Multiple Imputation

Multiple Imputation

Multiple imputation generates many complete data sets (e.g., $M > 20$), each with different imputations

X	Y	Z
4	4	3
3	NA	5
7	1	6
NA	1	6
5	9	3
3	NA	NA
1	6	7
9	4	9
2	NA	6

X	Y	Z
4	4	3
3	3.3	5
7	1	6
2.4	1	6
5	9	3
3	2.1	1.9
1	6	7
9	4	9
2	5.3	6

1

X	Y	Z
4	4	3
3	4.7	5
7	1	6
1.3	1	6
5	9	3
3	6.5	3.5
1	6	7
9	4	9
2	4.2	6

2 ⋯

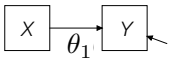
X	Y	Z
4	4	3
3	2.6	5
7	1	6
2.1	1	6
5	9	3
3	3.9	3.0
1	6	7
9	4	9
2	4.6	6

⋯ *M*

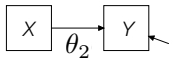
Analyzing Multiply Imputed Data

The model of interest is fit to each data set

X	Y	Z
4	4	3
3	3.3	5
7	1	6
2.4	1	6
5	9	3
3	2.1	1.9
1	6	7
9	4	9
2	5.3	6

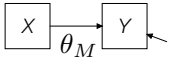


X	Y	Z
4	4	3
3	4.7	5
7	1	6
1.3	1	6
5	9	3
3	6.5	3.5
1	6	7
9	4	9
2	4.2	6



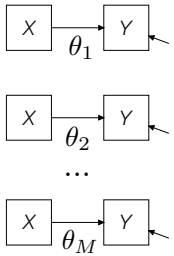
⋯

X	Y	Z
4	4	3
3	2.6	5
7	1	6
2.1	1	6
5	9	3
3	3.9	3.0
1	6	7
9	4	9
2	4.6	6



Combining Analysis Results

The pooling phase averages estimates and standard errors into a single set of analysis results



$$\hat{\theta} = (\theta_1 + \theta_2 + ... + \theta_M)\,/M$$

Benefits of Multiple Imputation

Accurate with MCAR and MAR mechanisms

Maximizes power by using all available data

Analyzing several data sets gives a method for incorporating missing data error into standard errors and significance tests

Readily accommodates different levels of measurement

Imputation Phase of Multiple Imputation

Bivariate Example

The substantive analysis is a simple regression model, where the covariate is incomplete

$$Y = \beta_0 + \beta_1 (X) + e$$

For example, efficacy to quit predicted by the number of cigarettes smoked, which is incomplete

$$Efficacy = \beta_0 + \beta_1(Cigs) + e$$

Bivariate Example

The imputation model is a reverse regression

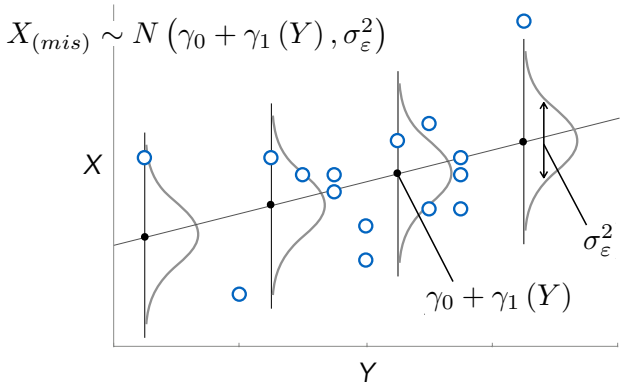
$$X_{(mis)} = \gamma_0 + \gamma_1 (Y) + \varepsilon$$

$$\phi = \left(\gamma_0, \gamma_1, \sigma_\varepsilon^2\right)$$

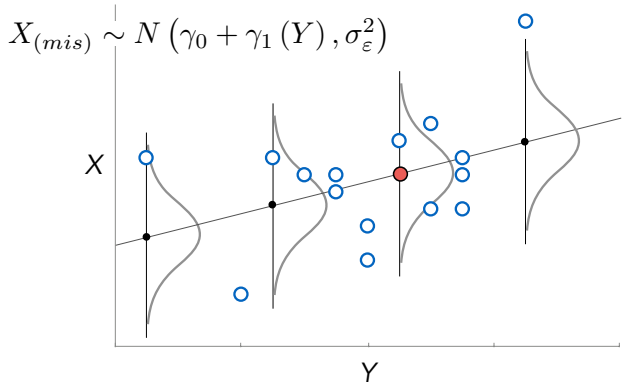
The imputation model parameters define a distribution of plausible replacement values

