MISSING DATA TECHNIQUES WITH SAS

IDRE Statistical Consulting Group

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 SAS Proc MI procedure
     - Assuming MVN
     - Assuming FCS
  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 unobserved value of the variable with missing predicts missingness.
   - Example: Individuals with very high incomes are more likely to decline to answer questions about their own income

OUR DATA

- High School and Beyond
- N=200
- 13 Variables
- Student Demographics and Achievement including test scores

ANALYSIS OF FULL DATA

REGRESSION ON FULL DATA

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10814.65527	2162.93105	41.53	<.0001
Error	194	10104.76473	52.08642		
Corrected Total	199	20919.42000			

R-Square	Coeff Var	Root MSE	read Mean
0.516967	13.81791	7.217092	52.23000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
write	1	1313.358758	1313.358758	25.21	<.0001
female	1	316.174687	316.174687	6.07	0.0146
math	1	1808.048867	1808.048867	34.71	<.0001
prog	2	119.465630	59.732815	1.15	0.3198

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9.623171965	B	3.40979657	2.82	0.0053
write	0.374741452		0.07462808	5.02	<.0001
female female	-2.698839662	B	1.09540766	-2.46	0.0146
female male	0.000000000	B	.	.	.
math	0.441863231		0.07499719	5.89	<.0001
prog academic	1.879263080	B	1.42306759	1.32	0.1882
prog general	0.232056170	B	1.51219473	0.15	0.8782
prog vocation	0.000000000	B	.	.	.

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**: Drop cases with missing data on any variable of interest
- **Appeal**: Nothing to implement – default method
- **Drawbacks:**
  - Loss of cases/data
  - Biased estimates unless MCAR

COMPLETE CASE ANALYSIS (LISTWISE DELETION)

```
proc means data = ats.hsb_mar nmiss N min max mean std;
  var _numeric_ ;
run;
```

The MEANS Procedure

Variable	Label	N Miss	N	Minimum	Maximum	Mean	Std Dev
ID	id	0	200	1.0000000	200.0000000	100.5000000	57.8791845
FEMALE	female	18	182	0	1.0000000	0.5549451	0.4983428
RACE	race	0	200	1.0000000	4.0000000	3.4300000	1.0394722
SES	ses	0	200	1.0000000	3.0000000	2.0550000	0.7242914
SCHTYP	type of school	0	200	1.0000000	2.0000000	1.1600000	0.3675260
PROG	type of program	18	182	1.0000000	3.0000000	2.0274725	0.6927511
READ	reading score	9	191	28.0000000	76.0000000	52.2879581	10.2107174
WRITE	writing score	17	183	31.0000000	67.0000000	52.9508197	9.2577729
MATH	math score	15	185	33.0000000	75.0000000	52.8972973	9.3608367
SCIENCE	science score	16	184	26.0000000	74.0000000	51.3097826	9.8178332
SOCST	social studies score	0	200	26.0000000	71.0000000	52.4050000	10.7357935

LISTWISE DELETION ANALYSIS DROPS OBSERVATIONS WITH MISSING VALUES

REGRESSION ON FULL DATA

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10814.65527	2162.93105	41.53	<.0001
Error	194	10104.76473	52.08642		
Corrected Total	199	20919.42000			

R-Square	Coeff Var	Root MSE	read Mean
0.516967	13.81791	7.217092	52.23000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
write	1	1313.358758	1313.358758	25.21	<.0001
female	1	316.174687	316.174687	6.07	0.0146
math	1	1808.048867	1808.048867	34.71	<.0001
prog	2	119.465630	59.732815	1.15	0.3198

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9.623171965	B	3.40979657	2.82	0.0053
write	0.374741452		0.07462808	5.02	<.0001
female female	-2.698839662	B	1.09540766	-2.46	0.0146
female male	0.000000000	B	.	.	.
math	0.441863231		0.07499719	5.89	<.0001
prog academic	1.879263080	B	1.42306759	1.32	0.1882
prog general	0.232056170	B	1.51219473	0.15	0.8782
prog vocation	0.000000000	B	.	.	.

LISTWISE REGRESSION

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5895.48143	1179.09629	23.69	<.0001
Error	124	6172.12627	49.77521		
Corrected Total	129	12067.60769			

R-Square	Coeff Var	Root MSE	READ Mean
0.488538	13.35231	7.055155	52.83846

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WRITE	1	1128.196639	1128.196639	22.67	<.0001
FEMALE	1	195.608602	195.608602	3.93	0.0496
MATH	1	566.769358	566.769358	11.39	0.0010
PROG	2	68.618278	34.309139	0.69	0.5038

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	13.02649943	B	4.12354544	3.16	0.0020
WRITE	0.44108340		0.09264775	4.76	<.0001
FEMALE female	-2.70633778	B	1.36519467	-1.98	0.0496
FEMALE male	0.00000000	B	.	.	.
MATH	0.32105246		0.09514356	3.37	0.0010
PROG academic	1.81115548	B	1.65485900	1.09	0.2759
PROG general	0.51774275	B	1.88083319	0.28	0.7836
PROG vocation	0.00000000	B	.	.	.

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, SD AND CORRELATION MATRIX OF 5 VARIABLES BEFORE & AFTER MEAN IMPUTATION

BEFORE MEAN IMPUTATION

The CORR Procedure

6 Variables:	WRITE READ FEMALE MATH progcat1 progcat2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
WRITE	183	52.95082	9.25777	9690	31.00000	67.00000	writing score
READ	191	52.28796	10.21072	9987	28.00000	76.00000	reading score
FEMALE	182	0.55495	0.49834	101.00000	0	1.00000	female
MATH	185	52.89730	9.36084	9786	33.00000	75.00000	math score
progcat1	182	0.52198	0.50089	95.00000	0	1.00000	academic
progcat2	182	0.22527	0.41892	41.00000	0	1.00000	general

Pearson Correlation Coefficients
Number of Observations

	WRITE	READ	FEMALE	MATH	progcat1	progcat2
WRITE writing score	1.00000 183	0.58719 174	0.25077 166	0.61825 170	0.34387 166	-0.06036 166
READ reading score	0.58719 174	1.00000 191	-0.01740 173	0.65890 176	0.39023 173	-0.10575 173
FEMALE female	0.25077 166	-0.01740 173	1.00000 182	-0.02408 168	0.05004 165	-0.03169 165
MATH math score	0.61825 170	0.65890 176	-0.02408 168	1.00000 185	0.44566 168	-0.16511 168
progcat1 academic	0.34387 166	0.39023 173	0.05004 165	0.44566 168	1.00000 182	-0.56349 182
progcat2 general	-0.06036 166	-0.10575 173	-0.03169 165	-0.16511 168	-0.56349 182	1.00000 182

