Multiple Imputation in Practice

Aprroaches for handing categorical and interaction variables

Aya A Mitani
amitani@stanford.edu

Quantitative Sciences Unit, Stanford University School of Medicine

April 17, 2013

Outline

- Background of multiple imputation (MI)
- Challenges to the user
- How SAS, Stata, and R handle these challenges
- Real world example using software
- General guidelines and conclusion

Some background: Patterns of missingness

There are 3 main categories for describing missing data pattern

1. Missing completely at Random (MCAR)

   Missingness is unrelated to any factor

2. Missing at Random (MAR)

   Missingness depends only on observed values

3. Not Missing at Random (NMAR)

   Missingness is related to unobserved values

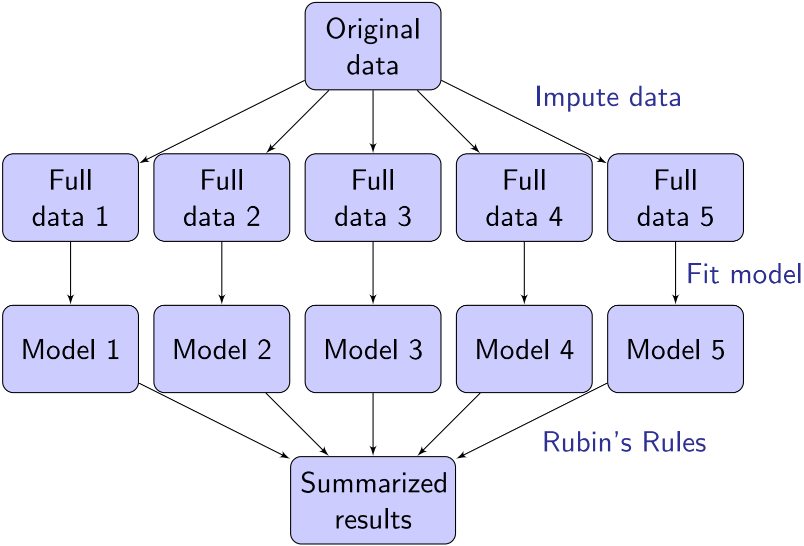
Some background: Multiple Imputation (MI)

MI is a simulation-based method for filling in missing values using observed data (valid if MAR)

Typical MI approach involves 3 basic steps

1. Imputation
2. Model fitting
3. Summarize estimates using Rubin's Rules (Rubin, 1987)

Some background: Multiple Imputation (MI)



Algorithm: JM vs FCS

Joint Modelling (JM)

- Specify joint model, usually under multivariate normal (MVN)
- Derive posterior predictive distribution i.e. distribution of unobserved values conditional on observed data

Fully Conditional Specification (FCS)

- Designed to handle variables of mixed type
- Specify conditional model for each missing variable
- Impute data on a variable-by-variable basis

van Buuren (2007) compares the two methods in greater detail

Challenges - User has to make decisions

- Variables to include in the imputation model
- Number of imputations
- Model selection
- Longitudinal data
- Variables of mixed type
    - Especially nominal categorical variables such as race
- Derived variables
    - Higher order terms (X^2)
    - Interaction effects ($X_1 \times X_2$)
    - Propensity scores

Challenges - User has to make decisions

- Variables to include in the imputation model
- Number of imputations
- Model selection
- Longitudinal data
- Variables of mixed type
  - Especially nominal categorical variables such as race
- Derived variables
  - Higher order terms (X^2)
  - Interaction effects ($X_1 \times X_2$)
  - Propensity scores

Why is nominal categorical variable a challenge?

Even though nominal categorical variables (e.g. race) are usually coded as numerical, the values are purely representative.

They are converted into dummy variables in the regression model.

In the imputation model, do we enter them as

- One class variable?
- Series of dummy variables?

How does each method handle categorical variables?

Joint Modelling (JM)

- If the imputation model is $Y, X_{Age}, X_{Male}, X_{Race}$
- The imputed value will not necessarily be a whole number

Original Data

ID	Age	Male	Race
1	60	1	2
2	50	.	3
3	.	0	.
4	40	0	.
5	55	1	1

$\rightarrow$

Imputed Data 1

ID	Age	Male	Race
1	60	1	2
2	50	1.2	3
3	44.4	0	-1.1
4	40	0	4.3
5	55	1	1

Options

1. Use these values (not an option for Race)
2. Round the values (rounding at 0.5 is not recommended for binary variables, Horton 2003)
3. Use dummy variables in the imputation model as you would in your regression model

Making use of dummy variables in the imputation model

Joint Modelling (JM)

What does your regression model look like?

$$Y = \beta_0 + \beta_1 Age + \beta_2 Male + \beta_3 Black + \beta_4 Asian$$

Original Data

ID	Age	Male	Black	Asian
1	60	1	1	0
2	50	.	0	1
3	.	0	.	.
4	40	0	.	.
5	55	1	0	0

→

Imputed Data 1

ID	Age	Male	Black	Asian
1	60	1	1	0
2	50	1.2	0	1
3	44.4	0	0.1	0.2
4	40	0	0.3	1.5
5	55	1	0	0

Options

1. Use these values
2. Round the values (rounding at 0.5 is not recommended for binary variables, Horton 2003)

How does each method handle categorical variables?

