Imputation for Interaction Effects

Interaction (Moderation)

Moderation (interaction) occurs when the magnitude of a bivariate relation depends on a third variable

In a regression analysis, the influence of the focal predictor depends on the value of the moderator

e.g., The influence of pain severity (focal) on daily stress (outcome) is different for males and females (moderator)

Moderated Regression

Moderated regression analysis where the influence of X_1 depends on X_2 (or vice versa)

$$Y = \beta_0 + \beta_1 (X_1) + \beta_2 (X_2) + \beta_3 (X_1)(X_2) + e$$

Bias-inducing incompatibilities arise when lower-order variables (and the product) are incomplete

Just-Another-Variable Imputation

Just-another-variable imputation inappropriately treats the product just like any normal variable

$$Z = (X_1)(X_2)$$

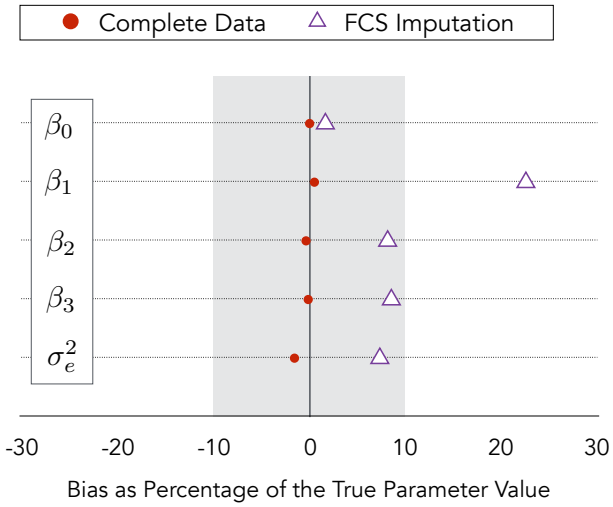
$$Z_{(mis)}^{(t)} = \gamma_0 + \gamma_1 (Y) + \gamma_2 (X_1) + \gamma_3 \left(X_2^{(t)}\right) + \varepsilon$$

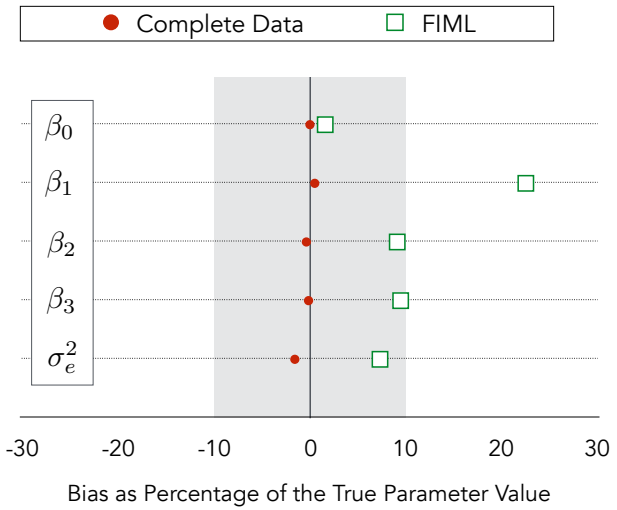
$$Z_{(mis)}^{(t)} \sim N\left(\gamma_0 + \gamma_1 (Y) + \gamma_2 (X_1) + \gamma_3 \left(X_2^{(t)}\right), \sigma_\varepsilon^2\right)$$

Computer Simulation

2000 replications of $N = 250$ with 25% of X_2 scores missing due to X_1 (missing at random)

$$Y = \beta_0 + \beta_1 (X_1) + \beta_2 (X_2) + \beta_3 (X_1)(X_2) + e$$
$$= 5 + 1 (X_1) + 1 (X_2) + 1 (X_1)(X_2)$$
$$\sigma_e^2 = 4$$





Substantive Model Compatible Imputation

Predictor variables are imputed only from other predictors (no product term), and the outcome is imputed from a model that matches the substantive analysis