AFTER MEAN IMPUTATION

The CORR Procedure

6 Variables:	WRITE READ FEMALE MATH progcat1 progcat2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
WRITE	200	52.95075	8.85351	10590	31.00000	67.00000	writing score
READ	200	52.28805	9.97715	10458	28.00000	76.00000	reading score
FEMALE	200	0.55450	0.47527	110.90000	0	1.00000	female
MATH	200	52.89750	9.00113	10580	33.00000	75.00000	math score
progcat1	200	0.49525	0.48524	99.05000	0	1.00000	academic
progcat2	200	0.25198	0.40849	50.39600	0	1.00000	general

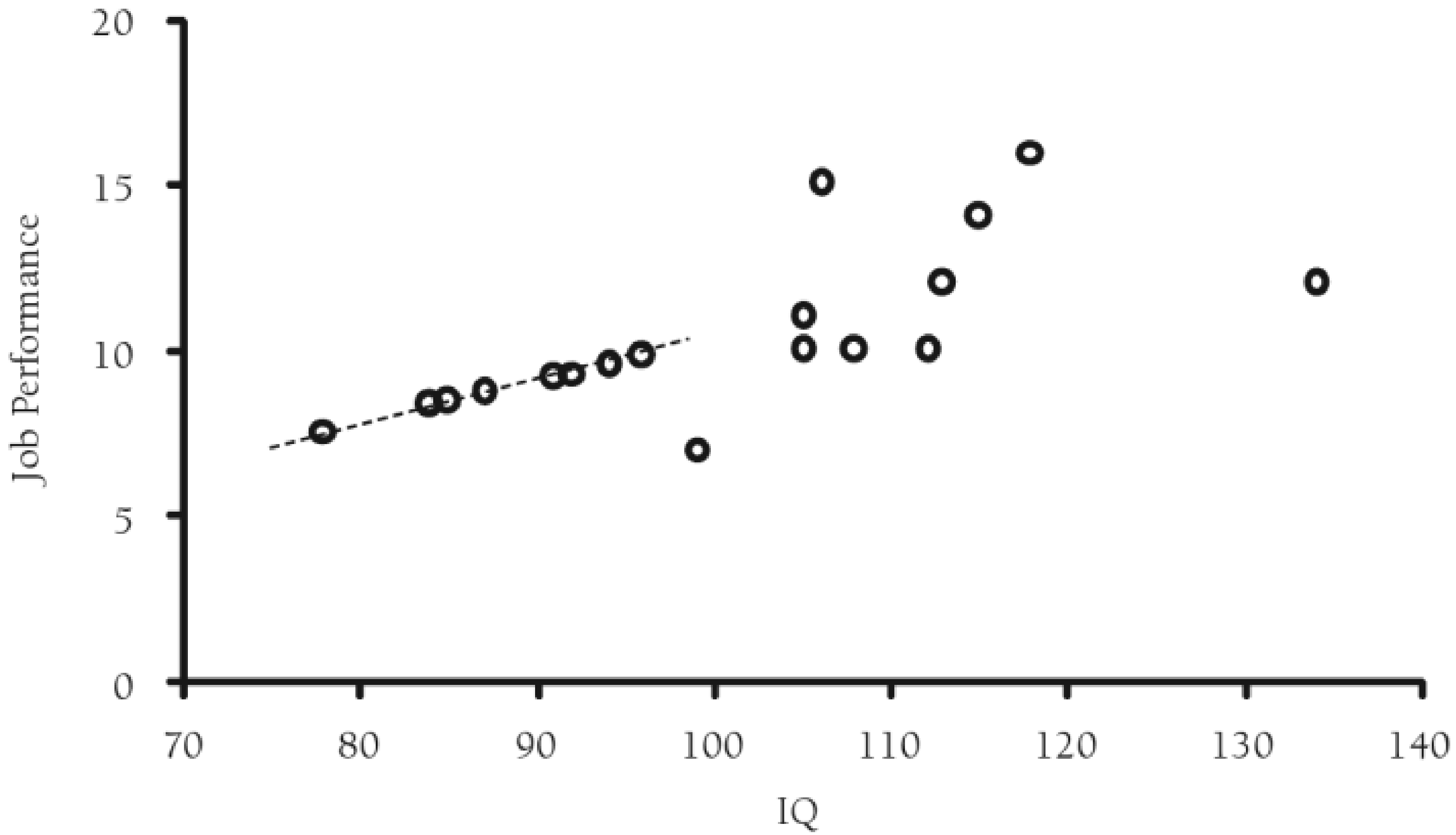
Pearson Correlation Coefficients, N = 200

	WRITE	READ	FEMALE	MATH	progcat1	progcat2
WRITE writing score	1.00000	0.54801	0.22903	0.54914	0.29206	-0.03377
READ reading score	0.54801	1.00000	-0.01461	0.61588	0.35526	-0.09629
FEMALE female	0.22903	-0.01461	1.00000	-0.02037	0.04740	-0.03131
MATH math score	0.54914	0.61588	-0.02037	1.00000	0.38741	-0.12928
progcat1 academic	0.29206	0.35526	0.04740	0.38741	1.00000	-0.57915
progcat2 general	-0.03377	-0.09629	-0.03131	-0.12928	-0.57915	1.00000

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.
- Also known as regression imputation