$$X_{(mis)} \sim N\left(\gamma_0 + \gamma_1 (Y), \sigma_\varepsilon^2\right)$$

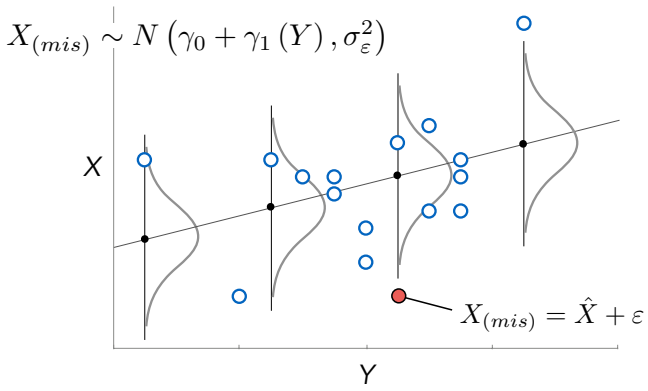
Distribution Of Missing Values



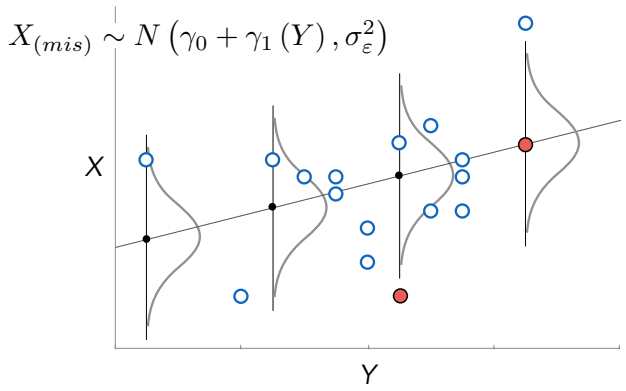
Sampling Imputations



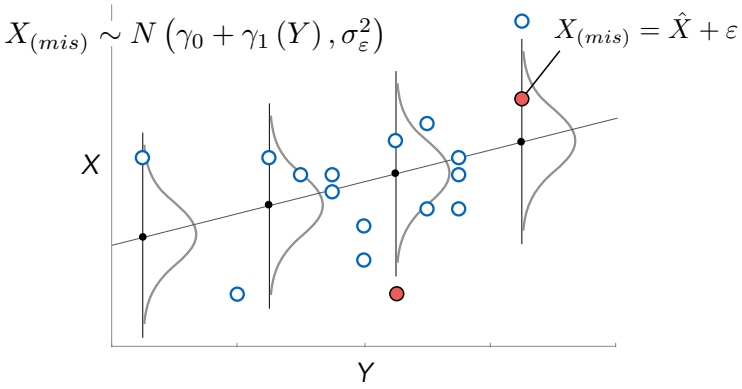
Imputation = Predicted Score + Noise



Sampling Imputations



Imputation = Predicted Score + Noise



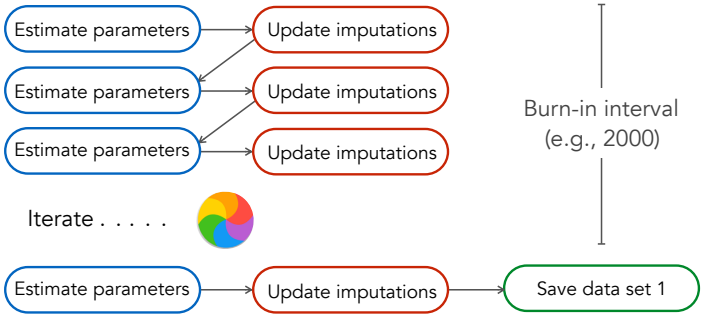
Markov Chain Monte Carlo

Each iteration *t* of the MCMC algorithm generates estimates of the regression model parameters, after which it samples imputations from a distribution based on those estimates

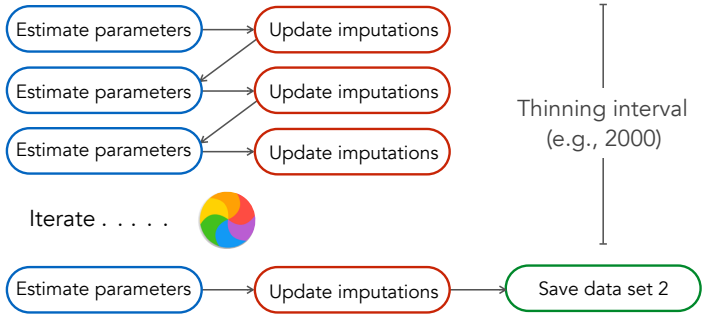
$$\phi^{(t)} \sim f\left(\gamma_0, \gamma_1, \sigma_\varepsilon^2 | Y, X^{(t-1)}\right)$$ ——— Estimation

$$X_{(mis)}^{(t)} \sim f\left(X_{(mis)} | Y, \phi^{(t)}\right)$$ ——— Imputation

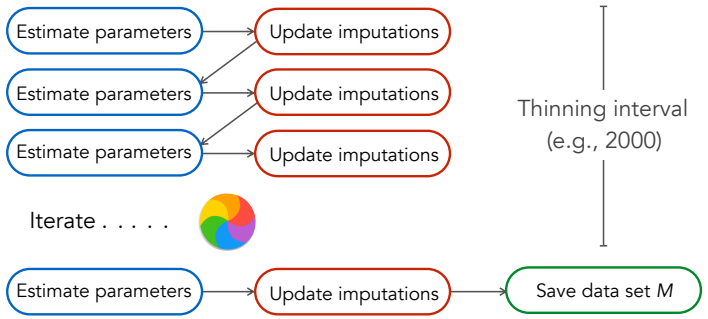
MCMC Burn-In Period



MCMC Thinning Period



Keep Thinning Until Finished



Recipe For One MCMC Iteration

Step 1: Sample regression coefficients from a distribution that conditions on the imputed data and the current residual variance estimate

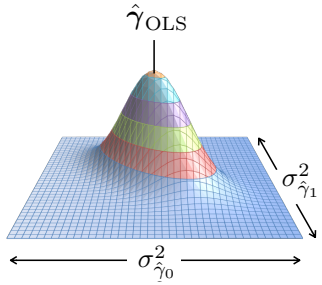
Step 2: Sample a residual variance from a distribution that conditions on the imputed data and the current regression coefficients

Step 3: Sample imputations from a distribution that conditions on the current regression coefficients and residual variance

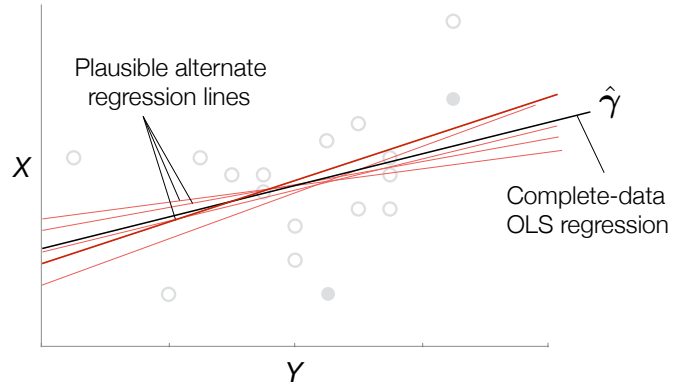
Estimating Regression Coefficients

Monte Carlo simulation "samples" coefficients from a normal distribution with center and spread equal to OLS estimates and covariance matrix

