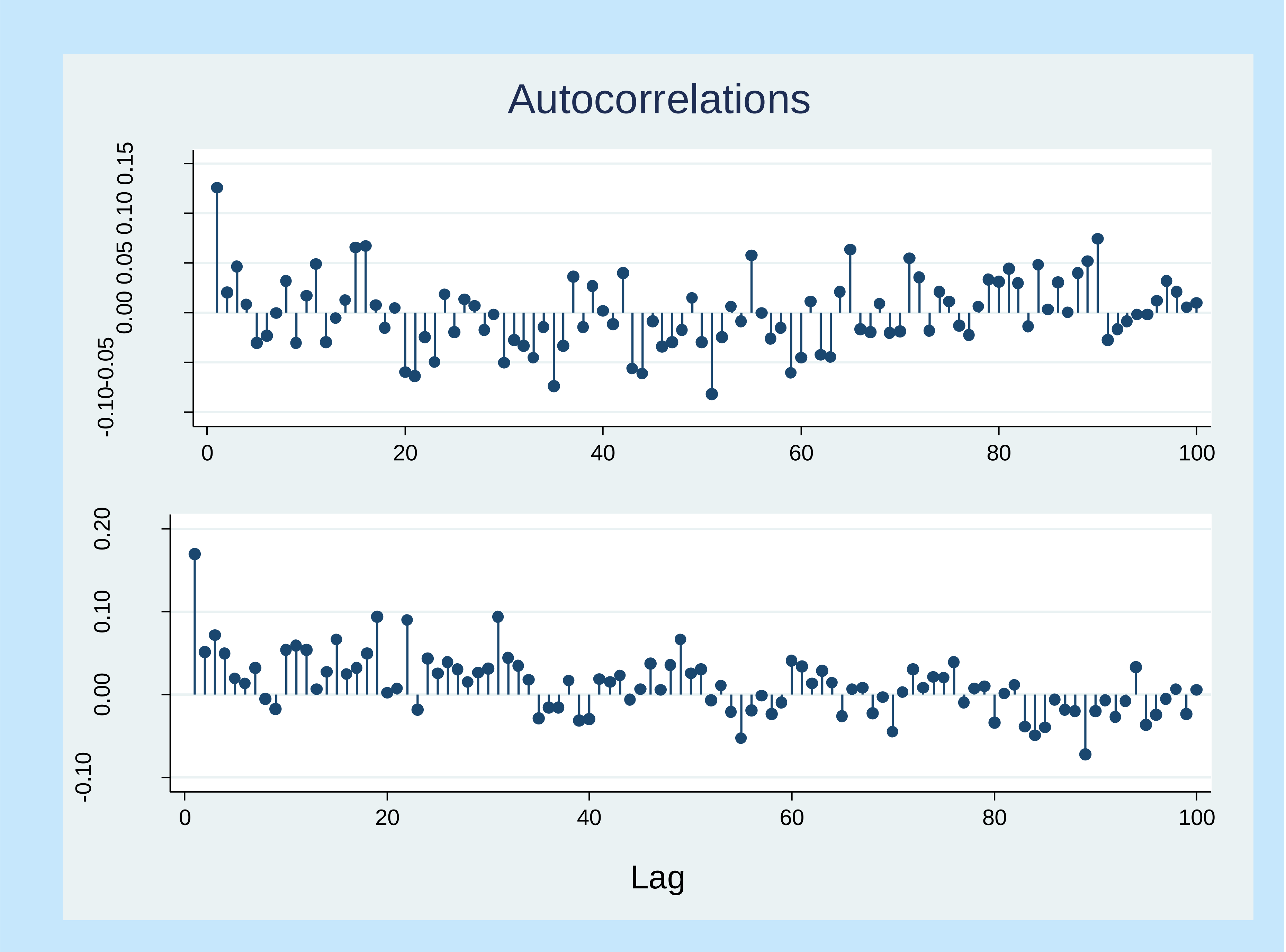
.015
.01
.005
0
-.005
-.01
-1000
-800
-600
-400
-200
0
iter
.4
.35
.3
.25
.2
-1000
-800
-600
-400
-200
0
iter
Iteration

EXAMPLE OF A POOR TRACE PLOT



AUTOCORRELATION PLOTS: DID MY IMPUTATION MODEL CONVERGE?