SINGLE OR DETERMINISTIC (REGRESSION) IMPUTATION



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

SINGLE OR DETERMINISTIC (REGRESSION) IMPUTATION

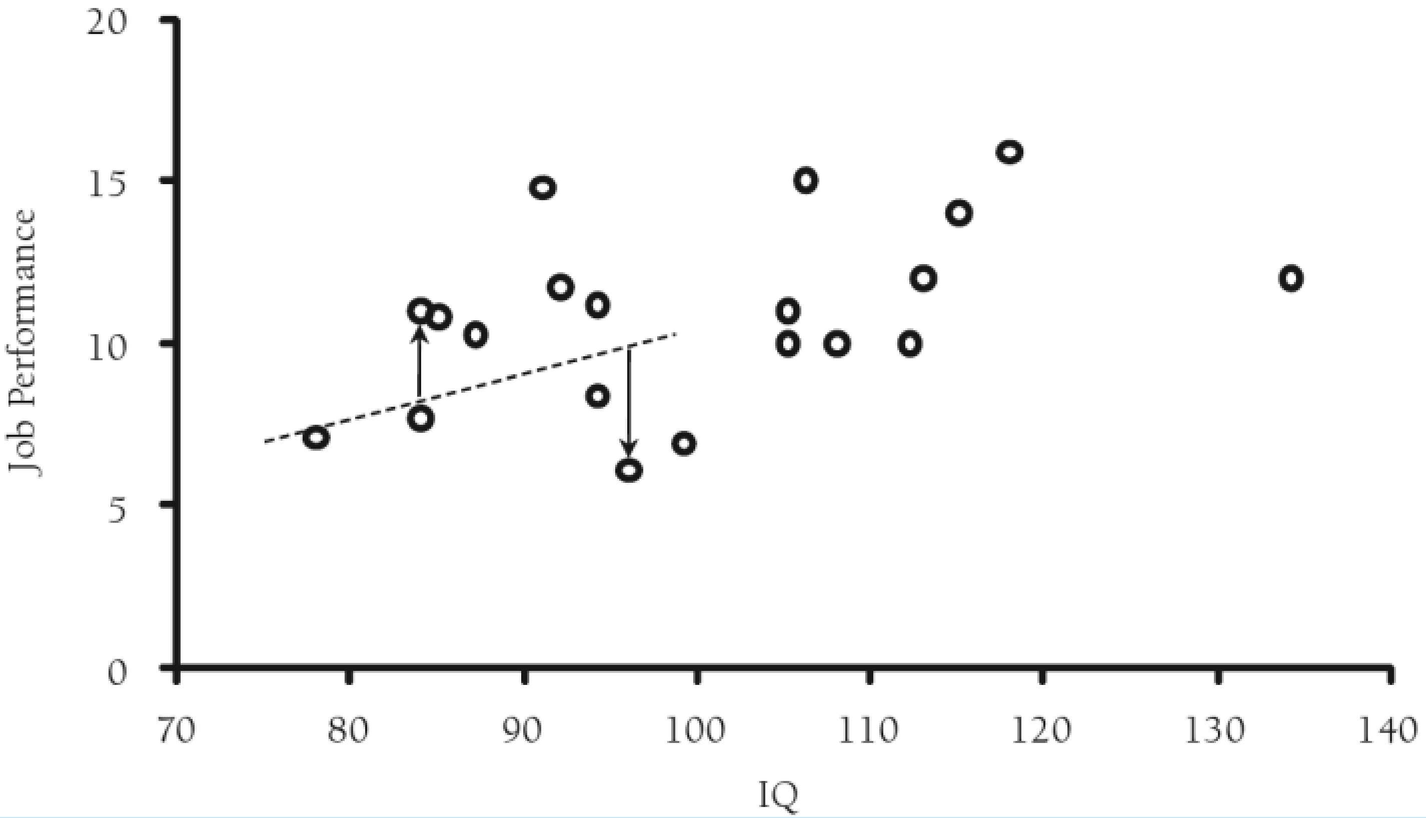
- Imputing values directly on the regression line:
  - Underestimates uncertainty
  - Inflates associations between variables because it imputes perfectly correlated values
  - Upwardly biases R-squared statistics, even under the assumption of MCAR

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 around the “true” value.
- 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 full data variance/covariance matrix

ISN'T MULTIPLE IMPUTATION JUST MAKING UP DATA?

- No.
- This is 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.

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
- 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: PROC MEANS NMISS OPTION

The MEANS Procedure

Variable	Label	N Miss
FEMALE	female	18
WRITE	writing score	17
READ	reading score	9
MATH	math score	15
PROG	type of program	18

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

FEMALE_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	182	91.00	182	91.00
1	18	9.00	200	100.00

WRITE_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	183	91.50	183	91.50
1	17	8.50	200	100.00

READ_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	191	95.50	191	95.50
1	9	4.50	200	100.00

MATH_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	185	92.50	185	92.50
1	15	7.50	200	100.00

PROG_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	182	91.00	182	91.00
1	18	9.00	200	100.00

EXAMINE MISSING DATA PATTERNS: SYNTAX

```
proc mi data=hsb_mar nimpute=0 ;
var write read female math prog ;
ods select misspattern;
run;
```

EXAMINE MISSING DATA PATTERNS

Missing Data Patterns												
								Group Means				
Group	WRITE	READ	FEMALE	MATH	PROG	Freq	Percent	WRITE	READ	FEMALE	MATH	PROG
1	X	X	X	X	X	130	65.00	53.200000	52.838462	0.600000	52.600000	2.046154
2	X	X	X	X	.	15	7.50	56.200000	52.733333	0.466667	55.400000	.
3	X	X	X	.	X	11	5.50	53.090909	51.363636	0.272727	.	2.000000
4	X	X	X	.	.	1	0.50	59.000000	52.000000	1.000000	.	.
5	X	X	.	X	X	15	7.50	49.933333	48.600000	.	49.866667	1.866667
6	X	X	.	X	.	1	0.50	44.000000	44.000000	.	40.000000	.
7	X	X	.	.	X	1	0.50	33.000000	44.000000	.	.	1.000000
8	X	.	X	X	X	9	4.50	51.333333	.	0.444444	53.444444	2.222222
9	.	X	X	X	X	13	6.50	.	54.230769	0.461538	57.076923	1.923077
10	.	X	X	X	.	1	0.50	.	55.000000	1.000000	66.000000	.
11	.	X	X	.	X	2	1.00	.	47.000000	0.500000	.	2.000000
12	.	X	.	X	X	1	0.50	.	39.000000	.	40.000000	3.000000