A special algorithm generates imputed predictor variables from a complex distribution that ensures the imputations "fit" well in a moderated regression

The product is not directly imputed, thus reducing bias

Full Bayesian Approach (Substantive Model Compatible Imputation)

Joint posterior distribution

$$p(X_{2(mis)}, \gamma, \sigma_\varepsilon^2 | Y, X_1, \beta, \sigma_e^2)$$

$$\propto p\left(Y | X_1, X_{2(mis)}, \boldsymbol{\beta}, \sigma_e^2\right) p\left(X_{2(mis)} | X_1, \gamma, \sigma_\varepsilon^2\right) p\left(\gamma, \sigma_\varepsilon^2\right)$$

Substantive model | Model for covariate | Prior

Distribution Of Missing Predictor Scores

$$p(X_{2(mis)}|\cdot)$$

$$\propto p\left(Y | X_1, X_{2(mis)}, \boldsymbol{\beta}, \sigma_e^2\right) p\left(X_{2(mis)} | X_1, \gamma, \sigma_\varepsilon^2\right)$$

$$\propto exp\left(-\frac{\left(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_{2(mis)} - \beta_3 X_1 X_{2(mis)}\right)^2}{2\sigma_e^2}\right)$$

$$\times exp\left(-\frac{\left(X_{2(mis)} - \gamma_0 - \gamma_1 X_1\right)^2}{2\sigma_\varepsilon^2}\right)$$

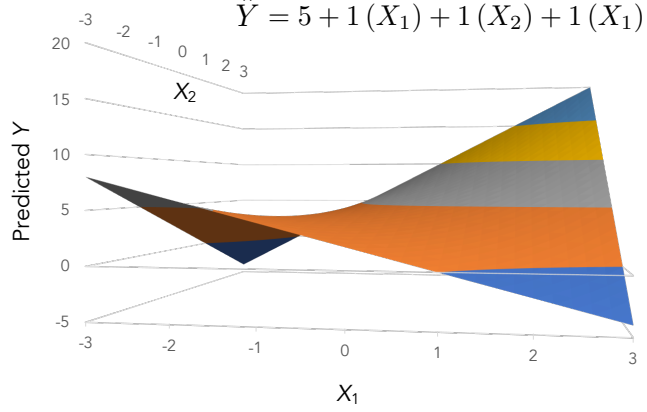
Distribution of Missing Values

Missing predictor scores must be sampled from a composite distribution that equals the product of two normal distributions

We can view the distribution or missing values as a normal distribution for the predictors, the points on which are weighted by how well each pair of scores fits a moderated regression model

Moderated Regression Predicted Scores

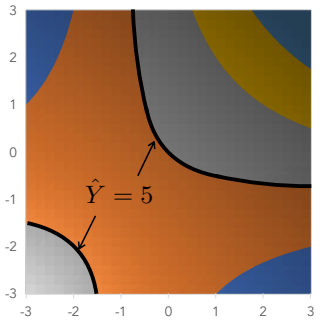
$$\hat{Y} = 5 + 1\left(X_1\right) + 1\left(X_2\right) + 1\left(X_1\right)\left(X_2\right)$$



Example: Imputing X_2 When $Y = 5$

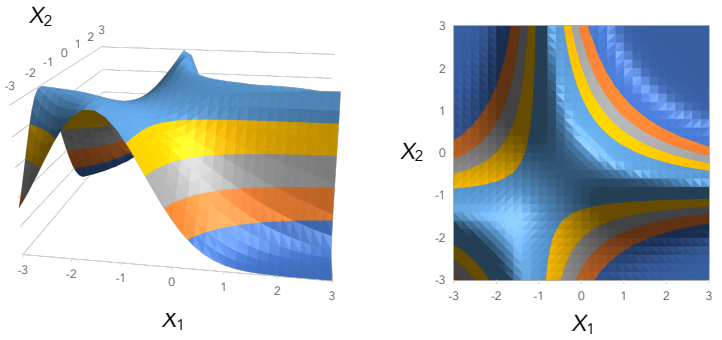
Consider a case with $Y = 5$

The fit (likelihood) for that case is maximized when its configuration of X scores give predicted values equal to 5



