7. MCMC Convergence Diagnostics

Burn-In Period

The examples thus far have discarded estimates from the first 1,000 iterations

This so-called burn-in period allows MCMC to recover from its (usually poor) starting values and to achieve a steady state

How do we determine burn-in length?

Convergence And Mixing

MCMC draws parameters at random from a distribution, so they continually change

Convergence ➠ How long does it take the algorithm to reach a steady state distribution?

Mixing ➠ How long does it take the algorithm to map the entire area under the curve?

Regression Analysis Model

Multiple regression with three predictors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i = E(Y|X) + \varepsilon_i$$

$$Y_i \sim N\left(E(Y|X), \sigma^2_\varepsilon\right)$$

Y is normally distributed around the regression line (predicted values) with constant variation

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 = learning disability, 2 = low achieving, 3 = average achieving
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

Math post-test scores regressed on pre-test scores, math self-efficacy, and anxiety

$$MATHPOST_i = \beta_0 + \beta_1(MATHPRE_i) + \beta_2(EFFICACY_i) + \beta_3(ANXIETY_i) + \varepsilon_i$$

Self-efficacy is a 6-point rating that we will treat as continuous for now

MCMC Recipe

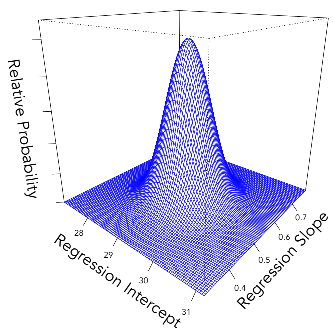
Do for t = 1 to T iterations

1. Estimate substantive model parameters (regression coefficients, residual variance), given the filled-in data
2. Impute missing Y values, given the regression model parameters
3. Estimate predictor model regression parameters (regression coefficients, residual variances), given the filled-in data
4. Impute missing predictors, given two sets of model parameters

Repeat

Conditional Distribution Of The Coefficients

$$f\left(\beta \mid \sigma^2_\varepsilon, data\right) \propto f(\beta) \times f\left(data \mid \beta, \sigma^2_\varepsilon\right) \propto MVN\left(\hat{\beta}_{OLS}, \Sigma_{\hat{\beta}}\right)$$



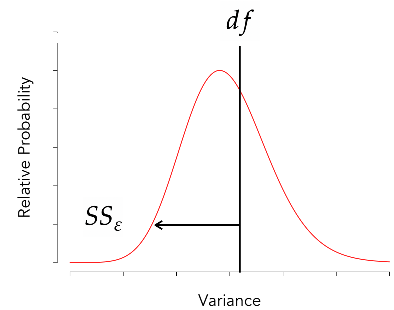
The posterior distribution of β is a multivariate normal distribution centered at OLS estimates

$$data = \left(Y_{(mis)}, Y_{(obs)}, X\right)$$

9

Conditional Distribution Of The Variance

$$f\left(\sigma^2_\varepsilon \mid \beta, data\right) \propto f\left(\sigma^2_\varepsilon\right) \times f\left(data \mid \beta, \sigma^2_\varepsilon\right) \propto IG\left(\frac{df}{2}, \frac{\Sigma\left(Y_i - E(Y|X)\right)^2}{2}\right)$$



The posterior distribution of σ^2_ε is a positively skewed inverse gamma distribution

$$data = \left(Y_{(mis)}, Y_{(obs)}, X\right)$$

10

Blimp Analysis Script

DATA: math.dat;
VARIABLES: id male lunchasst achievegrp stanread efficacy
anxiety mathpre mathpost;
MISSING: 999;
MODEL: mathpost ~ mathpre efficacy anxiety;
SEED: 90291;
BURN: 1000;
ITERATIONS: 10000;
CHAINS: 4 processors 4;
OPTIONS: psr;

11

Blimp Output

ANALYSIS MODEL ESTIMATES:

Missing outcome: mathpost

Parameters	Mean	Median	StdDev	Lower 2.5	Upper 97.5
Variances:					
Residual Var.	57.493	57.153	5.801	47.243	69.925
Coefficients:					
Intercept	38.774	38.774	4.532	29.865	47.568
mathpre	0.365	0.366	0.071	0.228	0.505
efficacy	1.196	1.199	0.374	0.453	1.933
anxiety	-0.243	-0.244	0.084	-0.409	-0.076
Standardized Coefficients:					
mathpre	0.341	0.342	0.062	0.218	0.459
efficacy	0.201	0.202	0.062	0.077	0.320
anxiety	-0.206	-0.208	0.070	-0.339	-0.065
Proportion Variance Explained					
by Fixed Effects	0.331	0.331	0.050	0.231	0.425
by Residual Variation	0.669	0.669	0.050	0.575	0.769

Summaries based on 10000 iterations using 4 chains

12

Mplus Analysis Script

DATA:
file = math.dat;
VARIABLE:
names = id male frlunch achgroup stanread matheff t1math t2math;
usevariables = mathpost mathpre efficacy anxiety;
missing = all(999);
ANALYSIS:
estimator = bayes;
bseed = 90291;
fbiterations = 10000;
MODEL:
mathpre efficacy anxiety with mathpre efficacy anxiety;
mathpost on mathpre efficacy anxiety;
OUTPUT:
tech8 stdyx;

Mplus Output

MODEL RESULTS

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%
MATHPOST ON					
MATHPRE	0.366	0.071	0.000	0.229	0.506
EFFICACY	1.195	0.381	0.002	0.444	1.937
ANXIETY	-0.244	0.085	0.002	-0.408	-0.077
...					
Intercepts					
MATHPOST	38.775	4.589	0.000	29.729	47.672
...					
Residual Variances					
MATHPOST	57.643	5.925	0.000	47.544	70.783

Mplus Output

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%
MATHPOST ON					
MATHPRE	0.344	0.064	0.000	0.218	0.467
EFFICACY	0.203	0.063	0.002	0.074	0.324
ANXIETY	-0.210	0.072	0.002	-0.346	-0.066
...					
Residual Variances					
MATHPOST	0.664	0.054	0.000	0.557	0.772

R-SQUARE

Variable	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%
MATHPOST	0.336	0.054	0.000	0.228	0.443

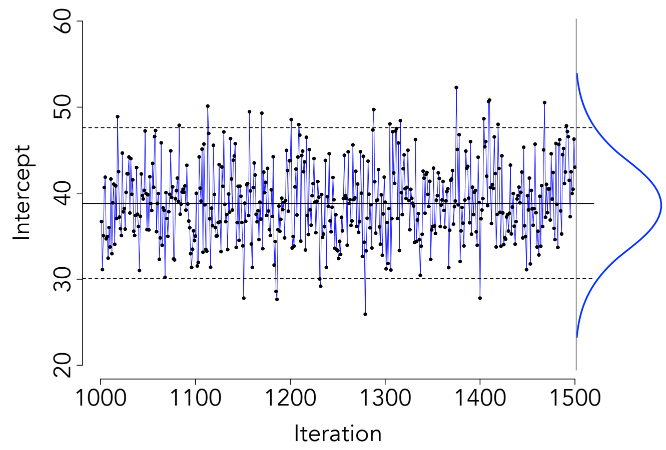
Trace Plots Revisited

A trace plot is a line graph that displays the iterations on the horizontal axis and the parameter estimates on the vertical axis

Trace plots are important tools for evaluating whether the MCMC algorithm is working well

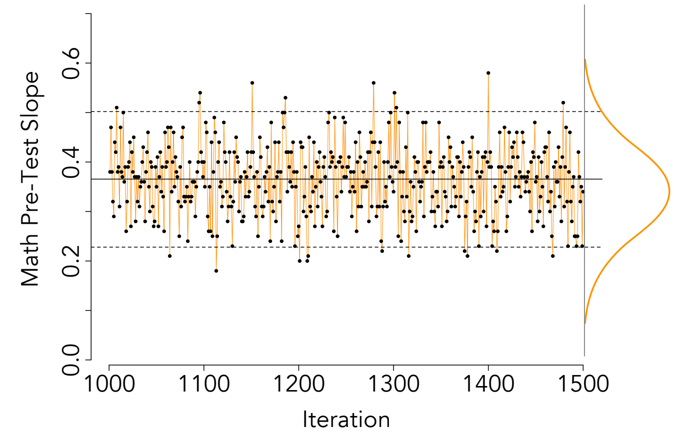
Trace plots provide visual verification that MCMC is converging and mixing well