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

Pearson Correlation Coefficients
Number of Observations

	SOCST	WRITE	READ	FEMALE	MATH	SCIENCE	progcat1	progcat2
SOCST social studies score	1.00000 200	0.59750 183	0.61604 191	0.08894 182	0.54509 185	0.45125 184	-0.07680 182	0.40956 182
WRITE writing score	0.59750 183	1.00000 183	0.58719 174	0.25077 166	0.61825 170	0.54977 168	-0.06036 166	0.34387 166
READ reading score	0.61604 191	0.58719 174	1.00000 191	-0.01740 173	0.65890 176	0.63288 176	-0.10575 173	0.39023 173
FEMALE female	0.08894 182	0.25077 166	-0.01740 173	1.00000 182	-0.02408 168	-0.09176 166	-0.03169 165	0.05004 165
MATH math score	0.54509 185	0.61825 170	0.65890 176	-0.02408 168	1.00000 185	0.62964 169	-0.16511 168	0.44566 168
SCIENCE science score	0.45125 184	0.54977 168	0.63288 176	-0.09176 166	0.62964 169	1.00000 184	0.05672 167	0.20379 167
progcat1	-0.07680 182	-0.06036 166	-0.10575 173	-0.03169 165	-0.16511 168	0.05672 167	1.00000 182	-0.56349 182
progcat2	0.40956 182	0.34387 166	0.39023 173	0.05004 165	0.44566 168	0.20379 167	-0.56349 182	1.00000 182

AUXILIARY VARIABLES ARE PREDICTORS OF MISSINGNESS

The TTEST Procedure

Variable: SOCST (social studies score)

MATH_FLAG	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	185	52.9784	10.4600	0.7690	26.0000	71.0000
1	15	45.3333	11.9323	3.0809	26.0000	66.0000
Diff (1-2)		7.6450	10.5709	2.8379		

MATH_FLAG	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		52.9784	51.4611	54.4956	10.4600	9.4918	11.6501
1		45.3333	38.7254	51.9412	11.9323	8.7360	18.8185
Diff (1-2)	Pooled	7.6450	2.0487	13.2414	10.5709	9.6243	11.7255
Diff (1-2)	Satterthwaite	7.6450	0.9063	14.3838			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	198	2.69	0.0077
Satterthwaite	Unequal	15.794	2.41	0.0287

IMPUTATION MODEL EXAMPLE 1:

MI USING MULTIVARIATE NORMAL DISTRIBUTION (MVN)

ASSUMING A JOINT MULTIVARIATE NORMAL DISTRIBUTION

- Probably the most common parametric approach for multiple imputation.
- Assumes variables are individually and jointly normally distributed
- Assuming a MVN distribution is robust to violations of normality given a large enough N.
- Uses the data augmentation (DA) algorithm to impute.

- Biased estimates may result when N is relatively small and the fraction of missing information is high.

IMPUTATION PHASE

```
proc mi data= new nimpute=10 out=mi_mvn seed=54321;
var socst science write read female math progcat1 progcat2;
run;
```

MULTIPLY IMPUTED DATASET

	Imputation	ID	FEMALE	RACE	SES
194	1	64	female	white	high
195	1	143	male	white	middle
196	1	77	female	white	low
197	1	162	female	white	middle
198	1	33	female	asian	low
199	1	57	female	white	middle
200	1	171	0.3999	white	middle
201	2	116	female	white	middle
202	2	170	male	white	high
203	2	97	male	white	high
204	2	104	male	white	high
205	2	121	female	white	middle
206	2	94	male	white	high

ANALYSIS PHASE: ESTIMATE MODEL FOR EACH IMPUTED DATASET

```
proc glm data = mi_mvn ;
model read = write female math progcat1
progcat2;
by _imputation_;
ods output ParameterEstimates=a_mvn;
run;
quit;
```

PARAMETER ESTIMATE DATASET

	Imputation	Dependent	Parameter	Estimate	StdErr	tValue	Probt
1	1	READ	Intercept	11.00754570	3.38556004	3.25	0.0014
2	1	READ	WRITE	0.39495686	0.07704234	5.13	<.0001
3	1	READ	FEMALE	-2.35626196	1.14135990	-2.06	0.0403
4	1	READ	MATH	0.38455797	0.07607936	5.05	<.0001
5	1	READ	progcat1	3.04400217	1.42233880	2.14	0.0336
6	1	READ	progcat2	-0.06368405	1.51733518	-0.04	0.9666
7	2	READ	Intercept	9.77748432	3.43430958	2.85	0.0049
8	2	READ	WRITE	0.39043178	0.07695156	5.07	<.0001
9	2	READ	FEMALE	-2.47879363	1.08779107	-2.28	0.0238
10	2	READ	MATH	0.42501898	0.07385623	5.75	<.0001
11	2	READ	progcat1	1.57396838	1.42301910	1.11	0.2701
12	2	READ	progcat2	-0.30087173	1.50717207	-0.20	0.8420
13	3	READ	Intercept	9.72160191	3.45457865	2.81	0.0054
14	3	READ	WRITE	0.40459015	0.07897935	5.12	<.0001
15	3	READ	FEMALE	-2.73504544	1.10315252	-2.48	0.0140
16	3	READ	MATH	0.39459091	0.07582946	5.20	<.0001
17	3	READ	progcat1	3.12186214	1.39688774	2.23	0.0266
18	3	READ	progcat2	0.79312586	1.52589004	0.52	0.6038

POOLING PHASE- COMBINING PARAMETER ESTIMATES ACROSS DATASETS

```
proc mianalyze parms=a_mvn;
  modeleffects intercept write female math
                       progcat1  progcat2;
  run;
```

MULTIPLE IMPUTATION REGRESSION - MVN

The MIANALYZE Procedure

Model Information	
PARMS Data Set	WORK.A_MVN
Number of Imputations	10

Variance Information

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
intercept	0.484830	11.599514	12.132827	4658	0.045977	0.044366	0.995583
write	0.000262	0.006014	0.006302	4311	0.047879	0.046134	0.995408
female	0.079066	1.264539	1.351512	2173.3	0.068778	0.065212	0.993521
math	0.000310	0.005865	0.006207	2973.8	0.058215	0.055648	0.994466
progcat1	0.239760	1.955043	2.218779	636.99	0.134900	0.121619	0.987984
progcat2	0.256880	2.339505	2.622073	774.97	0.120781	0.110059	0.989114

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	9.881994	3.483221	3.05323	16.71076	4658	8.523732	11.007546	0	2.84	0.0046
write	0.388897	0.079385	0.23326	0.54453	4311	0.346491	0.404590	0	4.90	<.0001
female	-2.424315	1.162545	-4.70413	-0.14450	2173.3	-2.830625	-2.059445	0	-2.09	0.0372
math	0.414454	0.078783	0.25998	0.56893	2973.8	0.384558	0.446180	0	5.26	<.0001
progcat1	2.332793	1.489557	-0.59224	5.25783	636.99	1.573968	3.121862	0	1.57	0.1178
progcat2	0.302432	1.619282	-2.87627	3.48113	774.97	-0.369425	0.868353	0	0.19	0.8519

COMPARE MIANALYZE ESTIMATES TO ANALYSIS WITH FULL DATA

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9.623171965	B	3.40979657	2.82	0.0053
write	0.374741452		0.07462808	5.02	<.0001
female female	-2.698839662	B	1.09540766	-2.46	0.0146
female male	0.000000000	B	.	.	.
math	0.441863231		0.07499719	5.89	<.0001
prog academic	1.879263080	B	1.42306759	1.32	0.1882
prog general	0.232056170	B	1.51219473	0.15	0.8782
prog vocation	0.000000000	B	.	.	.

