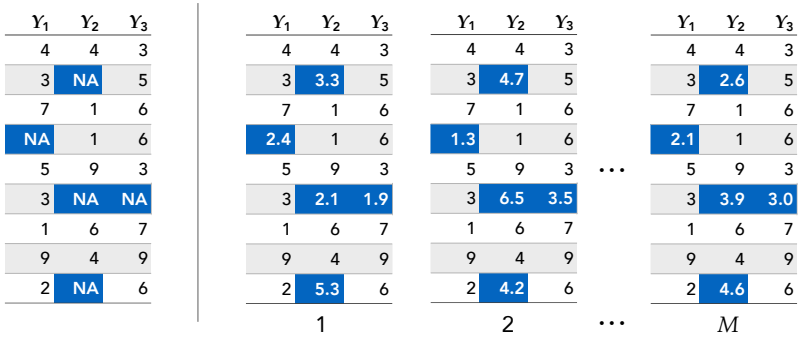
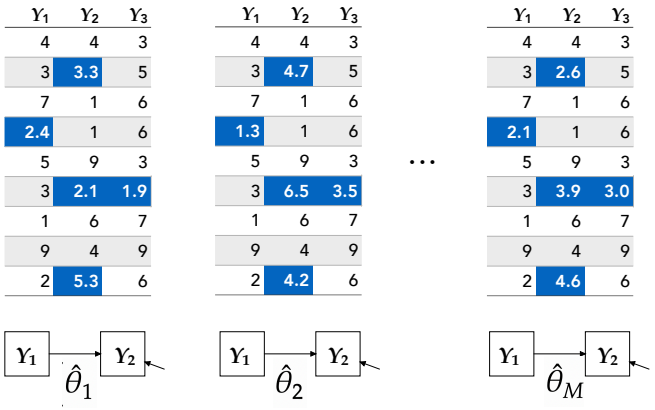
10: Multiple Imputation Imputation Phase

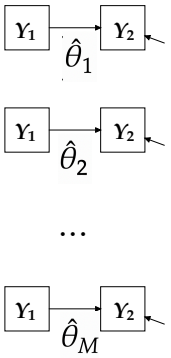
Multiple Imputation Step 1: Create M Complete Data Sets



Multiple Imputation Step 2: Perform Analysis on Each Data Set



Multiple Imputation Step 3: Combine Estimates and Standard Errors



$$\hat{\theta} = \left(\hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_M\right)/M$$

5

Bayesian vs. Multiple Imputation

The difference between multiple imputation and Bayesian estimation can get very blurry

Bayesian estimation is the mathematical machinery behind multiple imputation

The model used to create imputations may or may not be the analysis of substantive interest

6

Main Points Of Contrast

Bayesian Analysis	Multiple Imputation
Main goal is to estimate and interpret model parameters	Main goal is to perform secondary data analysis on imputed data sets
Imputation happens behind the scenes, is a means to an end	Imputations are the primary focus, we want to save them for future use

7

Why Use Multiple Imputation?

Don't have access to software that estimates the analysis model of scientific interest

Readily incorporates auxiliary variables

Perform several analyses on the same filled-in data

Prefer frequentist interpretations and significance tests

Analysis model involves composite scores

8

Paired-Samples t Test

Regression parameterization for a dependent-samples t test evaluating change between the two assessments, Y_1 and Y_2

Mean change

$$(Y_{2i} - Y_{1i}) = \beta_0 + \varepsilon_i = E(Y_\Delta) + \varepsilon_i$$

$$Y_{\Delta i} \sim N\left(E(Y_\Delta), \sigma^2_\varepsilon\right)$$

Change score ($Y_{\Delta i}$)

Math Achievement Data

Math achievement data for 250 students

The data set includes pre-test and post-test math achievement scores and academic-related variables such as math self-efficacy, math anxiety, standardized reading scores, socio-demographic variables

math.dat

Variable	Name	Missing %	Scaling
Identifier variable	ID	0	Integer index
Gender	MALE	0	0 = female, 1 = male
Free or reduced lunch	LUNCHASST	4.3	0 = none, 1 = assistance
Achievement group	ACHIEVEGRP	2.0	1 = typically achieving, 2 = low achieving, 3 = learning disability
Standardized reading	STANREAD	10.0	Continuous
Math self-efficacy	EFFICACY	9.7	6-point ordinal scale
Math anxiety	ANXIETY	9.3	Continuous
Pre-test math achievement	MATHPRE	0	Continuous
Post-test math achievement	MATHPOST	18.0	Continuous

