13. Model-Based Multiple Imputation

Model-Based Imputation

Standard imputation routines (e.g., FCS) are appropriate for additive regression models

They are known to produce bias when applied to analyses with interactive or non-linear effects

Use Bayesian estimation to tailor imputations around a specific analysis model of interest

Moderated Regression Revisited

Regression model with a product term

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i M_i + \varepsilon_i = E(Y|X) + \varepsilon_i$$
$$Y_i \sim N\left(E(Y|X, M, XM), \sigma_\varepsilon^2\right)$$

β_1 is the influence of X when M equals zero (i.e., a conditional effect), and β_3 captures the change in the β_1 slope for every one-unit increase in M

Chronic Pain Data

Pain-related data for 275 chronic pain patients

The data include psychological correlates of pain severity such as depression, pain interference with daily life, perceived control over pain, stress, and psychosocial disability

pain.dat

Variable	Name	Missing %	Scaling
Patient identifier	ID	0	Integer index
Gender	MALE	0	0 = female, 1 = male
Age	AGE	0	Continuous
Education level	EDUGROUP	0	1 = Some college or less, 2 = college, 3 = Post-BA
Work hours per week	WORKHRS	11.7	Continuous
Exercise	EXERCISE	1.7	8-point ordinal scale
Pain intensity rating	PAIN	7.3	1 = none/little, 2 = moderate, 3 = severe/very severe
Anxiety	ANXIETY	6.0	Continuous
Stress	STRESS	0	7-point ordinal scale
Perceived control over	CONTROL	0	Continuous
Pain interference with life	INTERFERE	13.3	Continuous
Depression	DEPRESS	13.3	Continuous
Psychosocial disability	DISABILITY	3.0	Continuous

Substantive Example

Does the influence of depression on psychosocial disability differ for males and females?

$$DISABILITY_i = \beta_0 + \beta_1(DEPRESS_i) + \beta_2(MALE_i) + \beta_3(DEPRESS_i)(MALE_i) + \varepsilon_i$$

Psychosocial disability measures pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc.

Indicator And Auxiliary Variable Correlations

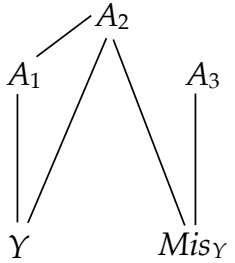
	Depression Missing Indicator	Depression
DISABILITY	0.08	0.45
MALE	-0.14	0.07
AGE	-0.08	-0.19
ANXIETY	0.03	**0.56**
STRESS	-0.04	**0.51**
CONTROL	-0.08	**-0.35**
INTERFERE	-0.09	**0.33**

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



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

Blimp Model-Based Imputation Script For Analysis In Mplus

```
DATA: pain.dat;
VARIABLES: id male age edugroup workhrs exercise pain anxiety
  stress control interfere depress disability;
ORDINAL: male stress;
FIXED: male;
MISSING: 999;
MODEL: disability ~  depress male depress*male anxiety  stress control interfere;
SEED: 90291;
NIMPS: 20;
BURN: 4000;
THIN: 2000;
CHAINS: 4 processors 4;
OPTIONS: psr;
SAVE: separate = imps_*.dat;
```


Blimp Model-Based Imputation Script For Analysis In R, SAS, SPSS, Stata

```
DATA: pain.dat;
VARIABLES: id male age edugroup workhrs exercise pain anxiety
  stress control interfere depress disability;
ORDINAL: male stress;
FIXED: male;
MISSING: 999;
MODEL: disability ~  depress male depress*male anxiety  stress control interfere;
SEED: 90291;
NIMPS: 20;
BURN: 4000;
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 age edugroup workhrs exercise pain anxiety stress control
    interfere depress disability
```

Separate file format (Mplus)

```
VARIABLE ORDER IN SAVED DATA:

   id male age edugroup workhrs exercise pain anxiety stress control
    interfere depress disability
```