FULL DATA ANALYSIS

MIANALYZE OUPUT

Parameter	Estimate	Std Error	95% Confidence Limits	
intercept	9.881994	3.483221	3.05323	16.71076
write	0.388897	0.079385	0.23326	0.54453
female	-2.424315	1.162545	-4.70413	-0.14450
math	0.414454	0.078783	0.25998	0.56893
progcat1	2.332793	1.489557	-0.59224	5.25783
progcat2	0.302432	1.619282	-2.87627	3.48113

HOW DOES PROC MIANALYZE WORK

- PROC MIANALYZE combines results across imputations
- Regression coefficients are averaged across imputations
- Standard errors incorporate uncertainty from 2 sources:
  - "within imputation" - variability in estimate expected with no missing data
    - The usual uncertainty regarding a regression coefficient
  - "between imputation" - variability due to missing information.
    - The uncertainty surrounding missing values

OUTPUT FROM MIANALYZE

Variance Information							
	Variance						
Parameter	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
intercept	0.484830	11.599514	12.132827	4658	0.045977	0.044366	0.995583
write	0.000262	0.006014	0.006302	4311	0.047879	0.046134	0.995408
female	0.079066	1.264539	1.351512	2173.3	0.068778	0.065212	0.993521
math	0.000310	0.005865	0.006207	2973.8	0.058215	0.055648	0.994466
progcat1	0.239760	1.955043	2.218779	636.99	0.134900	0.121619	0.987984
progcat2	0.256880	2.339505	2.622073	774.97	0.120781	0.110059	0.989114

VARIANCE WITHIN

- Sampling variability expected with **no missing data.**
- Average of variability of coefficients within an imputation
- Equal to arithmetic mean of sampling variances (SE^2)
- Example: Add together 10 estimated SE^2 for **write** and divide by 10
- $V_w = 0.006014$

VARIANCE BETWEEN

- Variability in estimates across imputations
  - i.e. the variance of the *m* coefficients
- 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
- $V_B = 0.000262.$

TOTAL VARIANCE

- The total variance is sum of 3 sources of variance.
  - Within,
  - Between
  - Additional source of sampling variance.

- What is the sampling variance?
  - Variance Between divided by number of imputations
  - Represents sampling error associated with the overall coefficient estimates.
  - Serves as a correction factor for using a specific number of imputations.

DEGREES OF FREEDOM

- DF for combined results are determined by the number of imputations.
- *By default DF = infinity, typically not a problem with large N but can be with smaller samples
- The standard formula to estimate DF can yield estimates that are fractional or that far exceed the DF for complete data.
- Correction to adjust for the problem of inflated DF has been implemented
- Use the **EDF** option on the proc mianalyze line to indicate to SAS what is the proper adjusted DF.

RELATIVE INCREASES IN VARIANCE (RVI)

- Proportional increase in total sampling variance due to missing information

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

- For example, the RVI for write coefficient is 0.048, meaning that the sampling variance is 4.8% larger than it would have been with complete data.

FRACTION OF MISSING INFORMATION (FMI)

- Directly related to RVI.
- Proportion of total sampling variance that is due to missing data

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

- For a given variable, FMI based on percentage missing and correlations with other imputation model variables.
- Interpretation similar to an R^2.
  - Example: FMI=.046 for **write** means that 4.6% of sampling variance is attributable to missing data.

RELATIVE EFFICIENCY: IS 5 IMPUTATIONS ENOUGH?

- Captures how well true population parameters are estimated
- Related to both the FMI and *m*
- Low FMI + few *m* = high efficiency
- As FMI increase so should *m:*
  - Better statistical power and more stable estimate

Table 62.2: Relative Efficiencies

	λ				
m	**10%**	**20%**	**30%**	**50%**	**70%**
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

DIAGNOSTICS:
HOW DO I KNOW IF IT WORKED?