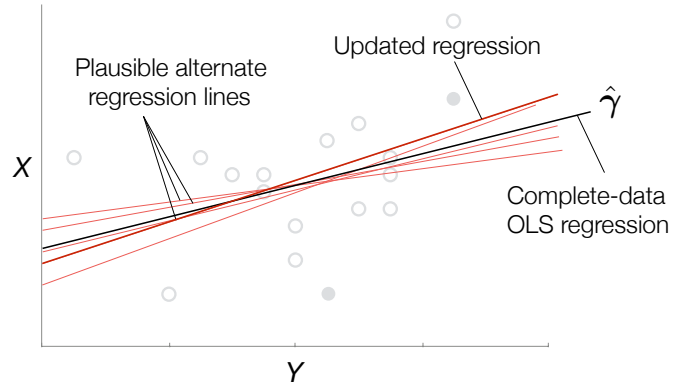
$$\boldsymbol{\gamma}^{(t)} \sim N\left(\hat{\boldsymbol{\gamma}}_{\text{OLS}}^{(t)}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}_{\text{OLS}}}\right)$$



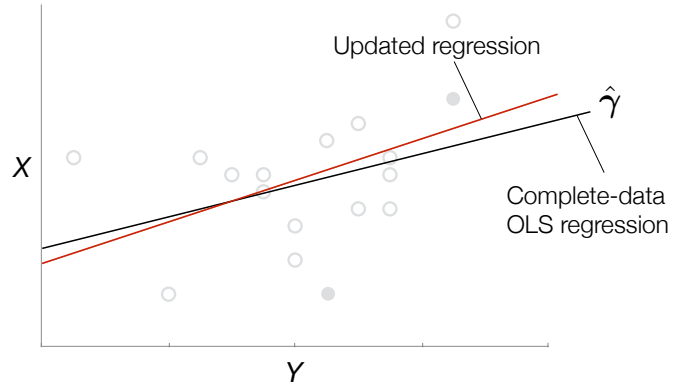
Distribution Of Plausible Regressions



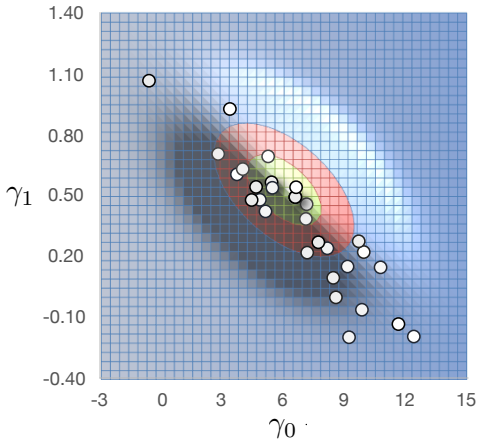
Updated Regression Line



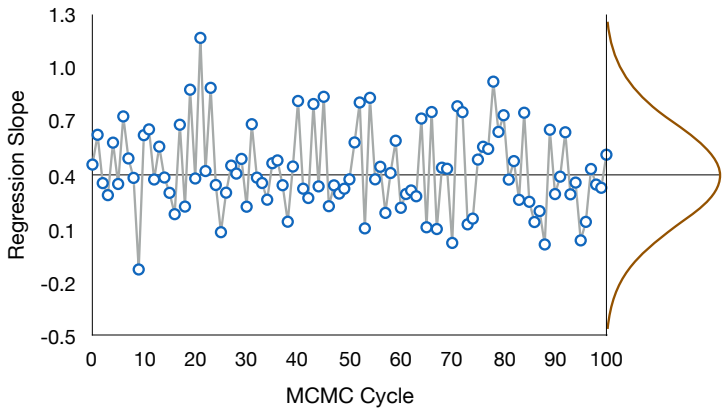
Updated Regression Line



Regression Coefficients from 30 MCMC Iterations



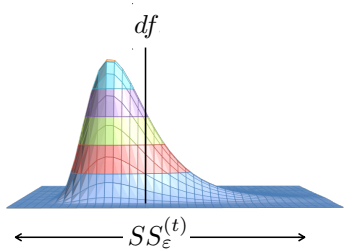
Trace Plot of Slopes from 100 MCMC Iterations



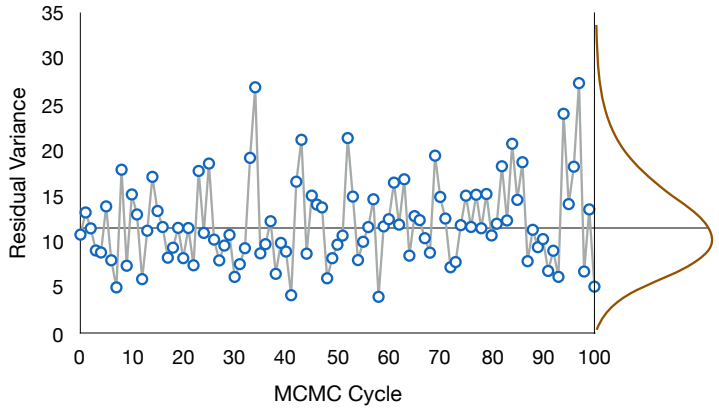
Estimating The Residual Variance

Monte Carlo simulation "samples" coefficients from an inverse gamma distribution with center and spread determined by the degrees of freedom and residual sum of squares

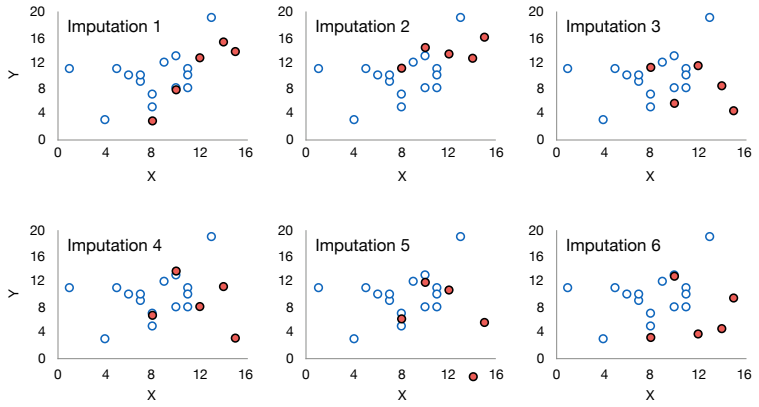
$$\sigma_{\varepsilon}^{2(t)} \sim IG\left(df, SS_{\varepsilon}^{(t)}\right)$$