Joint Modelling (JM)

Rounding options for binary variables

- Bernaards (2006) – Calculate cut-off value based on a normal approximation to the binomial distribution
- Yucel (2008) – Observed proportions
- Demirtas (2009) – Round at 0.5, run logistic regression model to use as a refinement

Rounding options for categorical variables

- Allison (2000) – Missing Data
- Song (2009) – Correction of Bias in Imputing Missing Values of Categorical Variables
- Yucel (2011) – Gaussian-Based Routines to Impute Categorical Variables in Health Surveys

These methods are not implemented in most software.

How does each method handle categorical variables?

Fully Conditional Specification (FCS)

- Specify a set of imputation models for each variable
    - Use linear regression to impute age
    - Use logistic regression to impute sex
    - Use multinomial regression to impute race

Original Data

ID	Age	Male	Race
1	60	1	2
2	50	.	3
3	.	0	.
4	40	0	.
5	55	1	1

$\rightarrow$

Imputed Data 1

ID	Age	Male	Race
1	60	1	2
2	50	1	3
3	44.4	0	1
4	40	0	3
5	55	1	1

Imputed data are ready to be analyzed.

How to handle interaction variables?

Options for JM and FCS

- Impute then transform
- Transform then impute
    - Active imputation

Additional options for FCS

- Passive imputation

von Hippel (2009) suggests to *transform then impute*, rather than *impute then transform*

Options for Interactions: Active vs Passive imputation

Active Imputation

- Assumes the interaction variable to be another independent variable
- Include the interaction variable in the imputation model with all other variables including the main effects

Passive Imputation

- Passively imputes the interaction variable
- interaction variables are used to impute other missing values but not the main effects

MI by FCS	*Vars in Imp Model*	
Variable	*Active*	*Passive*
X_1	Y, X_2, I_{12}, Z	Y, X_2, Z
X_2	Y, X_1, I_{12}, Z	Y, X_1, Z
$I_{12} : X_1 \times X_2$	Y, X_1, X_2, Z	–
Z	Y, X_1, X_2, I_{12}	Y, X_1, X_2, I_{12}

How does each method handle interaction variables?

Active Imputation

- The relationship between the interaction effect and the main effects are not necessarily internally consistent

Original Data					Imputed Data 1			
ID	Age	Male	Age $\times$ Male		ID	Age	Male	Age $\times$ Male
1	60	1	60		1	60	1	60
2	50	.	.		2	50	1	40
3	.	0	.	$\rightarrow$	3	44.4	0	2
4	40	0	0		4	40	0	0
5	55	1	55		5	55	1	55

How does each method handle interaction variables?

Passive Imputation

- The relationship between the interaction effect and the main effects is preserved

Original Data

ID	Age	Male	Age × Male
1	60	1	60
2	50	.	.
3	.	0	.
4	40	0	0
5	55	1	55

→

Imputed Data 1

ID	Age	Male	Age × Male
1	60	1	60
2	50	1	50
3	44.4	0	0
4	40	0	0
5	55	1	55

Another Challenge

- Interaction between two nominal categorical variables

Interaction variables

Simple case

Binary $\times$ Binary

Slightly more complicated case

Binary $\times$ Multi-level Categorical

Most complicated case

Multi-level Categorical $\times$ Multi-level Categorical

When interaction is Binary $\times$ Binary

Race: $X_1 = \begin{cases} 1, & \text{if White} \\ 0, & \text{Other} \end{cases}$ Drug: $X_2 = \begin{cases} 1, & \text{if drug A} \\ 0, & \text{if drug B} \end{cases}$

Interaction: $I_{12} = X_1 \times X_2 = \begin{cases} 1, & \text{if White \& drug A} \\ 0, & \text{All other} \end{cases}$

Regression model: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} I_{12} + e$

Form of Interaction Variables in Imputation Model

	JM	*FCS*
Active	I_{12}	specify 'logistic' for I_{12}
Passive	–	derive from X_1 and X_2

When interaction is Categorical $\times$ Binary

If no interaction...

$$\text{Race: } X_1 = \begin{cases} 1, & \text{if White} \\ 2, & \text{if Black} \\ 0, & \text{Other} \end{cases} \qquad \text{Drug: } X_2 = \begin{cases} 1, & \text{if drug A} \\ 0, & \text{if drug B} \end{cases}$$

We create dummy variables for X_1

$$X_\mathrm{W} = \begin{cases} 1, & \text{if White} \\ 0, & \text{Otherwise} \end{cases} \quad X_\mathrm{B} = \begin{cases} 1, & \text{if Black} \\ 0, & \text{Otherwise} \end{cases}$$

Regression model without interaction:
$Y = \alpha + \beta_1 X_W + \beta_2 X_B + \beta_3 X_2 + e$

When interaction is Categorical $\times$ Binary

By introducing an interaction term between X_1 and X_2...