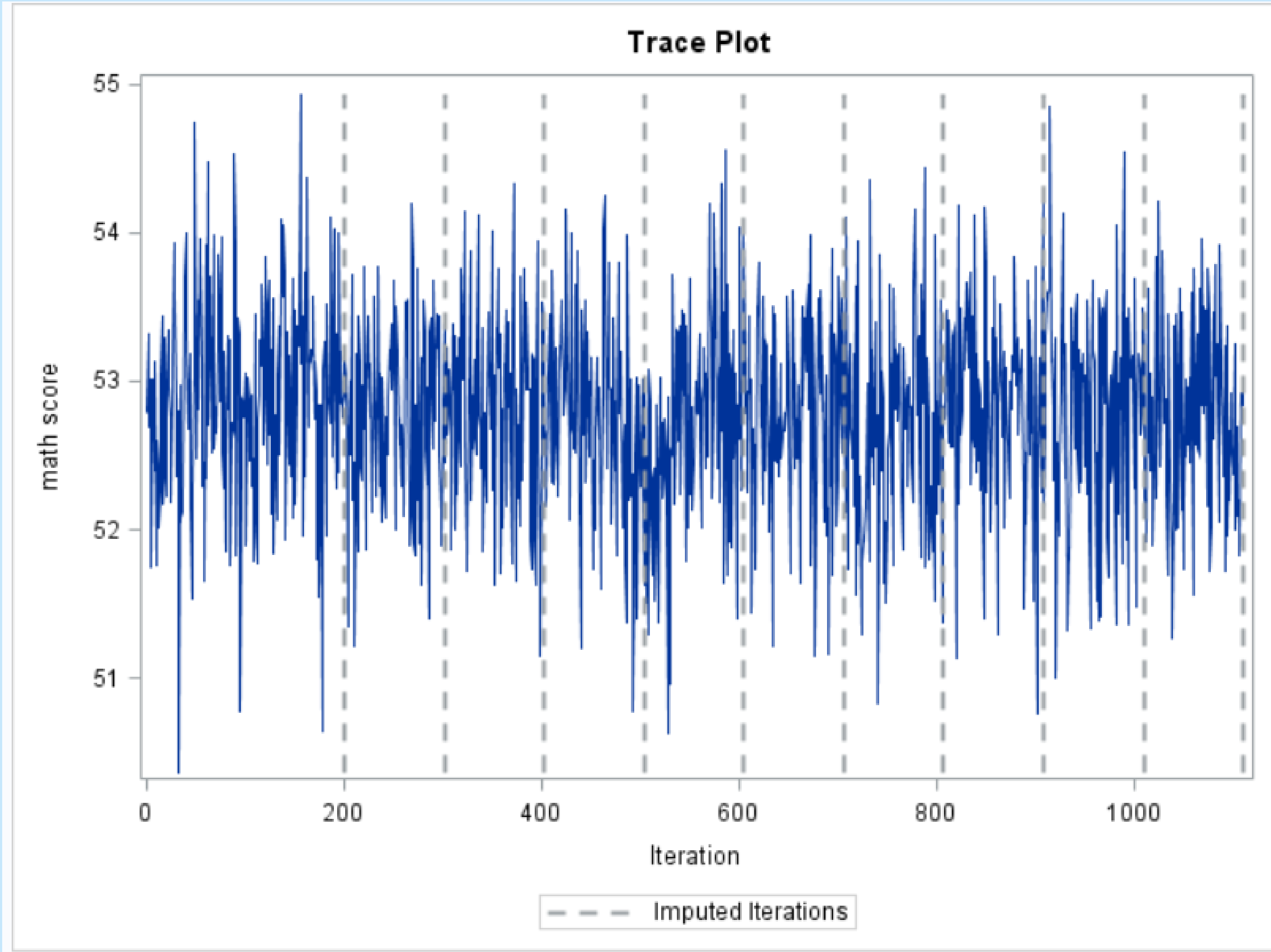
- Compare means and frequencies of observed and imputed values.
  - Use boxplots to compare distributions

- Look at "**Variance Information**" tables from the proc mianalyze output

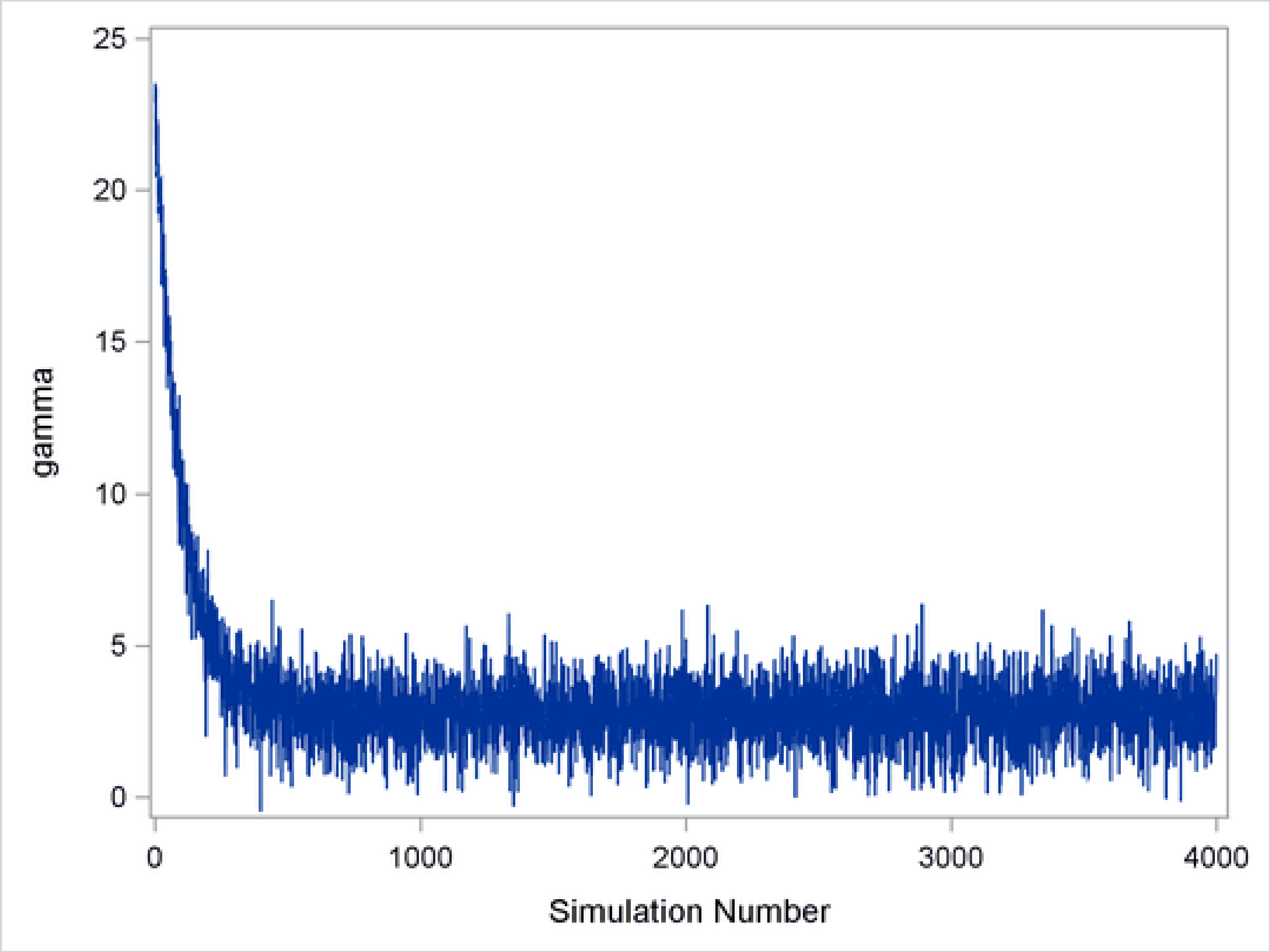
- **Plots - Assess convergence of DA algorithm**

TRACE PLOTS:
DID MY IMPUTATION MODEL CONVERGE?