Trace Plot of Residual Variances from 100 Iterations



Imputed Data Sets



Multivariate Missing Data

The substantive analysis is a multiple regression model, where any variable could be incomplete

$$Y = \beta_0 + \beta_1 (X_1) + \beta_2 (X_2) + e$$

For example, efficacy to quit predicted by years smoking and number of cigarettes smoked

$$Efficacy = \beta_0 + \beta_1 (Years) + \beta_2 (Cigs) + e$$

Fully Conditional Specification (FCS)

Fully conditional specification (FCS) imputes variables one at a time in a sequence

FCS uses a series of univariate imputation models, each of which consists of an estimation step that updates parameters and an imputation step

An imputed variable from one model functions as a predictor in all other models

Y Imputation Model

The *Y* imputation model conditions on the current values of X_1 and X_2

$$Y_{(mis)}^{(t)} = \gamma_{0(Y)} + \gamma_{1(Y)}\left(X_1^{(t-1)}\right) + \gamma_{2(Y)}\left(X_2^{(t-1)}\right) + \varepsilon_{(Y)}$$

Imputations are drawn from a normal distribution

$$Y_{(mis)}^{(t)} \sim N\left(\gamma_{0(Y)} + \gamma_{1(Y)}\left(X_1^{(t-1)}\right) + \gamma_{2(Y)}\left(X_2^{(t-1)}\right), \sigma^2_{\varepsilon(Y)}\right)$$

X_1 Imputation Model

The X_1 imputation model conditions on the current values of *Y* and X_2

$$X_{1(mis)}^{(t)} = \gamma_{0(X_1)} + \gamma_{1(X_1)}\left(Y^{(t)}\right) + \gamma_{1(X_1)}\left(X_2^{(t-1)}\right) + \varepsilon_{(X_1)}$$

Imputations are drawn from a normal distribution

$$X_{1(mis)}^{(t)} \sim N\left(\gamma_{0(X_1)} + \gamma_{1(X_1)}\left(Y^{(t)}\right) + \gamma_{1(X_1)}\left(X_2^{(t-1)}\right), \sigma^2_{\varepsilon(X_1)}\right)$$

X_2 Imputation Model

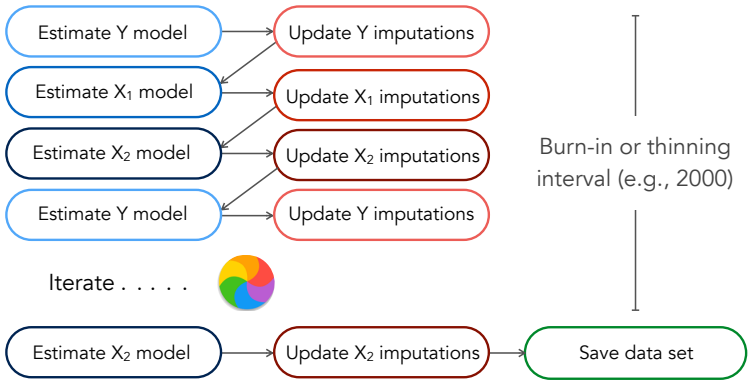
The X_2 imputation model conditions on the current values of *Y* and X_1

$$X_{2(mis)}^{(t)} = \gamma_{0(X_2)} + \gamma_{1(X_2)}\left(Y^{(t)}\right) + \gamma_{1(X_2)}\left(X_1^{(t)}\right) + \varepsilon_{(X_2)}$$

Imputations are drawn from a normal distribution

$$X_{2(mis)}^{(t)} \sim N\left(\gamma_{0(X_2)} + \gamma_{1(X_2)}\left(Y^{(t)}\right) + \gamma_{1(X_2)}\left(X_1^{(t)}\right), \sigma^2_{\varepsilon(X_2)}\right)$$

MCMC Steps



MCMC Convergence

MCMC gives parameters that constantly change across iterations, estimates never converge on a single value

MCMC converges when the parameter values achieve a stable distribution (the mean and variance of the distribution no longer changes across iterations)

Determining the number of cycles needed to achieve convergence is an important diagnostic step

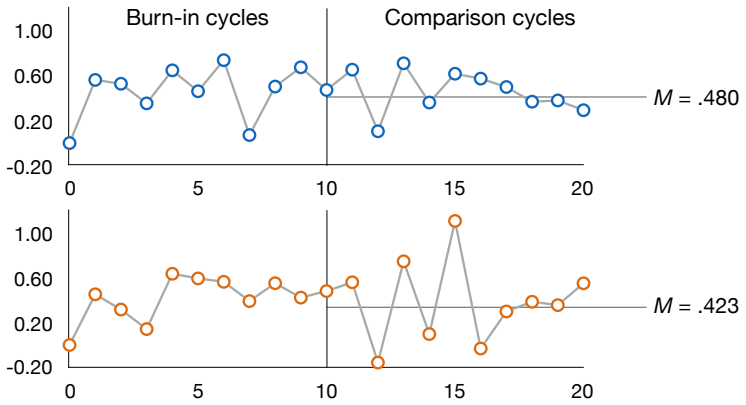
Evaluating Convergence

Run MCMC twice and determine how many iterations are needed for the two runs to give similar distributions

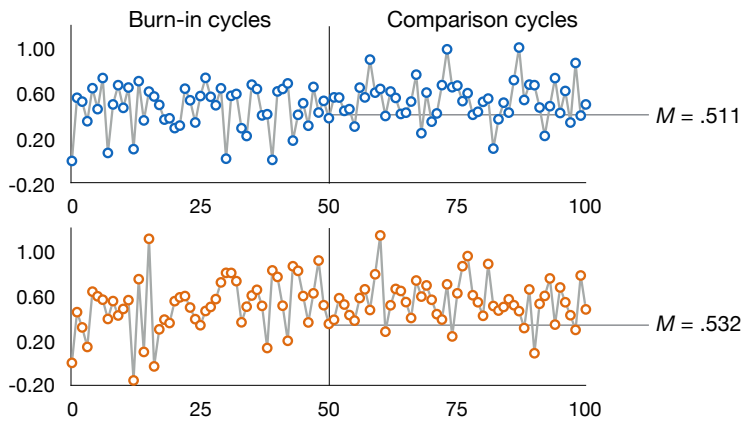
Similarity is assessed by looking at mean differences between the two runs