Model:

$$Y = \alpha + \beta_1 X_W + \beta_2 X_B + \beta_3 X_2 + \beta_4 X_W X_2 + \beta_5 X_B X_2 + e$$

Interaction is also the product of two main effects

$$\text{Interaction: } I_{12} = X_1 \times X_2 = \begin{cases} 1, & \text{if White \& drug A} \\ 2, & \text{if Black \& drug A} \\ 0, & \text{All other} \end{cases}$$

Form of Interaction Variables in Imputation Model

	JM	*FCS*
Active	$X_W X_2, X_B X_2$	specify 'multinomial' for I_{12}
Passive	–	derive from X_1 and X_2

When interaction is Categorical × Categorical

$$\text{Race: } X_1 = \begin{cases} 1, & \text{if White} \\ 2, & \text{if Black} \\ 0, & \text{Other} \end{cases} \qquad \text{Drug: } X_2 = \begin{cases} 1, & \text{if drug A} \\ 2, & \text{if drug B} \\ 0, & \text{if drug C} \end{cases}$$

Model without interaction:

$$Y = \alpha + \beta_1 X_W + \beta_2 X_B + \beta_3 X_{DA} + \beta_4 X_{DB} + e$$

When interaction is Categorical × Categorical

$$\text{Race: } X_1 = \begin{cases} 1, & \text{if White} \\ 2, & \text{if Black} \\ 0, & \text{Other} \end{cases} \qquad \text{Drug: } X_2 = \begin{cases} 1, & \text{if drug A} \\ 2, & \text{if drug B} \\ 0, & \text{if drug C} \end{cases}$$

Model without interaction:

$$Y = \alpha + \beta_1 X_W + \beta_2 X_B + \beta_3 X_{DA} + \beta_4 X_{DB} + e$$

Model with interaction:

$$\begin{aligned} Y = \alpha + \beta_1 X_W + \beta_2 X_B + \beta_3 X_{DA} + \beta_4 X_{DB} \\ + \beta_5 X_W X_{DA} + \beta_6 X_W X_{DB} \\ + \beta_7 X_B X_{DA} + \beta_8 X_B X_{DB} + e \end{aligned}$$

When interaction is Categorical $\times$ Categorical

Interaction is NOT the product of two main effects

$$\text{Interaction: } I^*_{12} = \begin{cases} 1, & \text{if White \& drug A} \\ 2, & \text{if White \& drug B} \\ 3, & \text{if Black \& drug A} \\ 4, & \text{if Black \& drug B} \\ 0, & \text{All other} \end{cases} \neq X_1 \times X_2$$

When interaction is Categorical $\times$ Categorical

Interaction is NOT the product of two main effects

$$\text{Interaction: } I^*_{12} = \begin{cases} 1, & \text{if White \& drug A} \\ 2, & \text{if White \& drug B} \\ 3, & \text{if Black \& drug A} \\ 4, & \text{if Black \& drug B} \\ 0, & \text{All other} \end{cases} \neq X_1 \times X_2 = \begin{cases} 1 \\ 2 \\ 4 \\ 0 \end{cases}$$

When interaction is Categorical $\times$ Categorical

Interaction is NOT the product of two main effects

$$\text{Interaction: } I_{12}^{*} = \begin{cases} 1, & \text{if White \& drug A} \\ 2, & \text{if White \& drug B} \\ 3, & \text{if Black \& drug A} \\ 4, & \text{if Black \& drug B} \\ 0, & \text{All other} \end{cases} \neq X_1 \times X_2 = \begin{cases} 1 \\ 2 \\ 4 \\ 0 \end{cases}$$

Form of Interaction Variables in Imputation Model

	JM	*FCS*
Active	$X_W X_{DA}, X_W X_{DB}, X_B X_{DA}, X_B X_{DB}$	specify 'multinomial' for I_{12}
Passive	–	derive from X_1 and X_2

Interaction variables - Summary

JM

1. Convert interaction variable into dummy variables
2. Multiply impute the data containing dummy variables
3. Use imputed dummy variables in the scientific model

FCS

1. Compute interaction variable
2. Multiply impute the interaction variable as one class variable
3. Convert interaction variable into dummy variables to include in the scientific model

Software choices

SAS - PROC MI and PROC MIANALYZE

- JM
- Recently introduced FCS in version 9.3
- No option for passive imputation

Stata - ice and micombine

- FCS
- Option for passive imputation

R - mice and pool

- FCS
- Option for passive imputation

There are many more choices but we will focus on these commands

SAS - PROC MI JM

- Pros
  - User-friendly
  - Computationally efficient
- Cons
  - No passive option

SAS - PROC MI FCS

- FCS Modelling Options
  - discrim (discriminant function method)
  - logistic (logistic regression method)
  - regress (linear regression method)
- Pros
  - User-friendly
- Cons
  - FCS option is still in experimental stage
  - Logistic option only for ordinal logistic regression
  - Discrim option only utilizes continuous variables as predictors
  - Computationally inefficient
  - No passive option

Stata - ice

- FCS Modelling Options
  - regress (regression method)
  - logit (logistic regression method)
  - ologit (ordinal logistic regression method)
  - mlogit (multinomial logistic regression method)
- Pros
  - User-friendly
  - Passive option available
- Cons
  - For multi-level (usually 6 or more), ice often gives an error for mlogit option (can use the persist option to ignore the error)
  - Passive option in ice needs care for specifying interaction between two multi-level nominal categorical variables