Substantive Example

Do math scores improve between the pre-test and post-test assessments?

$$(MATHPOST_i - MATHPRE_i) = \beta_0 + \varepsilon_i$$

Pre-test scores are complete, 18% of the post-test scores are missing and need to be imputed

Minimally Sufficient Imputation Model

Imputation treats incomplete variables as outcomes and complete variables as predictors

$$Y_{2i} = \gamma_0 + \gamma_1(Y_{1i}) + e_i + E(Y_2|Y_1) + e_i$$
$$Y_{2i} \sim N\left(E(Y_2|Y_1), \sigma_e^2\right)$$

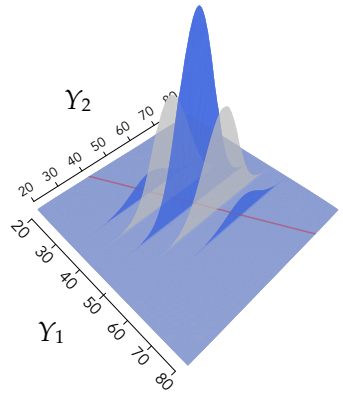
At a minimum, the imputation procedure should include all variables in the substantive analysis

Distribution Of The Post-Test Scores

The imputation model says Y_2 scores are normally distributed around predicted values

$$Y_{2i} \sim N\left(E(Y_2|Y_1), \sigma_e^2\right)$$

The residual variance defines the spread of the Y_2 scores around the regression line



We already know how to do this!

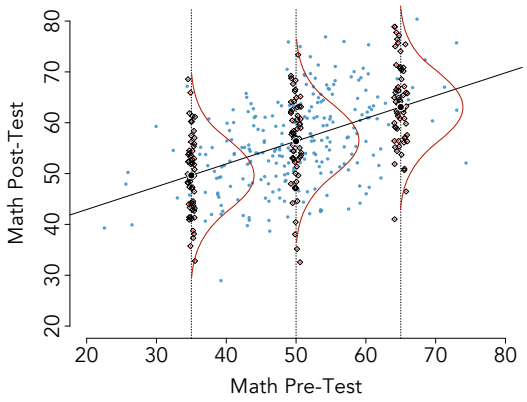
MCMC Recipe

Do for t = 1 to T iterations

1. Estimate the regression coefficients, given the filled-in data and current value of the residual variance
2. Estimate the residual variance, given the filled-in data and current values of the coefficients
3. Estimate (impute) missing values, given the regression model parameters

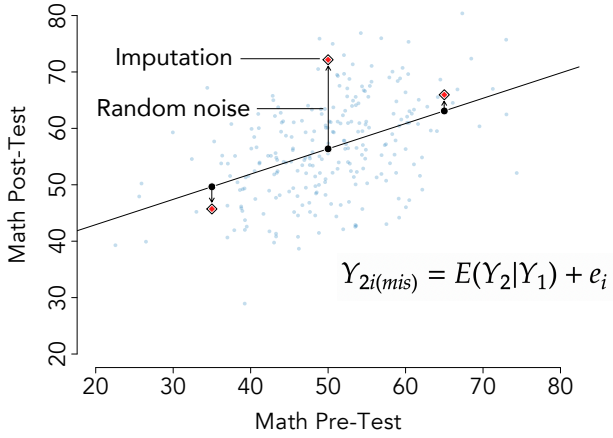
Repeat

Distribution Of Imputations At Three Pre-Test Values



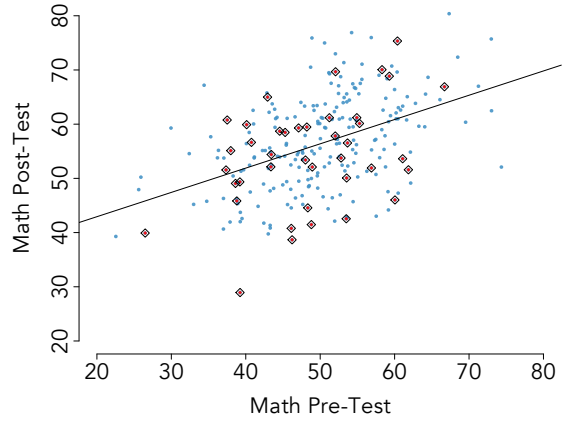
17

Imputation = Predicted Score + Noise



18

One Imputed Data Set



19

How Is This Different From A Bayesian Analysis?