The potential scale reduction (PSR) factor is a common diagnostic based on the mean parameter values from two or more separate MCMC chains

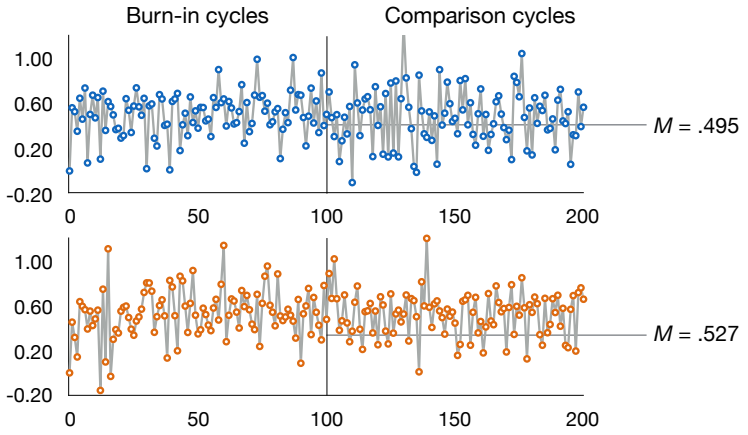
Two Chains after 20 Iterations



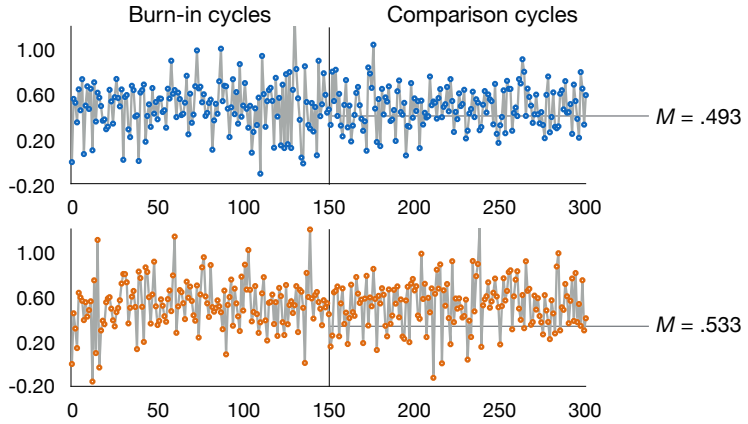
Two Chains after 100 Iterations



Two Chains after 200 Iterations



Two Chains after 300 Iterations



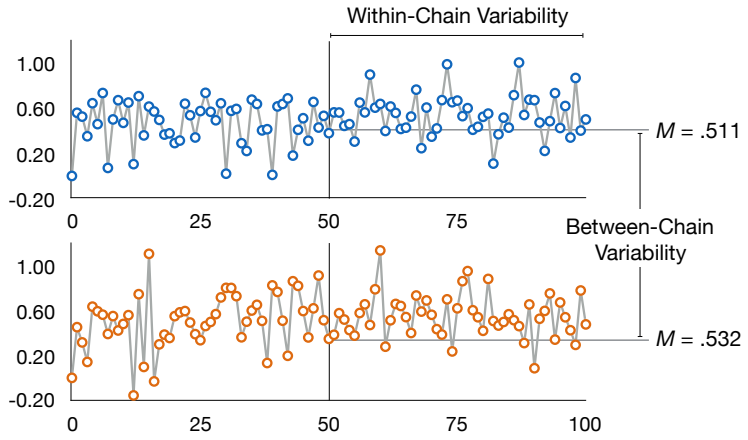
Reading the Trace Plots

The chains have noticeably different means at iteration 20, but the mean difference dissipates by cycle 100

The distribution of synthetic slope values stabilizes (converges) by 100 iterations and additional iterations do not affect the mean difference

The burn-in and/or thinning interval can safely be set to a value of 100 or greater

Within- and Between-Chain Variation



Potential Scale Reduction (PSR) Factor

The potential scale reduction (PSR) factor (Gelman & Rubin, 1992) uses ANOVA mean squares formulas to compute between- and within-chain variance

$$PSR = \sqrt{\frac{\sigma^2_{Within} + \sigma^2_{Between}}{\sigma^2_{Within}}} = \sqrt{\frac{1}{1 - R^2}}$$

The between-chain variance decreases as mean differences diminish, giving PSR values closer to 1

Recommendations

Parameters converge at different rates, so PSR values vary across parameters

Determine the number of MCMC iterations required for the worst (highest) PSR to fall below 1.05 to 1.10 (common rules of thumb in the literature)

Set the burn-in and thinning intervals to a value that exceeds the number of iterations required to converge

Analysis Example

Motivating Example

Analysis involving number of years smoking (Years), number of cigarettes smoked per week (Cigs), and self-efficacy to quit smoking (Efficacy)

Imputation assumes that values are missing at random

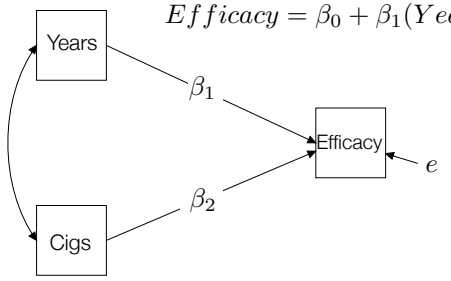
What does MAR require in this context?