Passive option in ice

Command:

```
*manually generate interaction variable
gen int = 0
replace int = 1 if x1 == 1 & x2 == 1
replace int = 2 if x1 == 1 & x2 == 2
replace int = 3 if x1 == 2 & x2 == 1
replace int = 4 if x1 == 2 & x2 == 2

ice x1 x2 int z y, passive(int:x1*x2)
 cmd(x1 x2:mlogit) saving(imp.dta) m(10)
```

Even if we explicitly create the interaction beforehand, because the passive option computes `int` by multiplying `x1` and `x2`, `int` will only have 4 levels (0, 1, 2, 4) after imputation

Passive option in ice

Reassign values for X_2

$$X_1 = \begin{cases} 0 \\ 1 \\ 2 \end{cases} \qquad X_2 = \begin{cases} 0 \\ 3 \\ 5 \end{cases} \qquad X_1 \times X_2 = \begin{cases} 0 \\ 3 \\ 5 \\ 6 \\ 10 \end{cases}$$

```
*passive imputation
ice x1 x2 int z y, passive(int:x1*x2)
cmd(x1 x2:mlogit) saving(imp.dta) m(10)
*analytic model
micombine logit y b0.x1 b0.x2 b0.int
```

R - mice

- FCS Modelling Options
  - norm.nob (Linear regression)
  - norm (Bayesian linear regression)
  - logreg (Logistic regression)
  - polyreg (Polytomous/unordered regression)
  - lda (Linear discriminant analysis)
- Pros
  - Flexible
  - Passive option available
  - Automatically creates dummy variables for factor variables
- Cons
  - Categorical variables need to be factor variables [as.factor()]
  - Passive option requires intricate coding

Example Data

Breast Cancer Care from Two Different Hospitals

ID	Hosp	Mastectomy	AgeDx	Stage	YearDx	Race
1	1	1	64	1	2008	1
2	0	0	47	NA	1999	NA
3	0	0	80	NA	2009	1
4	1	1	55	3	2003	1
5	1	0	60	NA	2009	1
6	1	0	58	1	2009	1

- Total $N = 7747$
- Missing variables – Stage (26%), Race(13%)
- Overall missing proportion – 31%
- Logistic regression model
  - Outcome – Mastectomy
  - Predictors – Hospital, Age, Stage, Year, Race, Hospital $\times$ Stage

JM with *Impute then Transform* Approach in SAS

```
/* Proc MI JM Impute then Transform */

data dummy; set rawdat;
  if Stage ne . then do; *** create dummy variables for stage (ref: Stage 0);
   Stage1=(Stage=1);
   Stage2=(Stage=2);
   Stage3=(Stage=3);
   Stage4=(Stage=4);
  end;
  drop Stage;

proc mi data=dummy out=impdat_jm nimpute=5; *** multiply impute data;
 var Mastectomy Hosp AgeDx Stage1 Stage2 Stage3 Stage4 YearDx Race;
run;

proc genmod data=impdat_jm desc; *** build logistic regression model;
 by _imputation_;
 model Mastectomy = Hosp AgeDx Stage1 Stage2 Stage3 Stage4 YearDx Race
                    Hosp*Stage1 Hosp*Stage2 Hosp*Stage3 Hosp*Stage4/link=logit dist=binomial covb;
 ods output parameterestimates=gmparms covb=gmcovb parminfo=gmpinfo;
run;

proc mianalyze parms=gmparms covb=gmcovb parminfo=gmpinfo; *** summarize estimates from each model;
    modeleffects Hosp AgeDx Stage1 Stage2 Stage3 Stage4 YearDx Race
                 Hosp*Stage1 Hosp*Stage2 Hosp*Stage3 Hosp*Stage4;
run;
```

JM with *Impute then Transform* Approach in SAS

Obs	_Imputation_	id	Hosp	Mastectomy	Age Dx	Year Dx	race	Stage1	Stage2	Stage3	Stage4
1	1	1	1	1	64	2008	1	1.00000	0.00000	0.00000	0.00000
2	1	2	0	0	47	1999	1.29	0.32028	0.02254	0.11697	-0.14360
3	1	3	0	0	80	2009	1	0.15842	-0.05614	0.36258	0.06359
4	1	4	1	1	55	2003	1	0.00000	0.00000	1.00000	0.00000
5	1	5	1	0	60	2009	1	0.74354	-0.26356	-0.38134	0.14464
6	1	6	1	0	58	2009	1	1.00000	0.00000	0.00000	0.00000

Obs	_Imputation_	id	Hosp	Mastectomy	Age Dx	Year Dx	race	Stage1	Stage2	Stage3	Stage4
7748	2	1	1	1	64	2008	1	1.00000	0.00000	0.00000	0.00000
7749	2	2	0	0	47	1999	0.46	0.65210	-0.45562	0.09514	0.05591
7750	2	3	0	0	80	2009	1	-0.26870	0.70456	-0.49028	0.66006
7751	2	4	1	1	55	2003	1	0.00000	0.00000	1.00000	0.00000
7752	2	5	1	0	60	2009	1	0.16455	0.57457	-0.15339	-0.02116
7753	2	6	1	0	58	2009	1	1.00000	0.00000	0.00000	0.00000