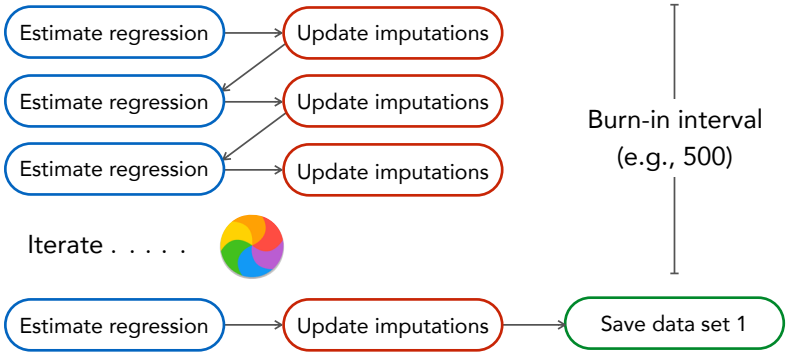
The Bayesian regression is a device to fill in the data, it is not the same as the analysis model

We need to save imputations for a secondary data analysis in a statistical software package

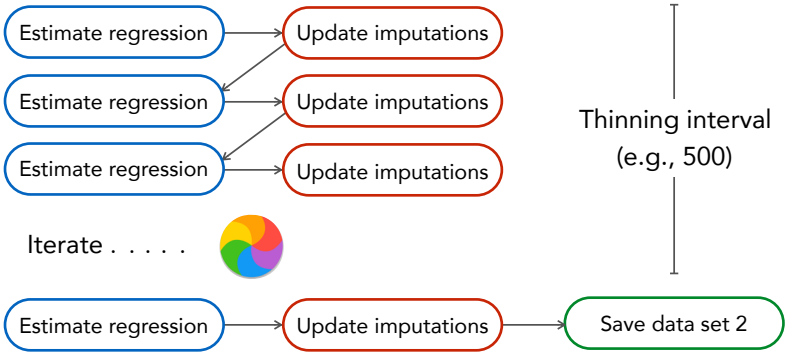
Recommendations suggest at least 20 data sets