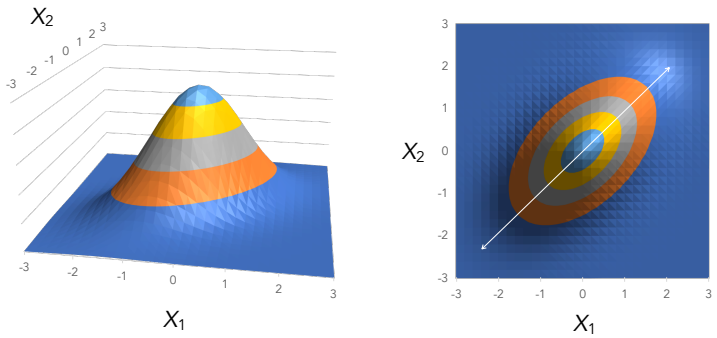
Likelihood Of $Y = 5$ Given X_1 And X_2

$$p\left(Y|X_1, X_{2(mis)}, \boldsymbol{\beta}, \sigma_e^2\right)$$



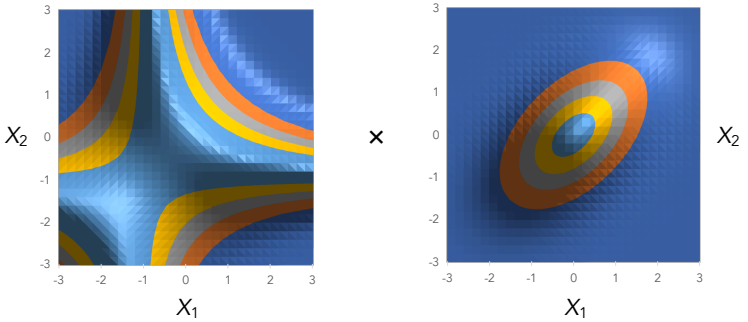
Conditional Distribution Of X_2

$$p\left(X_{2(mis)}|X_1, \gamma, \sigma_\varepsilon^2\right)$$



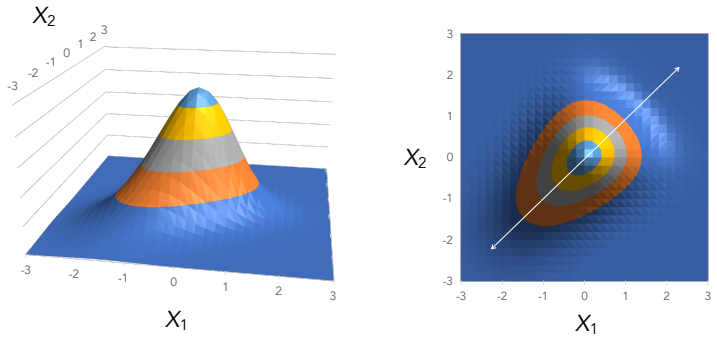
Component Distributions

$$p\left(Y|X_1, X_{2(mis)}, \boldsymbol{\beta}, \sigma_e^2\right) \times p\left(X_{2(mis)}|X_1, \gamma, \sigma_\varepsilon^2\right)$$



Distribution Of Missing Values

$$p\left(Y|X_1, X_{2(mis)}, \boldsymbol{\beta}, \sigma_e^2\right) p\left(X_{2(mis)}|X_1, \gamma, \sigma_\varepsilon^2\right)$$