Trace Plot Of 500 Intercept Estimates



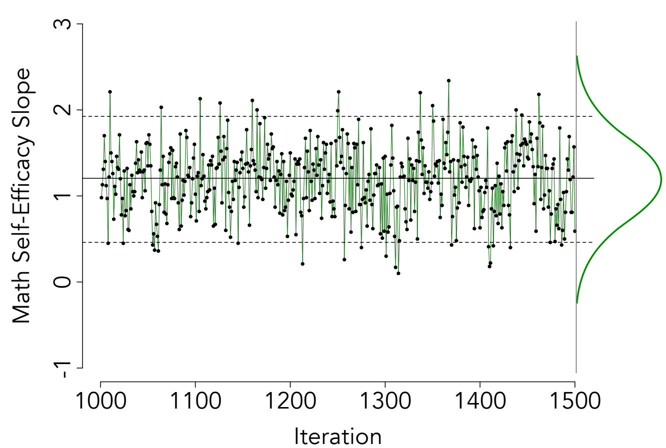
17

Trace Plot Of 500 Slope Estimates



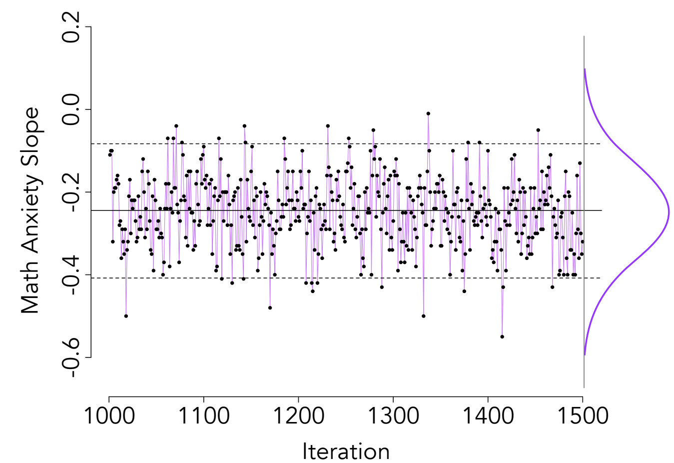
18

Trace Plot Of 500 Slope Estimates



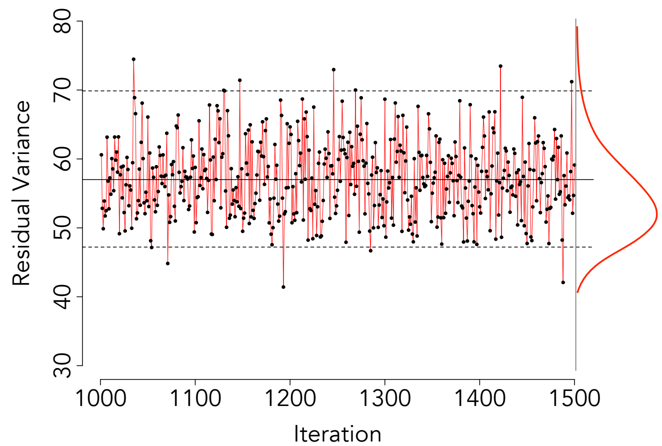
19

Trace Plot Of 500 Slope Estimates



20

Trace Plot Of 500 Variance Estimates



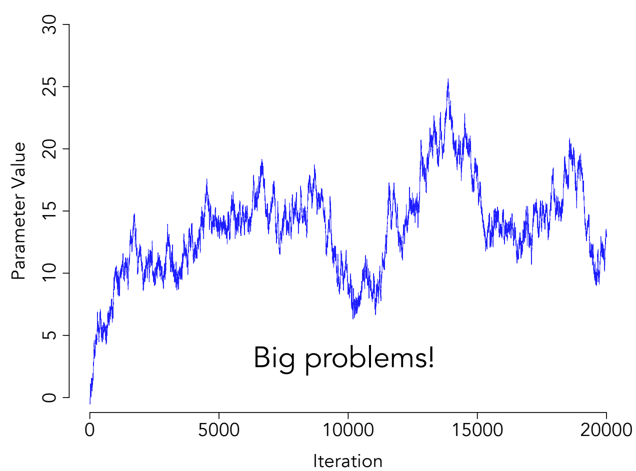
Interpretations

Trace plots appear optimal after the 1,000 iteration burn-in period