- Assess possible auto correlation of parameter values between iterations.
- Assess the magnitude of the observed dependency of imputed values across iterations.
- To produce these you will use the **ac** command on the same “trace” file you used to create the Trace plots



Autocorrelations
-0.10
-0.05
0.00
0.05
0.10
0.15
0
20
40
60
80
100
-0.10
0.00
0.10
0.20
Lag

IMPUTATION MODEL EXAMPLE 2:
MI USING IMPUTATION BY CHAINED EQUATIONS

WHAT IF I DON'T WANT TO ASSUME A MULTIVARIATE NORMAL DISTRIBUTION?

- Alternative method is (Multiple) Imputation by Chained Equates (ICE or MICE)
- Does not assume a joint distribution
- Allows different distributions for each variable
- Example uses:
  - Logistic model for binary outcome
  - Poisson model for count variable
  - Other bounded values

AVAILABLE DISTRIBUTIONS

- ICE methods available:
  - Regress (OLS, results similar to MVN)
  - Truncreg (Truncated)
  - Intreg (Interval)
  - Logit (Logistic)
  - Ologit (Ordinal Logistic)
  - Mlogit (Multinomial Logistic)
  - Poisson
  - Nbreg (Negative Binomial)

  - PMM (Predictive Mean Matching)
    - Don’t use Stata’s default knn

CHAINED SYNTAX

- **mi set mlong**
- **mi register imputed** female write read math prog science
- **mi impute chained (logit)** female **(mlogit)** prog **(regress)** write read math science = socst, **add**(10) **rseed** (53421)
- **mi estimate: regress** read write **i.**female math **i.**prog

IMPUTATION PHASE

- Commands are almost the same as the MVN example

- **mi set mlong**
  - The same internal Stata variables are created

- **mi register imputed** female write read math prog science

- **mi impute chained (logit)** female **(mlogit)** prog **(regress)** write read math science = socst, **add**(10) **rseed** (53421)
  - Specify type of distribution to be used for imputation
  - By default, the variables will be imputed in order from the most observed to the least observed

MI ESTIMATE OUTPUT

mi impute chained (logit) female **(mlogit)** prog **(regress)** write read math science =

```
Conditional models:
              read: regress read math science write i.female i.prog socst
              math: regress math read science write i.female i.prog socst
           science: regress science read math write i.female i.prog socst
             write: regress write read math science i.female i.prog socst
            female: logit female read math science write i.prog socst
              prog: mlogit prog read math science write i.female socst
```

ANALYSIS PHASE/POOLING PHASE

- **mi estimate: regress** read write **i.**female math **i.**prog
  - Imputed values for **female** and **prog** will now be true integer values and can be treated as indicator variables

```
. mi estimate: regress read write female math ib3.prog

Multiple-imputation estimates                   Imputations       =         10
Linear regression                               Number of obs     =        200
                                                Average RVI       =     0.1649
                                                Largest FMI       =     0.2121
                                                Complete DF       =        194
DF adjustment:   Small sample                   DF:     min       =      90.00
                                                        avg       =     117.29
                                                        max       =     146.03
Model F test:       Equal FMI                   F(   5,  170.7)   =      35.22
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
        read |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       write |   .4028188   .0827066     4.87   0.000     .2391084    .5665291
      female |  -2.650018   1.201493    -2.21   0.029    -5.026381    -.273656
        math |   .4089138   .0844608     4.84   0.000     .2414949    .5763326
             |
        prog |
    general  |   .0134051   1.710516     0.01   0.994    -3.384835    3.411645
   academic  |   2.341625   1.558824     1.50   0.136      -.75001    5.433259
             |
       _cons |   9.647476      3.617     2.67   0.009     2.499048     16.7959
------------------------------------------------------------------------------
```

PARAMETER ESTIMATES COMPARISON

	Full	Listwise	MVN	ICE	Full	Listwise	MVN	ICE
Parameter	β	β	β	β	SE	SE	SE	SE
Intercept	9.62	13.03	10.35	9.65	3.410	4.124	3.687	3.620
Write	0.37	0.44	0.39	0.40	0.075	0.093	0.082	0.083
Female	-2.70	-2.71	-2.74	-2.65	1.095	1.365	1.144	1.201
Math	0.44	0.32	0.40	0.41	0.075	0.095	0.087	0.084
PROG academic	1.88	1.81	2.81	2.34	1.423	1.655	1.602	1.559
PROG general	0.23	0.52	0.52	0.01	1.512	1.881	1.685	1.711

DIAGNOSTICS:
HOW DO I KNOW IF IT WORKED?