Years	Cigs	Efficacy
7	9	NA
8	NA	NA
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	NA	10
15	NA	11
5	11	13
9	12	11
11	11	16
14	NA	10
13	19	9
12	NA	5
11	8	7
10	13	10
10	8	NA
7	10	7
11	10	6

Analysis Model

The analysis model is a multiple regression predicting self-efficacy to quit based on years smoking and number of cigarettes smoked

$$Efficacy = \beta_0 + \beta_1(Years) + \beta_2\,(Cigs) + e$$



Imputation Highlights

The analysis model is linear and additive (e.g., no interactions), so FCS imputation based on reverse regression models is appropriate

e.g., Number of cigarettes is imputed from a linear regression with efficacy and years smoking as predictors

The imputation process should include all variables in the analysis, although extra variables can be used

Ex3.1.imp
Blimp Diagnostic Script

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
  efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 3000;
THIN: 1;
NIMPS: 2;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```

Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

Comparing iterations 51 to 100 for 2 chains.
                    ---------------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    ---------------------------------------------------------
          Max PSR   |       1.027|         nan|       1.008|         nan|
  Missing Variable  |        cigs|            |        cigs|            |
                    ---------------------------------------------------------

 Comparing iterations 101 to 200 for 2 chains.
                    ---------------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    ---------------------------------------------------------
          Max PSR   |       1.043|         nan|       1.002|         nan|
  Missing Variable  |    efficacy|            |        cigs|            |
                    ---------------------------------------------------------
```

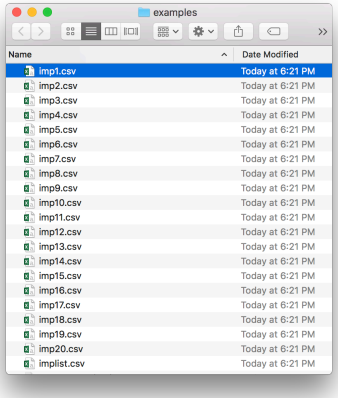
Ex3.2.imp
Blimp Imputation Script (Mplus Format)

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
  efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 100;
THIN: 100;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Imputed Data Sets

The imputation phase generates a set of imputed data sets

The next step is to analyze the data …



Ex3.3.imp
Blimp Imputation Script (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
  efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 100;
THIN: 100;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Multiple Imputation Analysis and Pooling Phase

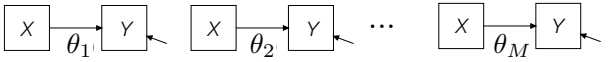
Analyzing Multiply Imputed Data

The model of interest is fit to each data set

X	Y	Z
4	4	3
3	**3.3**	5
7	1	6
2.4	1	6
5	9	3
3	**2.1**	**1.9**
1	6	7
9	4	9
2	**5.3**	6

X	Y	Z
4	4	3
3	**4.7**	5
7	1	6
1.3	1	6
5	9	3
3	**6.5**	**3.5**
1	6	7
9	4	9
2	**4.2**	6

X	Y	Z
4	4	3
3	**2.6**	5
7	1	6
2.1	1	6
5	9	3
3	**3.9**	**3.0**
1	6	7
9	4	9
2	**4.6**	6



Analysis Phase

The multiple regression analysis is fit to each imputed data set (e.g., 20 multiple regression analyses)

$$Efficacy = \beta_0 + \beta_1(Years) + \beta_2\,(Cigs) + e$$

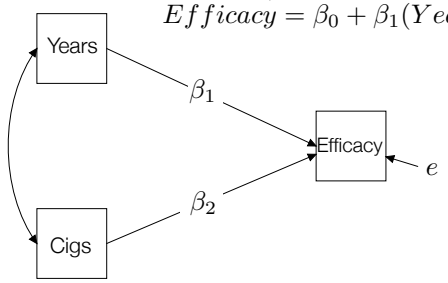


Illustration With Three Imputed Data Sets

Years	Cigs	Efficacy
7	9	**15.50**
8	**8.96**	**15.78**
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	**9.92**	10
15	**13.62**	11
5	11	13
9	12	11
11	11	16
14	**14.42**	10
13	19	9
12	**18.04**	5
11	8	7
10	13	10
10	8	**9.18**
7	10	7
11	10	6

Years	Cigs	Efficacy
7	9	**15.38**
8	**8.28**	**9.25**
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	**13.41**	10
15	**6.99**	11
5	11	13
9	12	11
11	11	16
14	**15.31**	10
13	19	9
12	**12.75**	5
11	8	7
10	13	10
10	8	**14.85**
7	10	7
11	10	6

Years	Cigs	Efficacy
7	9	**5.86**
8	**10.04**	**13.88**
1	11	16
4	3	21
6	10	17
8	5	10
8	7	13
10	**12.95**	10
15	**14.40**	11
5	11	13
9	12	11
11	11	16
14	**14.47**	10
13	19	9
12	**11.46**	5
11	8	7
10	13	10
10	8	**6.79**
7	10	7
11	10	6

Pooling Estimates

The multiple imputation point estimate is the arithmetic average of the *M* complete-data estimates

Estimate from one data set

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M}$$

Pooled estimate

Number of data sets

Pooling Descriptive Statistics

Data Set 1

	M	SD	N
Years	9.00	3.45	20
Cigs	10.59	3.84	20
SE	11.62	4.19	20

Data Set 2

	M	SD	N
Years	9.00	3.45	20
Cigs	10.19	3.61	20
SE	11.57	4.14	20

Data Set 3

	Years	Cigs	SE
Years	9.00	3.45	20
Cigs	10.52	3.51	20
SE	10.93	4.28	20

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M} = \frac{10.59 + 10.19 + 10.52}{3} = 10.43$$

Pooled Estimates

	M	SD	N
Years	9.00	3.45	20
Cigs	10.43	3.65	20
SE	11.37	4.20	20

Pooling Correlations

Data Set 1

	Years	Cigs	SE
Years	1.00		
Cigs	0.54	1.00	
SE	-0.60	-0.45	1.00

Data Set 2

	Years	Cigs	SE
Years	1.00		
Cigs	0.38	1.00	
SE	-0.57	-0.38	1.00

Data Set 3

	Years	Cigs	SE
Years	1.00		
Cigs	0.54	1.00	
SE	-0.52	-0.26	1.00

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M} = \frac{-.45 - .57 - .26}{3} = -.37$$

Pooled Estimates