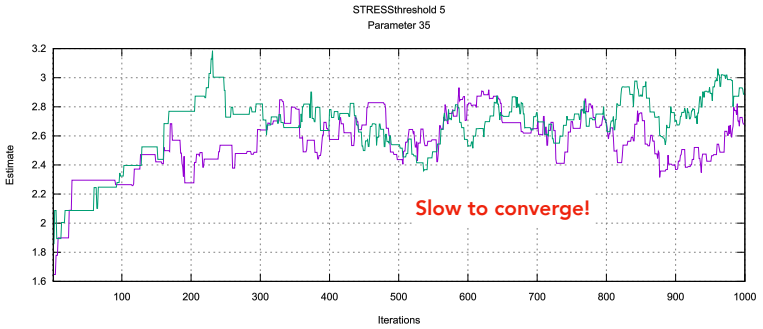
Blimp Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 4 chains   | Highest PSR | Parameter # |
                          51 to 100      |       3.517 |          33 |
                         101 to 200      |       1.411 |          33 |
                                         ...
                        1501 to 3000     |       1.138 |          35 |
                        1551 to 3100     |       1.113 |          35 |
                        1601 to 3200     |       1.087 |          35 |
                        1651 to 3300     |       1.075 |          36 |
                        1701 to 3400     |       1.060 |          36 |
                        1751 to 3500     |       1.044 |          36 |
                        1801 to 3600     |       1.035 |          36 |
                        1851 to 3700     |       1.030 |          36 |
                        1901 to 3800     |       1.029 |          36 |
                        1951 to 3900     |       1.029 |          36 |
                        2001 to 4000     |       1.035 |          36 |
```

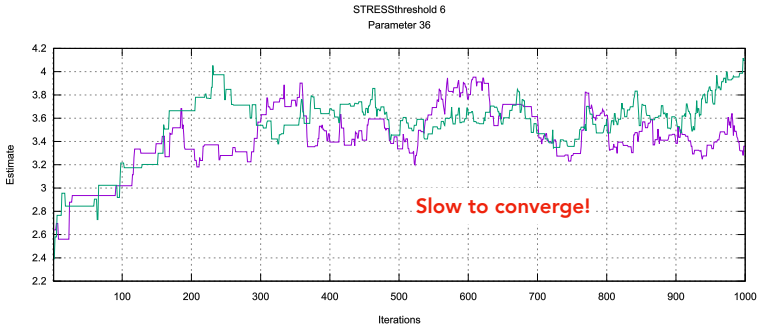

Blimp Trace Plots

Thresholds from 1000 iterations (parameter #35)



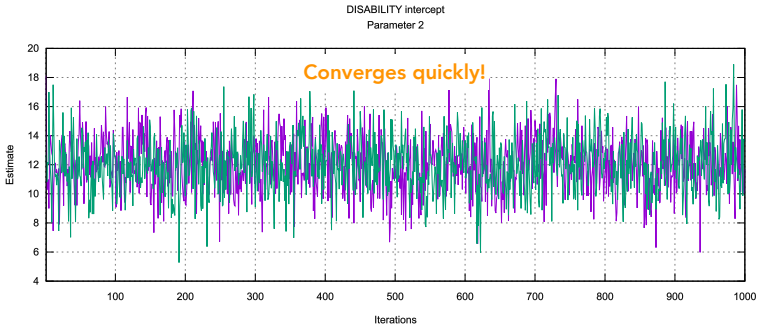
Blimp Trace Plots

Thresholds from 1000 iterations (parameter #36)



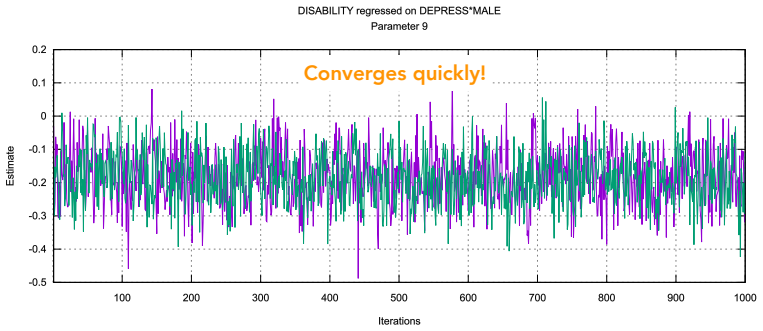
Blimp Trace Plots

Intercept estimates from 1000 iterations



Blimp Trace Plots

Interaction term estimates from 1000 iterations



17

Summary Of Multiple Imputation Estimates

Analysis results from 20 imputed data sets

Parameter	Est.	*SE*	*z*	*p*
Intercept	16.98	0.30	56.03	< .001
DEPRESS slope	0.40	0.05	7.39	< .001
MALE slope	-0.35	0.48	-0.72	0.470
DEPRESS*MALE	-0.19	0.08	-2.35	0.019

18

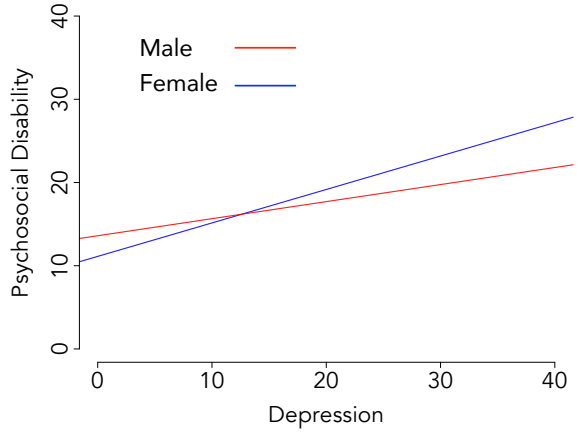
Bayesian Analysis Revisited

Bayesian analysis without auxiliary variables

Parameter	Mean	Std. Dev.	Lower 2.5%	Upper 97.5%
Intercept	16.97	0.35	16.28	17.67
DEPRESS slope	0.42	0.06	0.31	0.53
MALE slope	-0.35	0.50	-1.31	0.65
DEPRESS*MALE	-0.21	0.08	-0.37	-0.06

19

Simple Slopes



20

Interpretations

Depression is grand mean centered prior to computing the product term

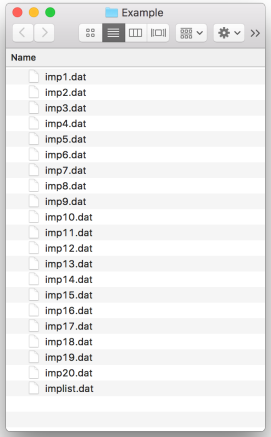
For females, the slope of depression on disability is is .402, indicating that an increase in depression increases psychosocial disability

The negative interaction coefficient means that the male slope is lower by .186 (p = .019)

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



Mplus Imputation Analysis Script