- Compare means and frequencies of observed and imputed values.
  - Use boxplots to compare distributions
  - Note choice of mi set style

- Look at “Variance Information” tables from the proc mianalyze output

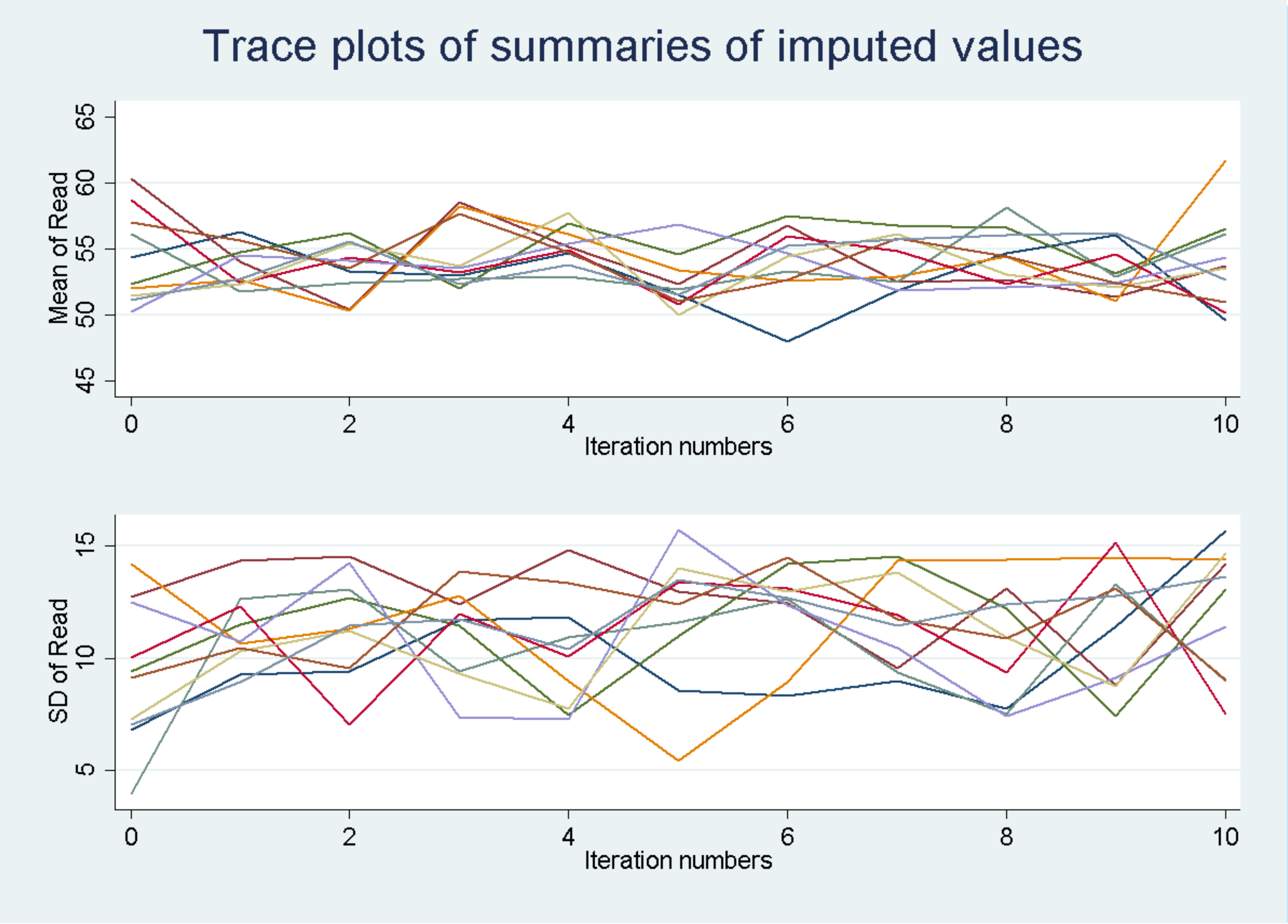
- **Plots - Assess convergence of imputation algorithm**

TRACE PLOTS: DID MY IMPUTATION MODEL CONVERGE?