Obs	_Imputation_	id	Hosp	Mastectomy	Age Dx	Year Dx	race	Stage1	Stage2	Stage3	Stage4
15495	3	1	1	1	64	2008	1	1.00000	0.00000	0.00000	0.00000
15496	3	2	0	0	47	1999	-.14	-0.21654	1.23289	0.06875	-0.03467
15497	3	3	0	0	80	2009	1	0.29618	0.51265	0.67214	0.14224
15498	3	4	1	1	55	2003	1	0.00000	0.00000	1.00000	0.00000
15499	3	5	1	0	60	2009	1	-0.14638	0.21437	0.12075	0.27207
15500	3	6	1	0	58	2009	1	1.00000	0.00000	0.00000	0.00000

Obs	_Imputation_	id	Hosp	Mastectomy	Age Dx	Year Dx	race	Stage1	Stage2	Stage3	Stage4
23242	4	1	1	1	64	2008	1	1.00000	0.00000	0.00000	0.00000
23243	4	2	0	0	47	1999	1.01	1.21119	-0.00646	0.07885	-0.17840
23244	4	3	0	0	80	2009	1	0.36106	0.17560	0.20136	0.14280
23245	4	4	1	1	55	2003	1	0.00000	0.00000	1.00000	0.00000
23246	4	5	1	0	60	2009	1	0.21547	0.26938	0.13176	0.42039
23247	4	6	1	0	58	2009	1	1.00000	0.00000	0.00000	0.00000

Obs	_Imputation_	id	Hosp	Mastectomy	Age Dx	Year Dx	race	Stage1	Stage2	Stage3	Stage4
30989	5	1	1	1	64	2008	1	1.00000	0.00000	0.00000	0.00000
30990	5	2	0	0	47	1999	1.77	0.32753	0.05600	0.16621	0.22796
30991	5	3	0	0	80	2009	1	1.14683	-0.01119	-0.18460	0.09322
30992	5	4	1	1	55	2003	1	0.00000	0.00000	1.00000	0.00000
30993	5	5	1	0	60	2009	1	0.91269	-0.56521	0.45316	0.24779
30994	5	6	1	0	58	2009	1	1.00000	0.00000	0.00000	0.00000

JM Using *Active Imputation* in SAS

```
/* Proc MI JM Active */

data active; set rawdat;
  if Stage ne . then do; *** create dummy variables for stage (ref: Stage 0);
   Stage1=(Stage=1);
   Stage2=(Stage=2);
   Stage3=(Stage=3);
   Stage4=(Stage=4);
  end;
  drop Stage;

  HospStage1 = Hosp*Stage1;
  HospStage2 = Hosp*Stage2;
  HospStage3 = Hosp*Stage3;
  HospStage4 = Hosp*Stage4;

proc mi data=active out=impdat_jmact nimpute=5; *** multiply impute data;
 var Mastectomy Hosp AgeDx Stage1 Stage2 Stage3 Stage4 YearDx Race
     HospStage1 HospStage2 HospStage3 HospStage4;
run;

proc genmod data=impdat_jmact desc; *** build logistic regression model;
 by _imputation_;
 model Mastectomy = Hosp AgeDx Stage1 Stage2 Stage3 Stage4 YearDx Race
                    HospStage1 HospStage2 HospStage3 HospStage4/link=logit dist=binomial covb;
 ods output parameterestimates=gmparms covb=gmcovb parminfo=gmpinfo;
run;

proc mianalyze parms=gmparms covb=gmcovb parminfo=gmpinfo; *** summarize estimates from each model;
    modeleffects Hosp AgeDx Stage1 Stage2 Stage3 Stage4 YearDx Race
                 HospStage1 HospStage2 HospStage3 HospStage4;
run;
```