	Years	Cigs	SE
Years	1.00		
Cigs	0.49	1.00	
SE	-0.57	-0.37	1.00

Pooling Regression Parameters

	Data Set 1	Data Set 2	Data Set 3	Pooled
B_0 (Intercept)	19.20	19.16	18.30	18.88
B_1 (Years)	-0.62	-0.59	-0.63	-0.61
B_2 (Cigarettes)	-0.19	-0.22	0.04	-0.12
Residual Variance	12.07	12.41	14.81	13.10
R^2	0.39	0.35	0.28	0.34

$$\hat{\theta} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M} = \frac{-.19 - .22 + .04}{3} = -.12$$

Pooling Standard Errors

Averaging standard errors underestimates sampling variability because the component standard errors are based on complete data sets

The pooling procedure incorporates an adjustment that reflects the additional error from missing data

The within-imputation component estimates complete-data sampling error, and the between-imputation component is additional missing data error

Within-Imputation Variance

The within-imputation variance is the average squared standard error (sampling variance)

$$Var_W = \frac{\sum_{m=1}^{M} SE_m^2}{M}$$

Within-imputation variance estimates sampling error in the hypothetically complete data

Within-Imputation Variance Example

Data Set 1

	Est.	SE	SE²
B_0	19.20	2.559	6.548
B_1	-0.62	0.275	0.076
B_2	-0.19	0.247	0.061

Data Set 2

	Est.	SE	SE²
B_0	19.16	2.762	7.629
B_1	-0.59	0.253	0.064
B_2	-0.22	0.242	0.059

Data Set 3

	Est.	SE	SE²
B_0	16.54	2.970	8.821
B_1	-0.67	0.304	0.092
B_2	0.04	0.299	0.089

$$Var_W = \frac{\sum_{m=1}^{M} SE_m^2}{M} = \frac{.061 + .059 + .089}{3} = .097$$

Pooled Estimates

	Est.	Var_W
B_0	18.30	10.745
B_1	-0.63	0.080
B_2	-0.12	0.097

Between-Imputation Variance

Variability in the estimates across the M data sets is a result of using different imputations (missing data error)

$$Var_B = \frac{\sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2}{M - 1}$$

Between-imputation variance applies the sample variance formula to the *M* estimates

Between-Imputation Variance Example

Data Set 1

	Est.	SE	SE²
B_0	19.20	2.559	6.548
B_1	-0.62	0.275	0.076
B_2	-0.19	0.247	0.061

Data Set 2

	Est.	SE	SE²
B_0	19.16	2.762	7.629
B_1	-0.59	0.253	0.064
B_2	-0.22	0.242	0.059

Data Set 3

	Est.	SE	SE²
B_0	16.54	2.970	8.821
B_1	-0.67	0.304	0.092
B_2	0.04	0.299	0.089

$$Var_B = \frac{\sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2}{M-1}$$

$$= \frac{(-.19-.12)^2 + (-.22+.12)^2 + (.04+.12)^2}{3-1} = .021$$

Pooled Estimates

	Est.	Var_W	Var_B
B_0	18.30	10.745	2.311
B_1	-0.63	0.080	0.002
B_2	-0.12	0.097	0.021

Pooled Standard Error

The pooled standard error combines complete-data sampling error and additional missing data error

$$SE = \sqrt{Var_W + Var_B + \frac{Var_B}{M}}$$

Complete data sampling error

Missing data error

Sampling error of the pooled estimate

Standard Error Example

	Est.	Var_W	Var_B	SE
B_0	18.30	10.745	2.311	3.72
B_1 (Years)	-0.63	0.080	0.002	0.29
B_2 (Cigs)	-0.12	0.097	0.021	0.35

$$SE = \sqrt{Var_W + Var_B + \frac{Var_B}{M}}$$

$$= \sqrt{.097 + .021 + \frac{.012}{3}} = .35$$

Significance Test

The single-parameter test statistic is based on the pooled estimate and standard error

Pooled estimate

Hypothesized value (e.g., 0)

$$t \ (or \ z) = \frac{\hat{\theta} - \theta_0}{SE}$$

Software packages use different reference distributions for *p*-values (e.g., *t* with various *df* adjustments, *z*)

Degrees of Freedom Adjustments

Rubin's (1987) original degrees of freedom expression for the *t* statistic often exceed the sample size

Use Barnard and Rubin (1999) or Reiter (2007) expressions for degrees of freedom when possible

Degrees of freedom expressions are complex functions of the complete-data degrees of freedom, sample size, missing data rates, and correlations among the variables

p-Value Comparison

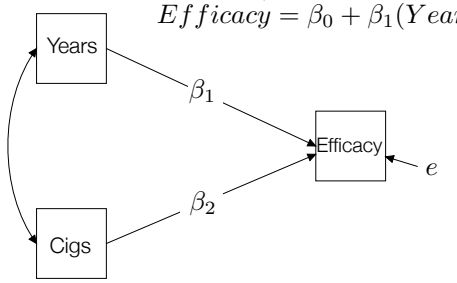
	Est.	SE	Est ÷ SE	df	p
***t* Distribution with Rubin (1987) degrees of freedom**					
B_0	18.16	2.95	6.17	1498.59	< .01
B_1 (Years)	-0.64	0.28	-2.31	8416.41	0.01
B_2 (Cigs)	-0.10	0.26	-0.38	1414.63	0.35
***t* Distribution with Barnard and Rubin (1999) degrees of freedom**					
B_0	18.16	2.95	6.17	13.46	< .01
B_1 (Years)	-0.64	0.28	-2.31	14.55	0.02
B_2 (Cigs)	-0.10	0.26	-0.38	13.40	0.36
Normal z Distribution					
B_0	18.16	2.95	6.17	NA	< .01
B_1 (Years)	-0.64	0.28	-2.31	NA	0.01
B_2 (Cigs)	-0.10	0.26	-0.38	NA	0.35

Analysis Example

Analysis Model

The analysis model is a multiple regression predicting self-efficacy to quit based on years smoking and number of cigarettes smoked

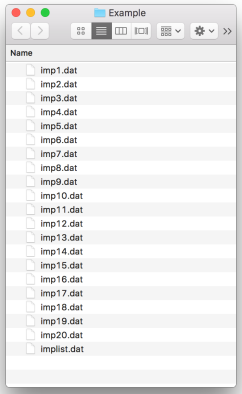
$$Efficacy = \beta_0 + \beta_1(Years) + \beta_2\,(Cigs) + e$$



Mplus Imputation Format

Mplus requires each imputed data set as a separate file

Blimp creates a list file (e.g., implist.dat) containing the names of the data sets, and this file serves as the input data for subsequent analyses



Ex3.2.imp
Blimp Imputation Script (Mplus Format)

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
  efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 100;
THIN: 100;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Blimp Output