Parameter estimates oscillate around a stable center line or mean, indicating convergence

MCMC samples estimates throughout the entire distribution, indicating good mixing

Trace Plot Of A Convergence Failure



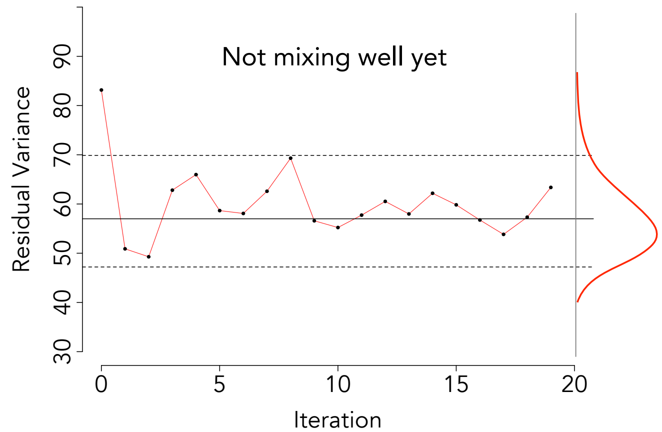
Burn-In Period

The start of the MCMC chain is what matters

How long does it take MCMC to achieve the optimal plots shown in the previous slides?

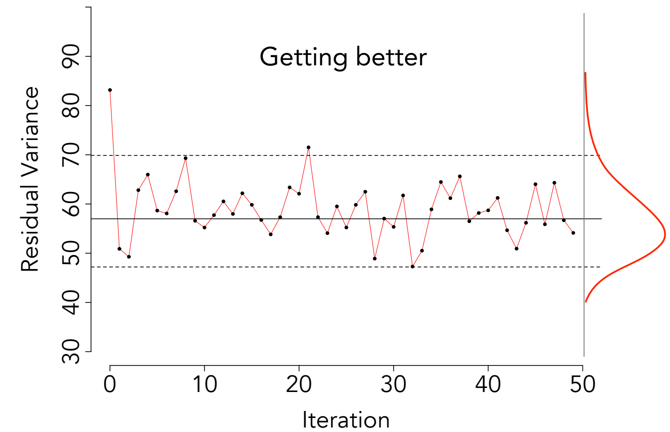
Some parameters achieve a steady state very quickly, while others converge more slowly