JM Using *Active Imputation* in SAS

Obs	_Imputation_	id	Hosp	Stage1	Stage2	Stage3	Stage4	Hosp Stage1	Hosp Stage2	Hosp Stage3	Hosp Stage4
1	1	1	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
2	1	2	0	0.43152	0.49869	-0.20955	-0.11642	0.52256	0.37097	-0.61139	0.00940
3	1	3	0	0.65656	0.37696	-0.25971	0.01811	0.31598	0.05759	-0.39495	-0.05765
4	1	4	1	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000
5	1	5	1	-0.47303	0.66659	-0.28363	0.41804	-0.29570	0.42075	0.14982	0.34148
6	1	6	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
7748	2	1	1	1.00000	0.00000	0.00000	0.000000	1.00000	0.00000	0.00000	0.00000
7749	2	2	0	0.52470	0.14513	-0.22259	0.066284	0.07992	0.11460	-0.10222	-0.10886
7750	2	3	0	0.54456	0.16606	0.12060	-0.083323	0.49956	-0.13215	-0.29400	0.01143
7751	2	4	1	0.00000	0.00000	1.00000	0.000000	0.00000	0.00000	1.00000	0.00000
7752	2	5	1	0.54132	0.46470	0.20298	-0.001120	0.40440	0.20949	0.27891	0.12870
7753	2	6	1	1.00000	0.00000	0.00000	0.000000	1.00000	0.00000	0.00000	0.00000
15495	3	1	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
15496	3	2	0	0.40537	0.99224	-0.06326	-0.08847	-0.12439	0.42246	0.04509	-0.03960
15497	3	3	0	0.19079	0.05812	0.21852	-0.03938	-0.25338	-0.02575	0.31272	-0.02103
15498	3	4	1	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000
15499	3	5	1	1.13960	0.23239	0.32978	-0.13984	0.47708	0.39006	0.03960	-0.19573
15500	3	6	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
23242	4	1	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
23243	4	2	0	0.25396	0.82490	0.37265	-0.14213	-0.18088	0.26161	0.35434	0.00790
23244	4	3	0	-0.00215	0.53356	-0.07251	-0.17300	-0.13492	0.54062	0.02605	-0.17474
23245	4	4	1	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000
23246	4	5	1	0.52080	-0.12762	0.22539	0.31735	0.59398	-0.14539	-0.02950	0.31308
23247	4	6	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
30989	5	1	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.000000
30990	5	2	0	0.97167	0.28859	0.01335	-0.05271	0.47491	-0.09917	-0.16830	-0.064626
30991	5	3	0	0.85360	-0.23736	-0.25223	-0.07449	0.15200	0.21280	-0.28863	-0.004471
30992	5	4	1	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	1.00000	0.000000
30993	5	5	1	1.59914	-0.35338	-0.56467	0.12126	1.12701	-0.05085	-0.57923	0.095463
30994	5	6	1	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.000000

Results from Two Methods in SAS

Proc MI JM Impute then Transform

Obs	Parm	Estimate	StdErr	LCLMean	UCLMean	Probt
1	Hosp	-0.020050	0.132546	-0.28908	0.24898	0.8806
2	AgeDx	-0.016363	0.001926	-0.02014	-0.01259	<.0001
3	Stage1	0.109555	0.127272	-0.15390	0.37301	0.3983
4	Stage2	0.972660	0.105035	0.76617	1.17915	<.0001
5	Stage3	1.889572	0.154569	1.58624	2.19290	<.0001
6	Stage4	-0.194886	0.351248	-0.94089	0.55112	0.5869
7	YearDx	-0.027180	0.008945	-0.04481	-0.00955	0.0027
8	Race	-0.179914	0.063441	-0.30427	-0.05556	0.0046
9	Hosp*Stage1	-0.194767	0.178784	-0.56382	0.17428	0.2868
10	Hosp*Stage2	-0.421166	0.156632	-0.73248	-0.10985	0.0086
11	Hosp*Stage3	-0.577543	0.235492	-1.05436	-0.10073	0.0189
12	Hosp*Stage4	0.095694	0.381707	-0.68329	0.87468	0.8037

Proc MI JM Active

Obs	Parm	Estimate	StdErr	LCLMean	UCLMean	Probt
1	Hosp	0.111996	0.141570	-0.18007	0.40406	0.4366
2	AgeDx	-0.016774	0.001947	-0.02059	-0.01296	<.0001
3	Stage1	0.224637	0.126132	-0.03404	0.48332	0.0861
4	Stage2	1.158222	0.138978	0.86739	1.44905	<.0001
5	Stage3	2.108979	0.206800	1.68308	2.53488	<.0001
6	Stage4	-0.375647	0.397006	-1.20777	0.45647	0.3562
7	YearDx	-0.027291	0.008481	-0.04392	-0.01066	0.0013
8	Race	-0.193142	0.063511	-0.31763	-0.06865	0.0024
9	HospStage1	-0.328696	0.185086	-0.71408	0.05669	0.0905
10	HospStage2	-0.644039	0.178030	-1.00974	-0.27834	0.0012
11	HospStage3	-0.806307	0.222029	-1.24675	-0.36586	0.0004
12	HospStage4	0.273622	0.459300	-0.68893	1.23618	0.5585

Recommendation in SAS

- Use JM – FCS option is still experimental
- No need to round
- Use active imputation

FCS Using *Active Imputation* in Stata

```
/* ice active */

*import data
insheet using "C:\example.csv", clear

*create interaction
gen theint=hosp*stage

*active imputation
ice mastectomy hosp agedx stage yeardx race theint, saving(imp_active.dta) m(5) replace
use imp_active.dta, clear

*build logistic regression model and summarize
micombine logit mastectomy hosp agedx b0.stage yeardx race b0.theint
```