```
-------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

   id quitmeth male age years cigs heavycig efficacy stress

-------------------------------------------------------------
```

Ex3.4.inp
Mplus Analysis Script

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = id quitmeth male age years
  cigs heavycig efficacy stress;
usevariables = years cigs efficacy;
MODEL:
efficacy on years (b1)
  cigs (b2);
MODEL TEST:
b1 = 0; b2 = 0;
OUTPUT:
standardized(stdyx);
```

Mplus Analysis Output

```
MODEL RESULTS

                                                    Two-Tailed    Rate of
                  Estimate       S.E.  Est./S.E.    P-Value    Missing

EFFICACY ON
   YEARS            -0.652      0.269     -2.423      0.015      0.129
   CIGS             -0.069      0.278     -0.249      0.803      0.322

Intercepts
   EFFICACY         17.815      2.818      6.323      0.000      0.178

Residual Variances
   EFFICACY         11.240      3.972      2.830      0.005      0.185
```

Mplus Analysis Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

                                                    Two-Tailed    Rate of
                  Estimate       S.E.  Est./S.E.    P-Value    Missing

EFFICACY ON
   YEARS            -0.540      0.201     -2.684      0.007      0.185
   CIGS             -0.055      0.254     -0.216      0.829      0.351

Intercepts
   EFFICACY          4.393      0.735      5.975      0.000      0.263

Residual Variances
   EFFICACY          0.677      0.178      3.791      0.000      0.087
```

Ex3.3.imp
Blimp Imputation Script (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
  efficacy stress;
MISSING: -99;
MODEL: ~ years cigs efficacy;
SEED: 90291;
BURN: 100;
THIN: 100;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Blimp Output

```
-------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

  imp# id quitmeth male age years cigs heavycig efficacy stress

-------------------------------------------------------------
```

Ex3.5.r
R Analysis Script

```
# Required packages
library(mitml)

# Read data
filepath <- "~/desktop/examples/imps.csv"
impdata <- read.csv(filepath, header = F)
names(impdata) <- c("imputation", "id", "quitmeth", "male", "age",
  "years", "cigs", "heavycig", "efficacy", "stress")

# Analyze data and pool estimates
implist <- as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lm(efficacy ~ years + cigs))
estimates <- testEstimates(analysis, var.comp = T, df.com = 17)
estimates

# Test full model with Wald test
emptymodel <- with(implist, lm(efficacy ~ 1))
testModels(analysis, emptymodel, method = "D1")
```

R Analysis Output

```
Final parameter estimates and inferences obtained from 20 imputed data sets.

             Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
(Intercept)    17.815     3.016     5.907    12.760     0.000     0.180     0.155
years          -0.652     0.289    -2.255    13.497     0.041     0.124     0.111
cigs           -0.069     0.294    -0.235    10.514     0.818     0.391     0.287

                  Estimate
Residual~~Residual   13.224

Hypothesis test adjusted for small samples with df=[17]
complete-data degrees of freedom.
```

R Analysis Output

```
Model comparison calculated from 20 imputed data sets.
Combination method: D1

   F.value     df1     df2   P(>F)     RIV
     3.215       2 948.316   0.041   0.222

Unadjusted hypothesis test as appropriate in larger samples.
```

Ex3.6.sps
SPSS Analysis Script

```
data list free file = '/users/craig/desktop/examples/imps.csv'
 /imputation_ id quitmeth male age years cigs heavycig
  efficacy stress.
exe.

* Initiate pooling routines.
sort cases by imputation_.
split file layered by imputation_.

* Analysis and pooling.
regression
  /descriptives mean stddev corr sig n
  /dependent efficacy
  /method enter years cigs.
```

SPSS Analysis Output

Coefficients[a]

imputation_	Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1.00	1	(Constant)	18.639	2.931		6.359	.000
		years	-.645	.278	-.519	-2.318	.033
		cigs	-.106	.279	-.085	-.381	.708
...							
20.00	1	(Constant)	17.606	3.246		5.424	.000
		years	-.578	.282	-.454	-2.053	.056
		cigs	-.135	.286	-.105	-.473	.642
Pooled	1	(Constant)	17.815	3.016		5.907	.000
		years	-.652	.289		-2.255	.024
		cigs	-.069	.294		-.235	.814

a. Dependent Variable: efficacy

Ex3.7.do
Stata Analysis Script

```
// Import and save original data
import delimited "~/desktop/examples/smoking.csv"
rename (v1 - v9)(id quitmeth male age years cigs heavycig
  efficacy stress)
generate imp=0

// Recode missing values
foreach var of varlist id - stress {
     replace `var' = . if `var'== -99
}
save original, replace

// Import and save imputed data
clear
import delimited "~/desktop/examples/imps.csv"
rename (v1 - v10)(imp id quitmeth male age years cigs heavycig
  efficacy stress)
save imputed, replace
```

Ex3.7.do
Stata Analysis Script

```
// Append original and imputed data
use original, clear
append using imputed

// Convert to mi data
mi import flong, m(imp) id(id) imputed(quitmeth - stress) clear

// Analyze data and pool results
mi estimate, cmdok: regress efficacy years cigs
```

Stata Analysis Output

```
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =         20
                                                Average RVI       =     0.1942
                                                Largest FMI       =     0.3196
                                                Complete DF       =         17
DF adjustment:   Small sample                   DF:     min       =      10.51
                                                        avg       =      12.26
                                                        max       =      13.50
Model F test:       Equal FMI                   F(   2,   14.0)   =       3.22
Within VCE type:          OLS                   Prob > F          =     0.0709

------------------------------------------------------------------------------
    efficacy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   -.651715   .2889468    -2.26   0.041    -1.273616   -.0298138
        cigs |   -.069252   .2941369    -0.24   0.818    -.7203108    .5818068
       _cons |    17.8151   3.015846     5.91   0.000     11.28726    24.34295
------------------------------------------------------------------------------
```