Trace Plot Of 20 Variance Estimates



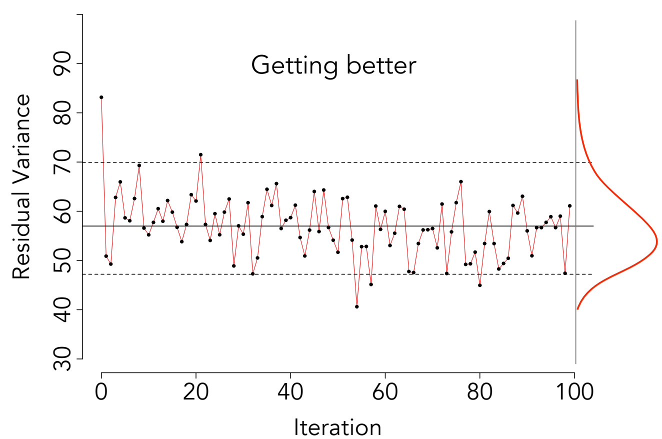
25

Trace Plot Of 50 Variance Estimates



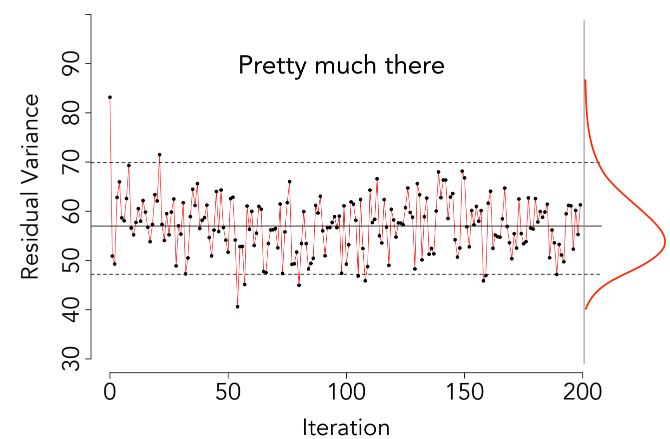
26

Trace Plot Of 100 Variance Estimates



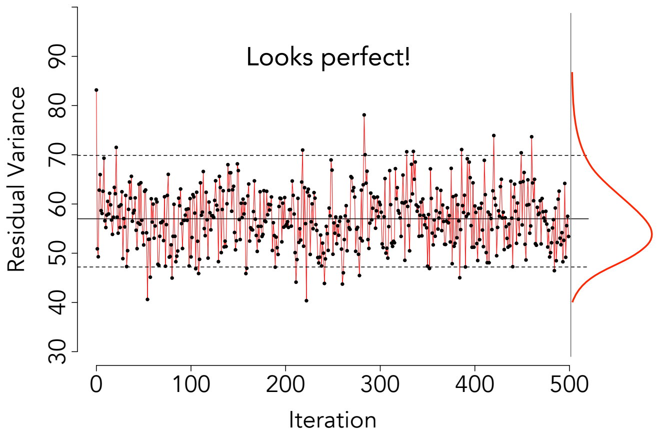
27

Trace Plot Of 200 Variance Estimates



28

Trace Plot Of 500 Variance Estimates



Numerical Convergence Diagnostics

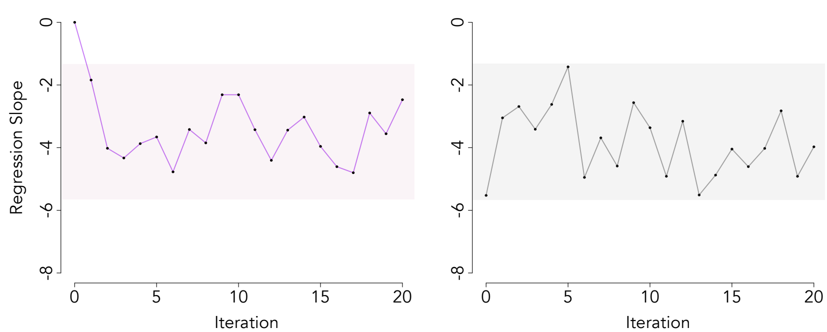
Run two or more MCMC chains

If the algorithm has achieved a steady state, two separate runs should give equivalent results

How many iterations are needed for two MCMC chains to produce the same distribution?

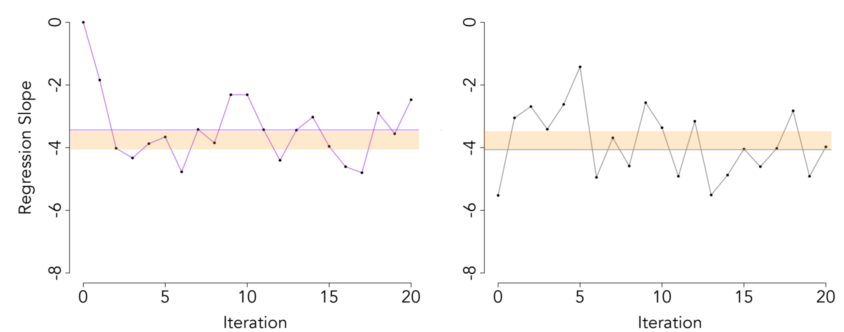
Within-Chain Variation

Estimates within each chain vary across iterations



Between-Chain Variation

The average estimate differs between chains



Potential Scale Reduction Factor (PSRF)