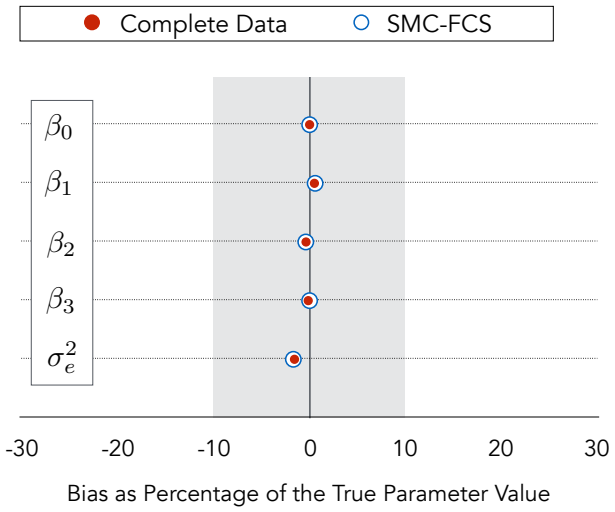
Computer Simulation Revisited

2000 replications of $N = 250$ with 25% of X_2 scores missing due to X_1 (missing at random)

$$\begin{aligned} Y &= \beta_0 + \beta_1\left(X_1\right) + \beta_2\left(X_2\right) + \beta_3\left(X_1\right)\left(X_2\right) + e \\ &= 5 + 1\left(X_1\right) + 1\left(X_2\right) + 1\left(X_1\right)\left(X_2\right) \\ \sigma_e^2 &= 4 \end{aligned}$$

Blimp Syntax

```
DATA: ~/desktop/example.dat;
VARNAMES: y x z;
MODEL: ~ y x z x*z;
OUTCOME: y;
BURN: 500;
THIN: 500;
NIMPS: 20;
MISSING: -99;
SEED: 90291;
OUTFILE: ~/desktop/imps.csv;
CHAINS: 2 processors 2;
```



Analysis Example

Analysis Model

Data from a sample of 250 chronic pain patients

The analysis is a regression that examines whether the influence of pain on stress differs for males and females

$$Stress = \beta_0 + \beta_1(Pain) + \beta_2(Female) + \beta_3(Pain)(Female) + e$$

Pain scores are incomplete and must be imputed

Ex7.1.imp
Blimp Diagnostic Script

```
DATA: ~/desktop/examples/pain.csv;
VARNAMES: id female diagnose sleep pain posaff
  negaff stress;
MISSING: -99;
MODEL: ~ stress pain female pain*female posaff;
ORDINAL: female;
OUTCOME: stress;
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  |       1.069|         nan|       1.073|         nan|
   Missing Variable |      stress|            |      stress|            |
                    ---------------------------------------------------

  Comparing iterations 101 to 200 for 2 chains.
                    ---------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    ---------------------------------------------------
           Max PSR  |       1.023|         nan|       1.011|         nan|
   Missing Variable |        pain|            |      posaff|            |
                    ---------------------------------------------------
```

Ex7.2.imp
Blimp Imputation Script (Mplus Format)

```
DATA: ~/desktop/examples/pain.csv;
VARNAMES: id female diagnose sleep pain posaff
  negaff stress;
MISSING: -99;
MODEL: ~ stress pain female pain*female posaff;
ORDINAL: female;
OUTCOME: stress;
SEED: 90291;
BURN: 200;
THIN: 200;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Blimp Output

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

VARIABLE ORDER IN SAVED DATA:

   id female diagnose sleep pain posaff negaff stress

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

Ex7.3.inp
Mplus Analysis Script

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = id female diagnose sleep pain posaff
  negaff stress;
usevariables = stress female pain femxpain;
DEFINE:
femxpain = female*pain;
MODEL:
stress on female pain femxpain;
OUTPUT:
standardized;
```

Mplus Analysis Output

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
STRESS ON					
FEMALE	-1.695	0.650	-2.610	0.009	0.286
PAIN	0.090	0.088	1.024	0.306	0.316
FEMXPAIN	0.354	0.136	2.590	0.010	0.283
Intercepts					
STRESS	3.343	0.374	8.950	0.000	0.347
Residual Variances					
STRESS	0.744	0.077	9.639	0.000	0.260