- Convergence for each imputed variable can be assessed using trace plots.
- Examine for each imputed variables
- Special attention to variables with a high FMI
- **proc mi data= ats.hsb_mar nimpute=10 out=mi_mvn;**
  mcmc plots=trace plots=acf ;
  var socst write read female math;
  run;

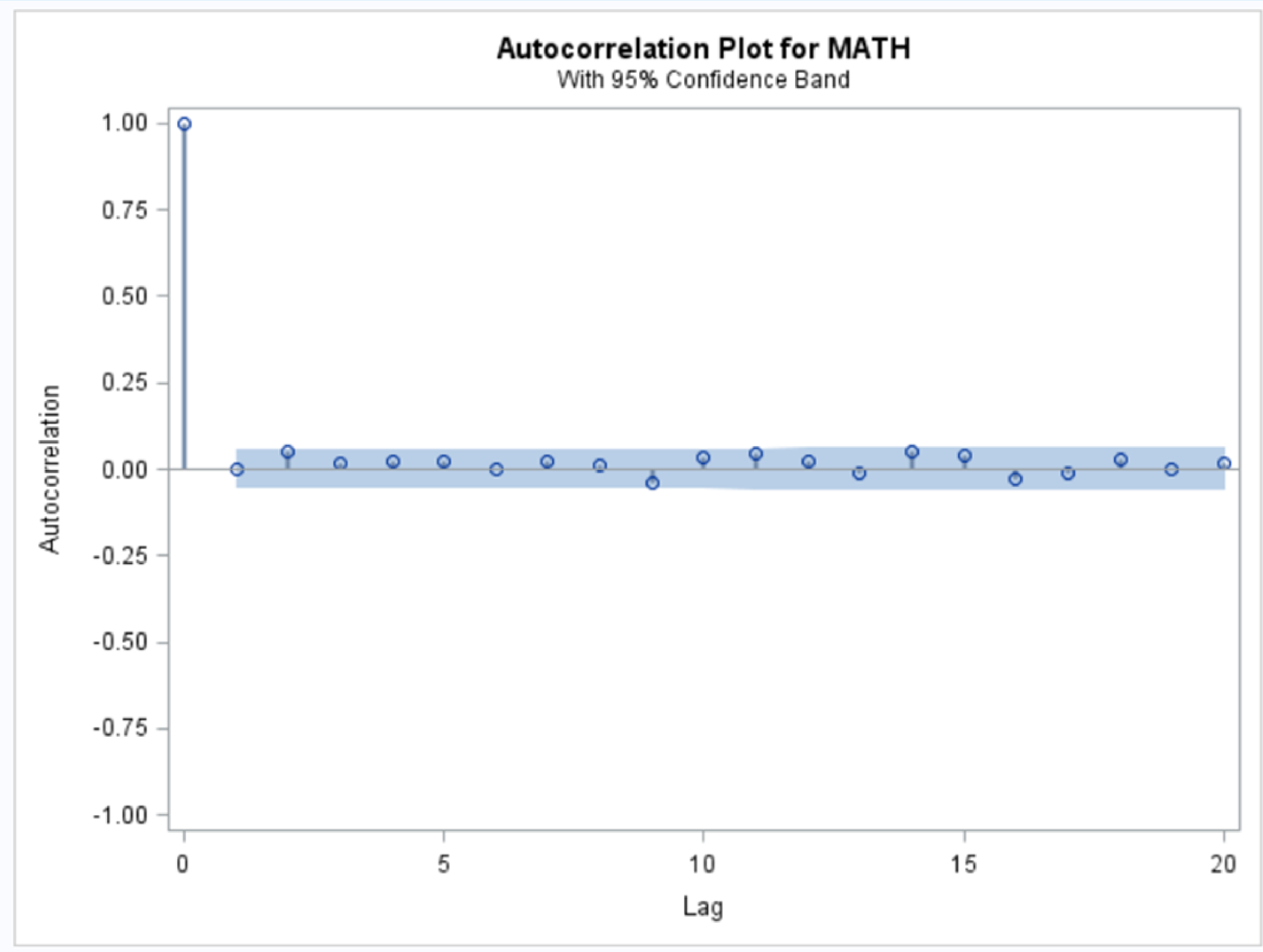

Trace Plot
math score
55
54
53
52
51
0
200
400
600
800
1000
Iteration
Imputed Iterations

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.
- ```
  proc mi data= ats.hsb_mar nimpute=10
  out=mi_mvn;
  mcmc plots=trace  plots=acf ;
  var socst write read female math;
  run;
  ```
  ```


Autocorrelation Plot for MATH
With 95% Confidence Band
Autocorrelation
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
0
5
10
15
20
Lag

IMPUTATION MODEL EXAMPLE 2:
MI USING FULLY CONDITIONAL SPECIFICATION (FCS)

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

- Alternative method for imputation is Fully Conditional Method (FCS)
- FCS does not assume a joint distribution and allows the use of different distributions across variables.
- Each variable with missing is allowed its own *type* of regression (linear, logistic, etc) for imputation
- Example uses:
  - Logistic model for binary outcome
  - Poisson model for count variable
  - Other bounded values

AVAILABLE DISTRIBUTIONS

- FCS methods available:
  - Discriminant function or logistic regression for binary/categorical variables
  - Linear regression and predictive mean matching for continuous variables.

- Properties to Note:

1. Discriminant function only continuous vars as covariates (default).
2. Logistic regression assumes ordering of class variables if more then two levels (default).
3. Regression is default imputation method for continuous vars.
4. PMM will provide “plausible” values.
   1. For an observation missing on X, finds cases in data with similar values on other covariates, then randomly selects an X value from those cases

IMPUTATION PHASE

```
proc mi data= ats.hsb_mar nimpute=20
out=mi_fcs ;
class female prog;
fcs plots=trace(mean std);
var socst write read female math science prog;
fcs discrim(female prog /classeffects=include)
   nbiter =100 ;
   run;
```

ALTERNATE EXAMPLE