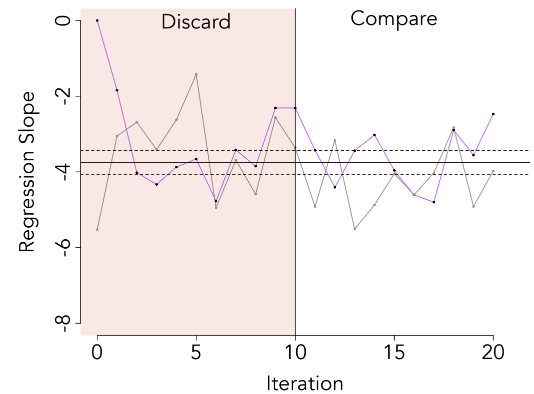
The best possible PSRF value of 1.0 occurs when there is no between-chain mean difference

$$\text{PSRF} = \sqrt{\frac{\text{Between-Chain} + \text{Within-Chain}}{\text{Within-Chain Variance}}}$$

PSRF less than 1.05 to 1.10 are considered adequate for practice

Split-Chain Comparison Method

Cut two chains in half and compare the second halves



PSRF Values For Substantive Model

Parameters converge at different rates, pay attention to the worst value across all parameters

Parameter	11 to 20	16 to 30	21 to 40	26 to 50	50 to 100	100 to 200
Intercept	1.514	1.057	1.008	1.043	1.000	1.004
Math Pre-Test	**1.521**	1.027	1.003	1.007	1.006	1.000
Self-Efficacy	1.038	1.007	**1.075**	**1.061**	1.061	1.001
Anxiety	1.197	**1.133**	1.007	1.031	1.001	1.037
Variance	1.003	1.009	1.065	1.004	**1.116**	1.001