Ex7.4.imp
Blimp Imputation Script (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/pain.csv;
VARNAMES: id female diagnose sleep pain posaff
  negaff stress;
MISSING: -99;
MODEL: ~ stress pain female pain*female posaff;
ORDINAL: female;
OUTCOME: stress;
SEED: 90291;
BURN: 200;
THIN: 200;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Blimp Output

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

VARIABLE ORDER IN SAVED DATA:

  imp# id female diagnose sleep pain posaff negaff stress

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

Ex7.5.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","female","diagnose","sleep","pain",
 "posaff","negaff","stress")

# Compute product variable
impdata$femxpain <- impdata$female * impdata$pain
```

Ex7.5.r
R Analysis Script

```
# Analyze data and pool estimates
implist <- as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lm(stress ~ female + pain + femxpain))
estimates <- testEstimates(analysis, var.comp = T, df.com = 246)
estimates

# Compare models with Wald test
emptymodel <- with(implist, lm(stress ~ 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)     3.343     0.376     8.902    82.791     0.000     0.504     0.343
female         -1.695     0.653    -2.594   103.078     0.011     0.383     0.283
pain            0.090     0.088     1.018    92.562     0.311     0.439     0.312
femxpain        0.354     0.137     2.575   104.306     0.011     0.377     0.279

                  Estimate
Residual~~Residual    0.756

Hypothesis test adjusted for small samples with df=[246]
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
     7.042       3 960.250   0.000   0.297

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

Ex7.6.sps
SPSS Analysis Script

```
data list free file =  '/users/craig/desktop/examples/imps.csv'
 /imputation_ id female diagnose sleep pain posaff negaff stress.
exe.

* Compute product variable.
compute femxpain = female * pain.
exe.

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

* Analysis and pooling.
regression
  /descriptives mean stddev corr sig n
  /dependent stress
  /method enter female pain femxpain.
```

SPSS Analysis Output

Coefficients[a]

imputation_	Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1.00	1	(Constant)	3.101	.324		9.576	.000
		female	-1.412	.574	-.760	-2.459	.015
		pain	.146	.077	.167	1.893	.060
		femxpain	.288	.121	.801	2.382	.018
...							
20.00	1	(Constant)	3.118	.322		9.679	.000
		female	-1.097	.576	-.586	-1.904	.058
		pain	.149	.078	.171	1.919	.056
		femxpain	.220	.122	.604	1.800	.073
Pooled	1	(Constant)	3.343	.376		8.902	.000
		female	-1.695	.653		-2.594	.010
		pain	.090	.088		1.018	.310
		femxpain	.354	.137		2.575	.011

Ex7.7.do
Stata Analysis Script

```
// Import and save original data
import delimited "~/desktop/examples/pain.csv"
rename (v1 - v8)(id female diagnose sleep pain posaff negaff 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 - v9)(imp id female diagnose sleep pain posaff
  negaff stress)
save imputed, replace
```

Ex7.7.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(female - stress) clear

// Compute product term
gen femxpain = female * pain

// Analyze data and pool results
mi estimate, cmdok: regress stress female pain femxpain
```

Stata Analysis Output

```
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =        250
                                                Average RVI       =     0.2728
                                                Largest FMI       =     0.3455
                                                Complete DF       =        246
DF adjustment:   Small sample                   DF:     min       =      82.79
                                                        avg       =      95.68
                                                        max       =     104.31
Model F test:       Equal FMI                   F(   3,  192.1)   =       7.04
Within VCE type:          OLS                   Prob > F          =     0.0002

------------------------------------------------------------------------------
      stress |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.695095   .6533503    -2.59   0.011    -2.990849   -.3993403
        pain |   .0896054   .0880383     1.02   0.311    -.0852321    .2644429
    femxpain |   .3535154   .1372886     2.57   0.011     .0812764    .6257544
       _cons |   3.343092   .3755457     8.90   0.000     2.596119    4.090065
------------------------------------------------------------------------------
```