```
DATA:
file = imps_list.dat;
type = imputation;
VARIABLE:
names = id male age edugroup workhrs exercise pain anxiety stress control interfere depress disability;
usevariables = male depress disability product;
DEFINE:
center depress (grandmean);
product = male * depress;
MODEL:
disability on depress (b1)
 male (b2)
 product (b3);
MODEL TEST:
0 = b1; b1 = b2; b2 = b3;
```


Mplus Output

```
MODEL FIT INFORMATION

Number of Free Parameters                       5

...

Wald Test of Parameter Constraints

          Value                             69.064
          Degrees of Freedom                     3
          P-Value                           0.0000
```

Mplus Output

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value
 DISABILI ON
    DEPRESS          0.402      0.054      7.393      0.000
    MALE            -0.345      0.477     -0.723      0.470
    PRODUCT         -0.186      0.079     -2.353      0.019
 Intercepts
    DISABILITY      16.980      0.303     56.032      0.000
 Residual Variances
    DISABILITY      14.594      1.321     11.051      0.000
```


R Imputation Analysis Script

```
library(mitml)
library(plyr)

# read stacked data
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
impdata <- read.table(paste0(getwd(),"/imps_stacked.dat"))
names(impdata) <- c("imputation", "id", "male", "age", "edugroup", "workhrs",
  "exercise", "pain", "anxiety", "stress", "control", "interfere", "depress", "disability")
impdata <- impdata[impdata$imputation > 0, ]

# center focal predictor
impdata <- ddply(impdata, c("imputation"), transform,
  depressc = scale(depress, center = T, scale = F))
```