20

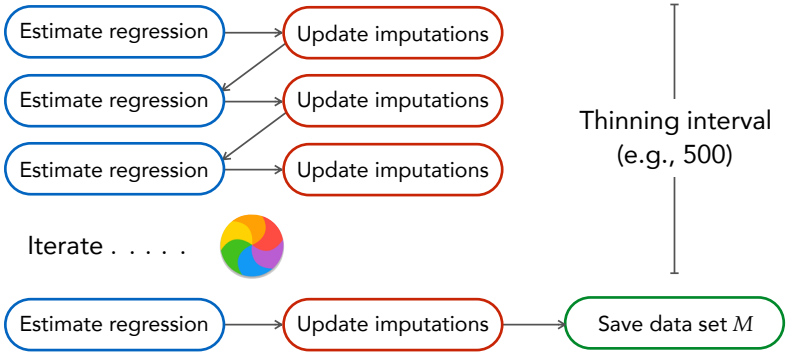
MCMC Burn-In Interval



MCMC Thinning Interval

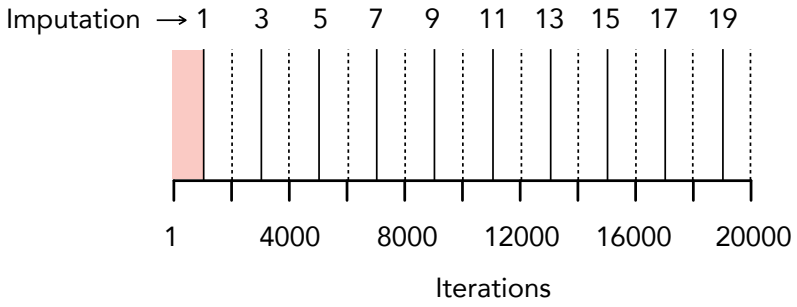


MCMC Thinning Interval Continued

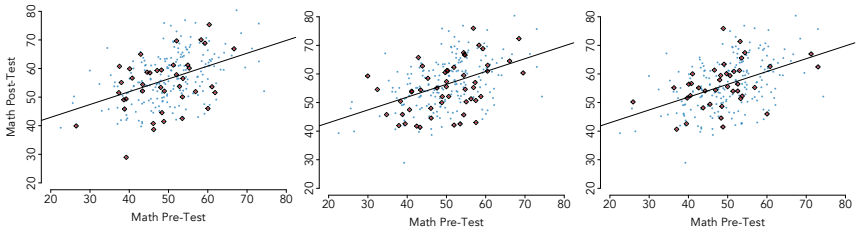


Generating 20 Imputations

20 imputations from an MCMC algorithm with 1000 burn-in iterations and 1000 thinning iterations



Three Imputed Data Sets



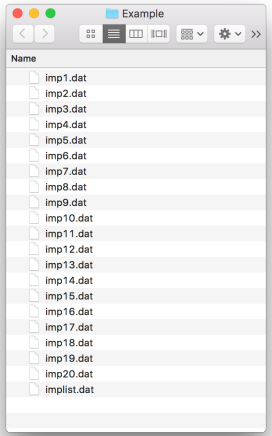
Blimp Imputation Script for Analysis in Mplus

```
DATA: math.dat;
VARIABLES: id male lunchasst achievegrp stanread efficacy
    anxiety mathpre mathpost;
MISSING: 999;
FCS: mathpre mathpost;
SEED: 90291;
NIMPS: 20;
BURN: 1000;
THIN: 1000;
CHAINS: 2 processors 2;
OPTIONS: psr;
SAVE: separate = imps_*.dat;
```


Mplus Imputation Format

Mplus requires imputed data sets as separate files

Blimp creates a text file containing the names of the data sets, and this file serves as the input data for subsequent analyses



Blimp Model-Based Imputation Script for Analysis in R, SAS, SPSS, Stata

```
DATA: math.dat;
VARIABLES: id male lunchasst achievegrp stanread efficacy
    anxiety mathpre mathpost;
MISSING: 999;
FCS: mathpre mathpost;
SEED: 90291;
NIMPS: 20;
BURN: 1000;
THIN: 1000;
CHAINS: 2 processors 2;
OPTIONS: psr;
SAVE: stacked0 = imps_stacked.dat;
```

Auxiliary Variables Revisited

An auxiliary variable is an ancillary variable that correlates with missingness or the analysis variables

Introducing auxiliary variables into imputation can improve power or reduce bias

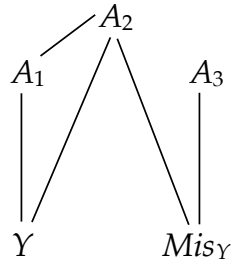
The benefit of an auxiliary variable depends on the pattern and magnitude of its correlations with the analysis variables and missing data indicators

Hierarchy Of Auxiliary Variables

Conditioning on A_1 can improve power but ignoring it does not introduce bias

Ignoring A_2 induces an NMAR mechanism and nonresponse bias

A_3 has no effect on bias and power and should not be used



Indicator And Auxiliary Variable Correlations

	Post-Test Missing Indicator	Math Post-Test
MALE	0.01	-0.18
LUNCHASST	-0.06	**-0.26**
STANREAD	-0.07	**0.49**
EFFICACY	-0.00	**0.34**
ANXIETY	0.05	**-0.43**
MATHPRE	-0.07	0.51
ACHGRP2	0.03	-0.15
ACHGRP3	0.13	-0.10

Practical Conclusions

There are no A_2 variables with strong enough correlations to introduce bias if ignored

Four A_1 variables are moderately or strongly correlated with depression (the focal predictor)

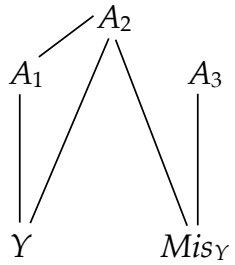
Imputing with these additional variables can reduce standard errors and improve power

Hierarchy Of Auxiliary Variables

Conditioning on A_1 improves power but ignoring it does not introduce bias

Ignoring A_2 induces an NMAR mechanism and nonresponse bias

A_3 cannot introduce bias nor can it increase power



Incorporating Auxiliary Variables

Auxiliary variables function as extra predictors in the imputation regression model