- **mi impute chained (logit) female (mlogit) prog (regress) write read math science = socst, add(10) rseed (53421) *savetrace*(trace1,replace)**

variable name	storage type	display format	value label	variable label
iter	long	%12.0g		Iteration numbers
m	long	%12.0g		Imputation numbers
read_mean	float	%9.0g		Mean of read
read_sd	float	%9.0g		Std. Dev. of read
math_mean	float	%9.0g		Mean of math
math_sd	float	%9.0g		Std. Dev. of math
science_mean	float	%9.0g		Mean of science
science_sd	float	%9.0g		Std. Dev. of science
write_mean	float	%9.0g		Mean of write
write_sd	float	%9.0g		Std. Dev. of write
female_mean	float	%9.0g		Mean of female
female_sd	float	%9.0g		Std. Dev. of female
prog_mean	float	%9.0g		Mean of prog
prog_sd	float	%9.0g		Std. Dev. of prog

TRACE PLOTS FOR MEAN AND SD OF READ



MICE HAS SEVERAL PROPERTIES THAT MAKE IT AN ATTRACTIVE ALTERNATIVE

1. **MICE** allows each variable to be imputed using its own conditional distribution

2. Different imputation models can be specified for different variables. However, this can also cause estimation problems.

Beware: Convergence issues such as complete and quasi-complete separation (e.g. zero cells) when imputing categorical variables.

COMMON QUESTIONS

- Why do I need auxiliary variables?
- How to determine the number of needed imputations?
- Should I bound imputed values or round to get “plausible” values?
- How do I treat variable transformations such as logs, quadratics and interactions?
- Should I include my dependent variable (DV) in my imputation model?

WHY AUXILIARY VARIABLES?

1. Help improve the likelihood of meeting the MAR assumption
2. Help yield more accurate and stable estimates and thus reduce the estimated SEs in analytic models.
    1. Especially for missing DV's.
3. Help to increase power.

- Bottom line: In general, there is almost always a benefit to adopting a more "inclusive analysis strategy".

SELECTING THE NUMBER OF IMPUTATIONS (*M*)

- **Historical** recommendation was 5
  - Fine when FMI is low and analysis is relatively simple
- **Current** recommendation: As many as 50+ imputations when the proportion of missing data is relatively high

- Why?
  1. Coefficients stabilize at much lower values of <u>*m*</u> than estimates of variances and covariances
  2. Superior RE of estimates
  3. ROT: Multiple highest FMI by 100 and use as approx. number of *m*

- Multiple runs of *m* imputations are recommended to assess the stability of the parameter estimates

MAXIMUM, MINIMUM AND ROUND

- Common issue when using MVN
- Appeal:
  - Makes sense intuitively
- Drawback:
  - Decrease efficiency and increase bias by altering the correlation or covariances
  - Often result in an underestimation of the uncertainty around imputed values
- Bottom line:
  - Imputed values are NOT equivalent to observed values
  - Leaving the imputed values “as is” is perfectly
  - If you need integer or bounded values used MICE

HOW DO I TREAT VARIABLE TRANSFORMATIONS SUCH AS LOGS, QUADRATICS AND INTERACTIONS?

- Treat variable transformations as "just another variable".
  - For example, if your analytic model is interested the modifying effect of Z on the association between X and Y (i.e. an interaction).
  - Properties of your data should be maintained in the resulting imputed values
- Less ideal is passive imputation, X, Z, and Y values are imputed under a model assuming that Z is not a moderator of the association between X an Y.
- Effect modification (e.g. interaction) of interest will be attenuated.

SHOULD I INCLUDE MY DEPENDENT VARIABLE (DV) IN MY IMPUTATION MODEL?

- The answer is ALWAYS yes!
- But opinions differ on how to use the imputed values:
  - Using imputed values of your DV is considered perfectly acceptable with good auxiliary variables
  - There are studies that show imputing DV's when auxiliary variables are not present can add unnecessary random variation into imputed values

MI IN STATA TIPS

Can't Do:

- Multilevel Imputation
  - Some options for 2 level
  - http://www.stata.com/support/faqs/statistics/clustering-and-mi-impute/
- Factor Analysis
- SEM/GSEM

Can Do:

- Multilevel commands
- Survey Data (mi svyset)
- Panel Data (mi xtset)
- Survival Data (mi stset)
- Robust SE's

REFERENCES

- The webpages has almost 30 citations so feel free to use these recourses as a starting off point to your foray into MI.
- A couple recommendations for introductory material:
  - Book
    - Enders (2010). Applied Missing Data Analysis. The Guilford Press.
  - Articles
    - Johnson and Young (2011). Towards Best Practices in analyzing Datasets with Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family,* 73(5): 926-45.
  - Websites:
    - Companion website to “Applied Missing Data Analysis”
    - Social Science Computing Cooperative – University of Wisconsin

BOTTOM LINE

- MI improves over single imputation methods because:
  - Single value never used
  - Appropriate estimates of uncertainty
- Data and model will determine if you choose MVN or ICE
- Several decisions to be made before performing a MI
- MI is not magic, and it should not be expected to provide "significant" effects
- MI is one tool to address a very common problem