R Imputation Analysis Script

```
# analyze data and pool estimates
implist <- as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lm(disability ~ depressc + male + depressc*male))
estimates <- testEstimates(analysis, var.comp = T, df.com = NULL)
estimates

# estimate empty model with no predictors
emptymodel <- with(implist, lm(disability ~ 1))

# compare models with Wald test (e.g., MI version of omnibus F test)
testModels(analysis, emptymodel, method = "D1")

# compare models with likelihood ratio test (e.g., MI version of chi-square diff)
testModels(analysis, emptymodel, method = "D3")
```


R Output

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

               Estimate Std.Error   t.value         df   P(>|t|)       RIV       FMI
(Intercept)      16.980     0.309    55.020   5573.767     0.000     0.062     0.059
depressc          0.402     0.055     7.346    923.755     0.000     0.167     0.145
male             -0.346     0.482    -0.719  34262.851     0.472     0.024     0.024
depressc:male    -0.186     0.080    -2.338   1304.064     0.020     0.137     0.122

                   Estimate
Residual~~Residual   14.810

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

R Output

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

    F.value      df1      df2    P(>F)      RIV
     23.798        3 6241.015    0.000    0.098

Unadjusted hypothesis test as appropriate in larger samples.


Model comparison calculated from 20 imputed data sets.
Combination method: D3

    F.value      df1      df2    P(>F)      RIV
     20.441        3 3287.743    0.000    0.140
```

Wald test (D1)

Likelihood ratio test (D3)

SAS Imputation Analysis Script

```
/* read stacked imputation data */
data impdata (where = (_imputation_ gt 0));
infile '/folders/myfolders/imps_stacked.dat';
input _imputation_ id male age edugroup workhrs exercise pain anxiety stress control interfere depress disability;
run;

/* center covariate at grand means in each data set */
proc means data = impdata noprint;
var depress;
by _imputation_;
output out = grandmeans (drop =_type_ _freq_) mean = depressmean; run;

data impdata;
merge impdata grandmeans;
by _imputation_;
depressc = depress - depressmean;
product = depressc * male; run;
```


SAS Imputation Analysis Script

```
/* analyze imputations */
proc reg data = impdata outest = estimates covout noprint;
model disability = depressc male product;
by _imputation_;
run;

/* pool estimates and standard errors */
proc mianalyze data = estimates;
modeleffects Intercept depressc male product;
run;

/* omnibus test of model fit */
proc mianalyze data = estimates mult;
modeleffects depressc male product;
run;
```


SAS Output

The MIANALYZE Procedure

Model Information	
Data Set	WORK.ESTIMATES
Number of Imputations	20

Variance Information (20 Imputations)

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Intercept	0.003353	0.089611	0.093132	13293	0.039292	0.037951	0.998106
depressc	0.000409	0.002566	0.002995	923.75	0.167428	0.145265	0.992789
male	0.004288	0.226559	0.231061	50050	0.019871	0.019523	0.999025
product	0.000731	0.005592	0.006360	1304.1	0.137276	0.122051	0.993934

Parameter Estimates (20 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	16.979666	0.305175	16.38148	17.57785	13293	16.892744	17.143617	0	55.64	<.0001
depressc	0.402062	0.054729	0.29465	0.50947	923.75	0.367223	0.441484	0	7.35	<.0001
male	-0.345152	0.480688	-1.28731	0.59700	50050	-0.515220	-0.254030	0	-0.72	0.4727
product	-0.186419	0.079751	-0.34287	-0.02997	1304.1	-0.240442	-0.138388	0	-2.34	0.0196

SAS Output

The MIANALYZE Procedure

Model Information	
Data Set	WORK.ESTIMATES
Number of Imputations	20

Variance Information (20 Imputations)

Parameter	Variance: Between	Variance: Within	Variance: Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
depressc	0.000409	0.002566	0.002995	923.75	0.167428	0.145265	0.992789
male	0.004288	0.226559	0.231061	50050	0.019871	0.019523	0.999025
product	0.000731	0.005592	0.006360	1304.1	0.137276	0.122051	0.993934

Parameter Estimates (20 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
depressc	0.402062	0.054729	0.29465	0.50947	923.75	0.367223	0.441484	0	7.35	<.0001
male	-0.345152	0.480688	-1.28731	0.59700	50050	-0.515220	-0.254030	0	-0.72	0.4727
product	-0.186419	0.079751	-0.34287	-0.02997	1304.1	-0.240442	-0.138388	0	-2.34	0.0196

Multivariate Inference
Assuming Proportionality of Between/Within Covariance Matrices

Avg Relative Increase in Variance	Num DF	Den DF	F for H0: Parameter=Theta0	Pr > F
0.096540	3	6411.1	23.83	<.0001

SPSS Imputation Analysis Script