FCS Using *Active Imputation* in Stata

```
. *active imputation
. ice mastectomy hosp agedx stage yeardx race theint, saving(imp_active.dta) m(5) replace

 #missing |
   values |      Freq.     Percent        Cum.
----------+-----------------------------------
        0 |      5,341       68.94       68.94
        1 |        386        4.98       73.93
        2 |      1,395       18.01       91.93
        3 |        625        8.07      100.00
----------+-----------------------------------
    Total |      7,747      100.00

   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------------
 mastectomy |         | [No missing data in estimation sample]
       hosp |         | [No missing data in estimation sample]
      agedx |         | [No missing data in estimation sample]
     yeardx |         | [No missing data in estimation sample]
       race | logit   | mastectomy hosp agedx stage yeardx theint
      stage | mlogit  | mastectomy hosp agedx yeardx race theint
     theint | mlogit  | mastectomy hosp agedx stage yeardx race
--------------------------------------------------------------------------------

Imputing
[Perfect prediction detected: using augmlogit to impute stage]
....
Error #430 encountered while running -uvis-
I detected a problem with running uvis with command mlogit on response theint
and covariates mastectomy hosp agedx stage yeardx race.

The offending command resembled:
uvis mlogit theint mastectomy hosp agedx stage yeardx race , gen([imputed])

With mlogit, try combining categories of theint, or if appropriate, use ologit

you may wish to try the -persist- option to persist beyond this error.
dumping current data to ./_ice_dump.dta
convergence not achieved
r(430);
```

FCS Using *Active Imputation* in Stata

```
/* ice active */

*import data
insheet using "C:\example.csv", clear

*create interaction
gen theint=hosp*stage

*active imputation
ice mastectomy hosp agedx stage yeardx race theint, saving(imp_active.dta) m(5) replace persist
use imp_active.dta, clear

*build logistic regression model and summarize
micombine logit mastectomy hosp agedx b0.stage yeardx race b0.theint
```

FCS Using *Active Imputation* in Stata

```
. *active imputation
. ice mastectomy hosp agedx stage yeardx race theint, saving(imp_active.dta) m(5) replace persist

  #missing |
    values |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |      5,341       68.94       68.94
         1 |        386        4.98       73.93
         2 |      1,395       18.01       91.93
         3 |        625        8.07      100.00
-----------+-----------------------------------
     Total |      7,747      100.00

   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------
 mastectomy |         | [No missing data in estimation sample]
       hosp |         | [No missing data in estimation sample]
      agedx |         | [No missing data in estimation sample]
     yeardx |         | [No missing data in estimation sample]
       race | logit   | mastectomy hosp agedx stage yeardx theint
      stage | mlogit  | mastectomy hosp agedx yeardx race theint
     theint | mlogit  | mastectomy hosp agedx stage yeardx race
------------------------------------------------------------------------

Imputing
[Perfect prediction detected: using augmlogit to impute stage]
..........1
[Perfect prediction detected: using augmlogit to impute stage]
..........2
[Perfect prediction detected: using augmlogit to impute stage]
......
[persist option: ignoring error #430, not updating theint in cycle 7]
....3
[Perfect prediction detected: using augmlogit to impute stage]
..
[persist option: ignoring error #430, not updating theint in cycle 3]
......
[persist option: ignoring error #430, not updating theint in cycle 9]
..4
[Perfect prediction detected: using augmlogit to impute stage]
.........
[persist option: ignoring error #430, not updating theint in cycle 10]
.5
(note: file imp_active.dta not found)
file imp_active.dta saved
```

FCS Using *Active Imputation* in Stata

```
. list in 7748/7753
```

	id	hosp	mastec~y	agedx	yeardx	stage	race	theint	_mi	_mj
7748.	1	1	1	64	2008	1	1	1	1	1
7749.	2	0	0	47	1999	1	1	0	2	1
7750.	3	0	0	80	2009	1	1	0	3	1
7751.	4	1	1	55	2003	3	1	3	4	1
7752.	5	1	0	60	2009	1	1	1	5	1
7753.	6	1	0	58	2009	1	1	1	6	1

```
. tab stage theint
```

	theint					
stage	0	1	2	3	4	Total
0	11,159	22	0	0	0	11,181
1	8,153	7,028	7	0	0	15,188
2	5,846	3	6,576	3	0	12,428
3	1,481	0	2	2,578	4	4,065
4	471	5	1	2	1,121	1,600
Total	27,110	7,058	6,586	2,583	1,125	44,462

FCS Using *Passive Imputation* in Stata

```
/* ice passive */

*import data
insheet using "C:\example.csv", clear

*create interaction
gen theint=hosp*stage

*passive imputation
ice mastectomy hosp agedx stage yeardx race theint, passive(theint:hosp*stage) saving(imp_passive.dta) m(5) rep
use imp_passive.dta, clear

*build logistic regression model and summarize
micombine logit mastectomy hosp agedx b0.stage yeardx race b0.theint
```

FCS Using *Passive Imputation* in Stata

```
/* ice passive */

*import data
insheet using "C:\example.csv", clear

*create interaction
gen theint=hosp*stage

*passive imputation
ice mastectomy hosp agedx stage yeardx race theint, passive(theint:hosp*stage) saving(imp_passive.dta) m(5) rep
use imp_passive.dta, clear

*build logistic regression model and summarize
micombine logit mastectomy hosp agedx b0.stage yeardx race b0.theint
```