If incomplete (as they are here), they also function as outcomes to be imputed

Include auxiliary variables in the imputation procedure, then ignore them during analysis

Fully Conditional Specification (FCS)

FCS is a round robin imputation scheme where each incomplete variable is regressed on all other variables

The imputed variable from one step serves as a predictor in all other imputation regressions

Mathematically, FCS is a sequence of univariate regression models estimated via MCMC

FCS Imputation Model Sequence

Current iteration — Previous iteration

Impute Y_1 conditional on current Y_2 and Y_3

$$Y_{1i}^{(t)} = \gamma_{10} + \gamma_{11}\left(Y_{2i}^{(t-1)}\right) + \gamma_{12}\left(Y_{3i}^{(t-1)}\right) + e_{1i}$$

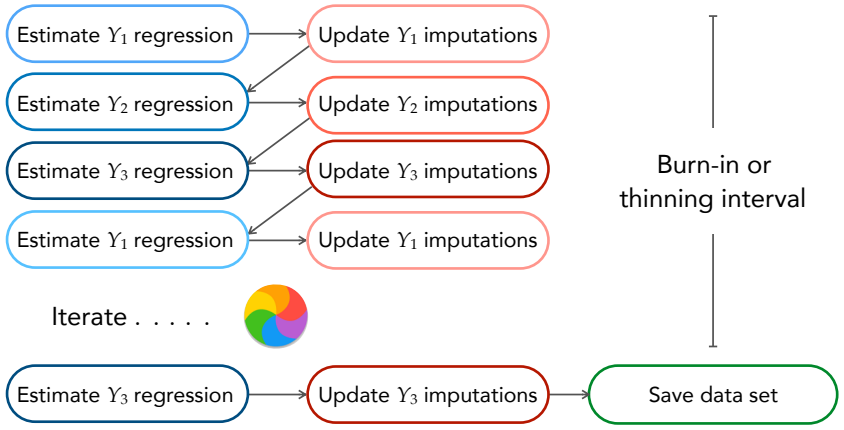
Impute Y_2 conditional on Y_3 and updated Y_1

$$Y_{2i}^{(t)} = \gamma_{20} + \gamma_{21}\left(Y_{1i}^{(t)}\right) + \gamma_{22}\left(Y_{3i}^{(t-1)}\right) + e_{2i}$$

Impute Y_3 conditional on updated Y_1 and Y_2

$$Y_{3i}^{(t)} = \gamma_{30} + \gamma_{31}\left(Y_{1i}^{(t)}\right) + \gamma_{32}\left(Y_{2i}^{(t)}\right) + e_{3i}$$

MCMC Algorithm



Imputation Regression Models

$$MATHPOST^{(t)} \sim MATHPRE + LUNCH^{(t-1)} + STANREAD^{(t-1)} + EFFICACY^{(t-1)} + ANXIETY^{(t-1)}$$

$$LUNCH^{(t)} \sim MATHPRE + MATHPOST^{(t)} + STANREAD^{(t-1)} + EFFICACY^{(t-1)} + ANXIETY^{(t-1)}$$

$$STANREAD^{(t)} \sim MATHPRE + MATHPOST^{(t)} + LUNCH^{(t)} + EFFICACY^{(t-1)} + ANXIETY^{(t-1)}$$

$$EFFICACY^{(t)} \sim MATHPRE + MATHPOST^{(t)} + LUNCH^{(t)} + STANREAD^{(t)} + ANXIETY^{(t-1)}$$

$$ANXIETY^{(t)} \sim MATHPRE + MATHPOST^{(t)} + LUNCH^{(t)} + STANREAD^{(t)} + EFFICACY^{(t)}$$

Blimp Imputation Script For Analysis In Mplus

```
DATA: math.dat;
VARIABLES: id male lunchasst achievegrp stanread efficacy
    anxiety mathpre mathpost;
MISSING: 999;
ORDINAL: lunchasst efficacy;
FCS: mathpre mathpost lunchasst stanread efficacy anxiety;
SEED: 90291;
NIMPS: 20;
BURN: 2000;
THIN: 2000;
CHAINS: 4 processors 4;
OPTIONS: psr;
SAVE: separate = imps_*.dat;
```