```
* set working directory.
CD "YOUR-FILE-PATH".

* read stacked imputation data.
DATA LIST free file = "imps_stacked.dat"
 /imputation_ id male lunchasst achievegrp stanread efficacy anxiety mathpre mathpost.
MISSING VALUES all (999).

* center focal predictor and compute product.
AGGREGATE
 /break = imputation_
 /depressmean = mean(depress).
COMPUTE depressc = depress - depressmean.
COMPUTE product = depressc * male.
```


SPSS Imputation Analysis Script

```
* initiate pooling routines.
SORT CASES by imputation_.
SPLIT FILE layered by imputation_.

* analysis and pooling.
REGRESSION
 /descriptives mean stddev corr sig n
 /dependent disability
 /method enter depressc male product.
```


SPSS Analysis Output

Coefficients[a]

imputation_	Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.
.00	1	(Constant)	17.051	.326		52.295	.000
		depressc	.423	.056	.615	7.537	.000
		male	-.562	.500	-.065	-1.122	.263
		product	-.235	.080	-.238	-2.926	.004
1.00	1	(Constant)	17.039	.293		58.092	.000
		depressc	.392	.047	.588	8.419	.000
		male	-.433	.467	-.049	-.929	.354
		product	-.148	.071	-.146	-2.092	.037
20.00	1	(Constant)	16.888	.300		56.268	.000
		depressc	.367	.050	.543	7.417	.000
		male	-.300	.478	-.034	-.627	.531
		product	-.138	.073	-.138	-1.886	.060
Pooled	1	(Constant)	16.980	.309		55.020	.000
		depressc	.402	.055		7.346	.000
		male	-.346	.482		-.719	.472
		product	-.186	.080		-2.338	.020

a. Dependent Variable: disability

Stata Imputation Analysis Script

```
// set working directory
cd "YOUR-FILE-PATH"

// read stacked data
clear
infile imp id male age edugroup workhrs exercise pain anxiety stress control interfere depress disability
  using "imps_stacked.csv"

// recode missing data in original data (imp = 0)
recode male - disability (999 = .)

// center focal predictor
egen impmeans = mean(depress), by(imp)
gen depressc = depress - impmeans

// convert to mi data
mi import flong, m(imp) id(id) imputed(male - depressc) clear

// analyze data and pool results
mi estimate, cmdok: regress disability depressc male c.depressc#c.male
```


Stata Output

```
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =        275
                                                Average RVI       =     0.0846
                                                Largest FMI       =     0.1463
                                                Complete DF       =        271
DF adjustment:   Small sample                   DF:     min       =     184.43
                                                        avg       =     221.91
                                                        max       =     260.69
Model F test:       Equal FMI                   F(   3,  257.2)   =      23.80
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
  disability |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    depressc |   .4020616   .0547289     7.35   0.000     .2940865    .5100368
        male |  -.3463137    .481765    -0.72   0.473    -1.29496     .6023326
             |
c.depressc#c.male |  -.1864192   .0797507    -2.34   0.020    -.3436782   -.0291603
             |
       _cons |   16.98037   .3086214    55.02   0.000     16.37244    17.58829
------------------------------------------------------------------------------
```