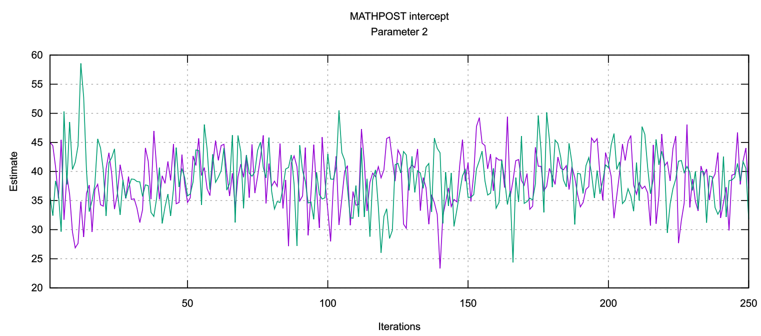
Blimp Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 4 chains    | Highest PSR | Parameter # |
                           51 to 100      |       1.083 |          10 |
                          101 to 200      |       1.018 |           5 |
                          151 to 300      |       1.015 |          10 |
                          201 to 400      |       1.013 |          14 |
                          251 to 500      |       1.015 |          14 |
                          301 to 600      |       1.014 |           4 |
                          351 to 700      |       1.005 |          11 |
                          401 to 800      |       1.008 |          11 |
                          451 to 900      |       1.005 |          11 |
                          501 to 1000     |       1.008 |           5 |
```

Blimp Trace Plots

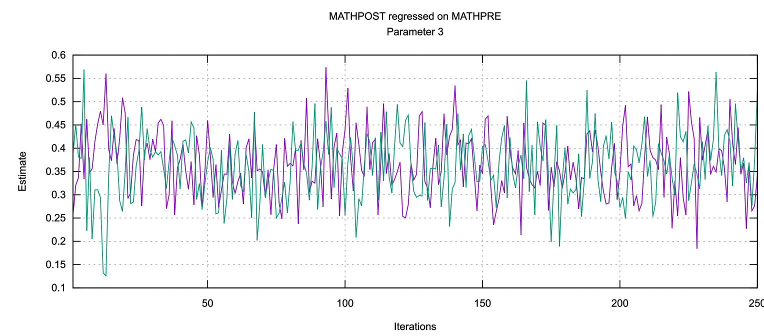
Intercept estimates from 500 iterations



37

Blimp Trace Plots

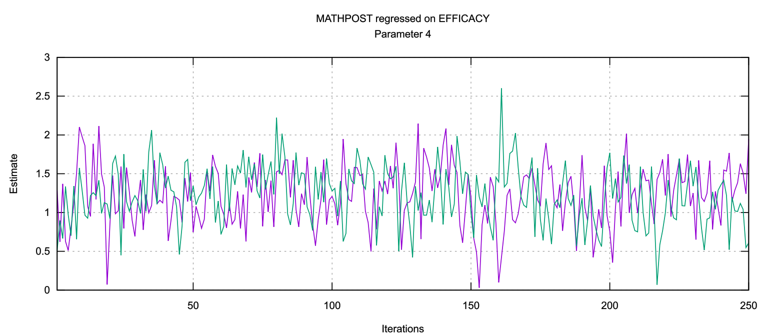
Pre-test math slope estimates from 500 iterations



38

Blimp Trace Plots

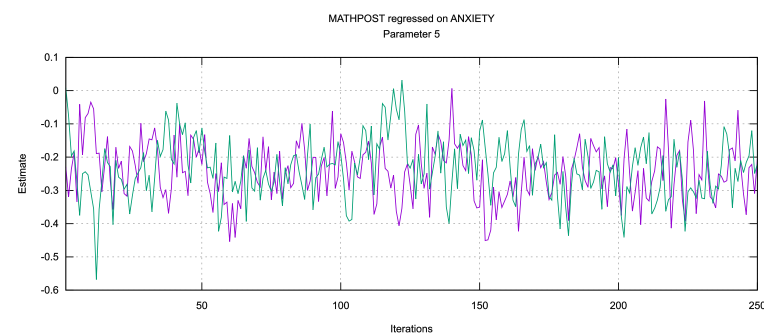
Self-efficacy slope estimates from 500 iterations



39

Blimp Trace Plots

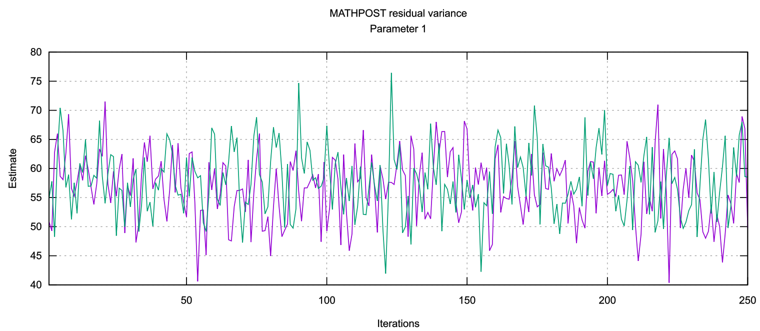
Anxiety slope estimates from 500 iterations



40

Blimp Trace Plots

Residual variance estimates from 500 iterations



Mplus Output

```
TECHNICAL 8 OUTPUT

  TECHNICAL 8 OUTPUT FOR BAYES ESTIMATION

    CHAIN    BSEED
    1        90291
    2        255458

                       POTENTIAL       PARAMETER WITH
    ITERATION    SCALE REDUCTION      HIGHEST PSR
    100              1.021              2
    200              1.013              4
    300              1.012              2
    400              1.006              2
    500              1.001              2
    600              1.001              4
    700              1.002              13
    800              1.005              13
    900              1.006              11
    1000             1.005              11
```


Take-Home Message

The MCMC algorithm converges to a steady state within relatively few iterations (< 200)

The PSRF diagnostics confirm the trace plots

Discard a conservative number of burn-in iterations and base analysis results on a large number of iterations (e.g., 10,000) after that