Blimp Imputation Script For Analysis In R, SAS, SPSS, Stata

```
DATA: math.dat;
VARIABLES: id male lunchasst achievegrp stanread efficacy
    anxiety mathpre mathpost;
MISSING: 999;
ORDINAL: lunchasst efficacy;
FCS: mathpre mathpost lunchasst stanread efficacy anxiety;
SEED: 90291;
NIMPS: 20;
BURN: 2000;
THIN: 2000;
CHAINS: 4 processors 4;
OPTIONS: psr;
SAVE: stacked0 = imps_stacked.dat;
```

Output Data Information

Stacked file format (R, SAS, SPSS, Stata)

```
VARIABLE ORDER IN SAVED DATA:

imp# id male lunchasst achievegrp stanread efficacy anxiety mathpre mathpost
```

Separate file format (Mplus)

```
VARIABLE ORDER IN SAVED DATA:

id male lunchasst achievegrp stanread efficacy anxiety mathpre mathpost
```

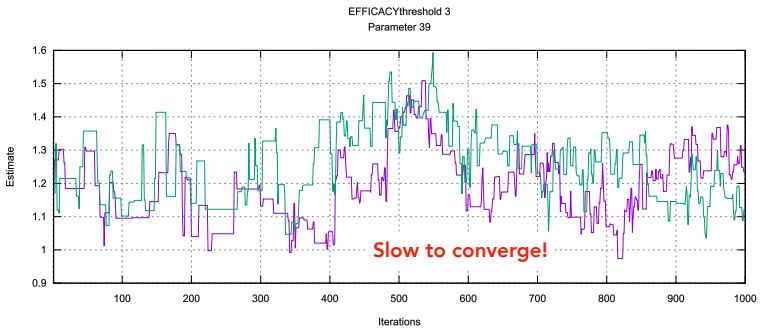

Blimp Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 4 chains   | Highest PSR | Parameter # |
                          51 to 100      |       1.396 |          40 |
                         101 to 200      |       1.216 |          40 |
                         151 to 300      |       1.373 |          40 |
                         201 to 400      |       1.402 |          39 |
                         251 to 500      |       1.169 |          39 |
                                ...
                         801 to 1600     |       1.076 |          41 |
                         851 to 1700     |       1.067 |          41 |
                         901 to 1800     |       1.046 |          41 |
                         951 to 1900     |       1.042 |          41 |
                        1001 to 2000     |       1.041 |          39 |
```

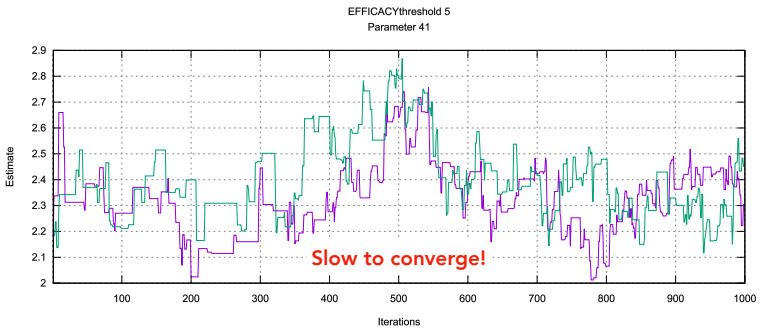

Blimp Trace Plots

Thresholds from 1000 iterations (parameter #39)



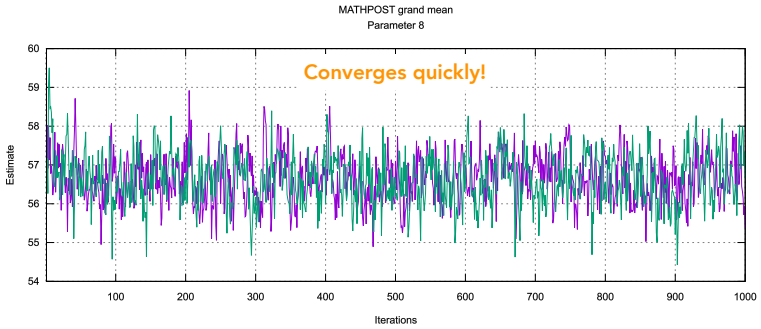
Blimp Trace Plots

Thresholds from 1000 iterations (parameter #41)



Blimp Trace Plots

Grand mean estimates from 1000 iterations



Blimp Trace Plots

Pre-test slope estimates from 1000 iterations