```
proc mi data= ats.hsb_mar nimpute=20
out=mi_new1;
class female prog;
var socst write read female math science prog;
fcs logistic (female= socst science) ;
fcs logistic (prog =math socst /link=glogit)
regpmm(math read write);
run;
```

ANALYSIS PHASE: ESTIMATE GLM MODEL USING EACH IMPUTED DATASET

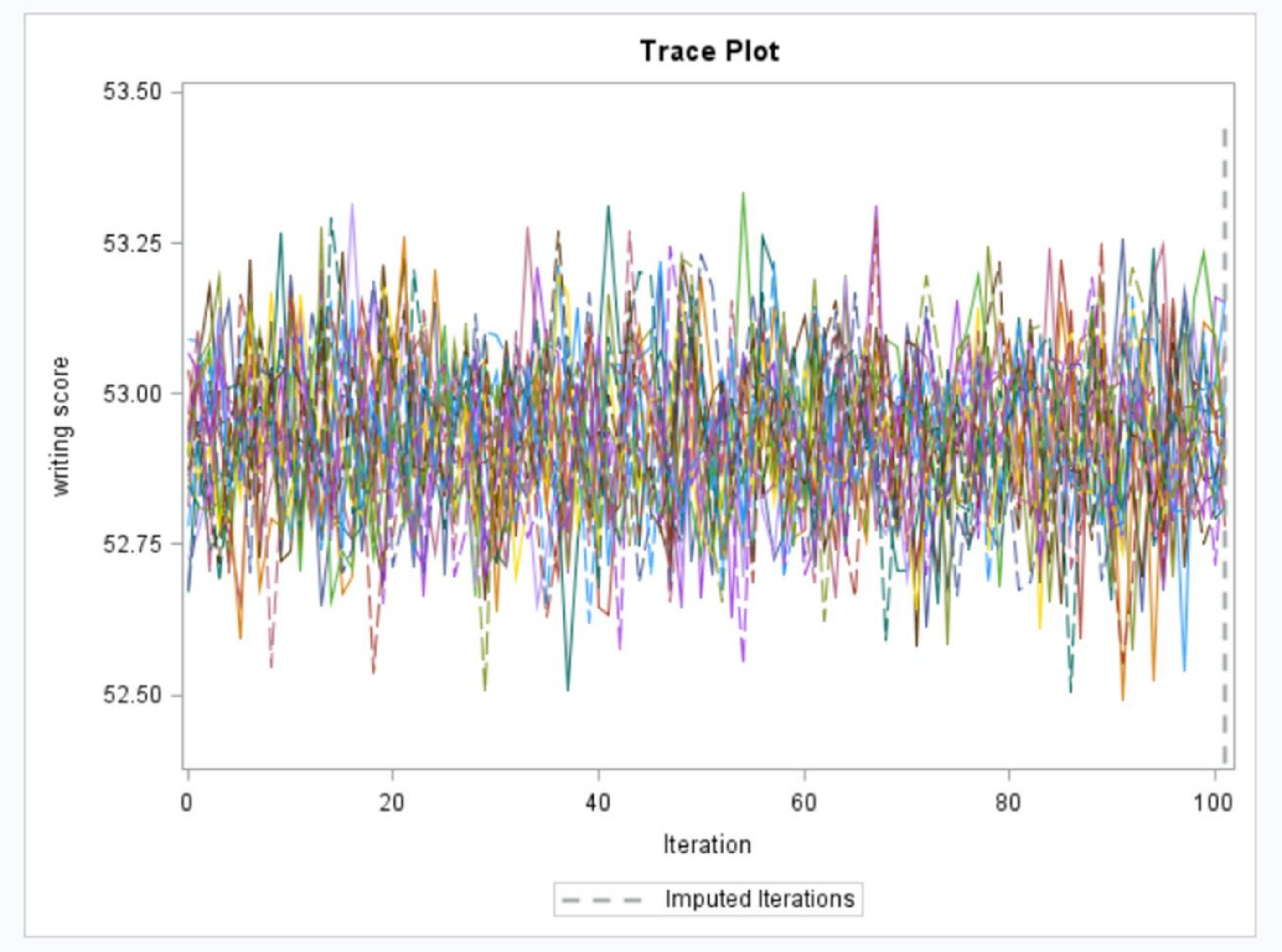
```
proc genmod data=mi_fcs;
class female prog;
model read=write female math prog/dist=normal;
by _imputation_;
ods output ParameterEstimates=gm_fcs;
run;
```

Obs	_Imputation_	Parameter	Level1	DF	Estimate	StdErr	LowerWaldCL	UpperWaldCL	ChiSq	ProbChiSq
1	1	Intercept		1	9.3341	3.3237	2.8197	15.8484	7.89	0.0050
2	1	WRITE		1	0.4064	0.0796	0.2504	0.5624	26.07	<.0001
3	1	FEMALE	female	1	-2.8816	1.1653	-5.1655	-0.5977	6.12	0.0134
4	1	FEMALE	male	0	0.0000	0.0000	0.0000	0.0000	.	.
5	1	MATH		1	0.3998	0.0766	0.2497	0.5498	27.26	<.0001
6	1	PROG	academic	1	2.9736	1.3657	0.2969	5.6502	4.74	0.0295
7	1	PROG	general	1	1.0303	1.4963	-1.9024	3.9631	0.47	0.4911
8	1	PROG	vocation	0	0.0000	0.0000	0.0000	0.0000	.	.
9	1	Scale		1	7.1943	0.3597	6.5227	7.9351	_	_
10	2	Intercept		1	9.3267	3.3800	2.7019	15.9514	7.61	0.0058
11	2	WRITE		1	0.4185	0.0797	0.2623	0.5748	27.55	<.0001
12	2	FEMALE	female	1	-2.2714	1.1349	-4.4959	-0.0470	4.01	0.0454
13	2	FEMALE	male	0	0.0000	0.0000	0.0000	0.0000	.	.
14	2	MATH		1	0.3893	0.0798	0.2329	0.5456	23.81	<.0001
15	2	PROG	academic	1	2.4804	1.4041	-0.2717	5.2324	3.12	0.0773
16	2	PROG	general	1	0.2041	1.5302	-2.7951	3.2032	0.02	0.8939
17	2	PROG	vocation	0	0.0000	0.0000	0.0000	0.0000	.	.
18	2	Scale		1	7.2457	0.3623	6.5693	7.9917	_	_

POOLING PHASE-
COMBINING PARAMETER ESTIMATES ACROSS DATASETS

```
PROC MIANALYZE parms(classvar=level)=gm_fcs;
class female prog;
MODELEFFECTS INTERCEPT write female math
prog;
RUN;
```

TRACE PLOTS:
DID MY IMPUTATION MODEL CONVERGE?



FCS HAS SEVERAL PROPERTIES THAT MAKE IT AN ATTRACTIVE ALTERNATIVE

1. **FCS** 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.

BOTTOM LINE

- MI improves over single imputation methods because:
  - Single value never used
  - Appropriate estimates of uncertainty
- The nature of your data and model will determine if you choose MVN or FCS
- Both are state of the art methods for handling missing data
  - Produce unbiased estimates assuming MAR