FCS Using *Passive Imputation* in Stata

```
. *passive imputation
. ice mastectomy hosp agedx stage yeardx race theint, passive(theint:hosp*stage) saving(imp_passive.dta) m(5) re

   #missing |
     values |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,341       68.94       68.94
          1 |        386        4.98       73.93
          2 |      1,395       18.01       91.93
          3 |        625        8.07      100.00
------------+-----------------------------------
      Total |      7,747      100.00

   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------------
 mastectomy |         | [No missing data in estimation sample]
       hosp |         | [No missing data in estimation sample]
      agedx |         | [No missing data in estimation sample]
     yeardx |         | [No missing data in estimation sample]
       race | logit   | mastectomy hosp agedx stage yeardx theint
      stage | mlogit  | mastectomy hosp agedx yeardx race
     theint |         | [Passively imputed from hosp*stage]
------------------------------------------------------------------------------

Imputing ..........1..........2..........3..........4..........5
file imp_passive.dta saved

. use imp_passive.dta, clear
```

FCS Using *Passive Imputation* in Stata

```
. *passive imputation
. ice mastectomy hosp agedx stage yeardx race theint, passive(theint:hosp*stage) saving(imp_passive.dta) m(5) re

   #missing |
     values |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,341       68.94       68.94
          1 |        386        4.98       73.93
          2 |      1,395       18.01       91.93
          3 |        625        8.07      100.00
------------+-----------------------------------
      Total |      7,747      100.00

   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------------
 mastectomy |         | [No missing data in estimation sample]
       hosp |         | [No missing data in estimation sample]
      agedx |         | [No missing data in estimation sample]
     yeardx |         | [No missing data in estimation sample]
       race | logit   | mastectomy hosp agedx stage yeardx theint
      stage | mlogit  | mastectomy hosp agedx yeardx race
     theint |         | [Passively imputed from hosp*stage]
------------------------------------------------------------------------------

Imputing ..........1..........2..........3..........4..........5
file imp_passive.dta saved

. use imp_passive.dta, clear
```

FCS Using *Passive Imputation* in Stata

```
. list in 7748/7753
```

	id	hosp	mastec~y	agedx	yeardx	stage	race	theint	_mi	_mj
7748.	1	1	1	64	2008	1	1	1	1	1
7749.	2	0	0	47	1999	0	1	0	2	1
7750.	3	0	0	80	2009	4	1	0	3	1
7751.	4	1	1	55	2003	3	1	3	4	1
7752.	5	1	0	60	2009	3	1	3	5	1
7753.	6	1	0	58	2009	1	1	1	6	1

```
. tab stage theint
```

	theint					
stage	0	1	2	3	4	Total
0	9,815	0	0	0	0	9,815
1	8,145	7,639	0	0	0	15,784
2	5,882	0	7,118	0	0	13,000
3	1,487	0	0	2,725	0	4,212
4	455	0	0	0	1,196	1,651
Total	25,784	7,639	7,118	2,725	1,196	44,462

Results from the Two Methods in Ice

```
. *active imputation
. micombine logit mastectomy hosp agedx

Multiple imputation parameter estimates
```

mastectomy	Coef.	Std. Err.
mastectomy		
hosp	-.1109177	.1080544
agedx	-.0173094	.0019449
stage		
0	0	(empty)
1	.1335282	.1025934
2	1.103616	.1100777
3	2.087673	.23944
4	-.4561715	.3856415
yeardx	-.0285017	.0084972
race	-.1872521	.0709139
theint		
0	0	(empty)
1	.0284646	.1421275
2	-.3623982	.1397179
3	-.5782175	.2582078
4	.5624794	.4661046
_cons	57.34566	17.03261

7747 observations (imputation 1).

```
. *passive imputation
. micombine logit mastectomy hosp agedx

Multiple imputation parameter estimates
```

mastectomy	Coef.	Std. Err.
mastectomy		
hosp	-.0400464	.1114586
agedx	-.0165242	.0019731
stage		
0	0	(empty)
1	.1033335	.1112763
2	.975499	.1098498
3	1.980216	.1814719
4	-.3110553	.2978437
yeardx	-.0279643	.0084428
race	-.1934281	.066012
theint		
0	0	(empty)
1	-.1572001	.148122
2	-.4194108	.1574878
3	-.6265811	.2145474
4	.2444306	.3547756
_cons	56.27944	16.92546

7747 observations (imputation 1).

Recommendation in Stata

- Study the output
  - Command
  - Predictor equation
- Use passive imputation
- Increase the number of imputation when using the persist option
- If using passive imputation, make sure all levels of interaction variable are present
- May be worthwhile to compare two approaches

FCS Using *Active Imputation* in R

```
### Active imputation

# Select variables used for imputation
toimp <- mydata[c("Hosp","Mastectomy", "AgeDx", "stage", "YearDx", "race")]

# Create interaction
toimp$theint <- toimp$Hosp * toimp$stage

# Change to factor variables
toimp$stage <- as.factor(toimp$stage)
toimp$race <- as.factor(toimp$race)
toimp$theint <- as.factor(toimp$theint)

# Impute
imp_active <- mice(toimp, m=5, method=c("", "", "", "polyreg", "", "logreg", "polyreg"))
summary(imp_active)

# Summarize scientific model
result <- pool(with(data=imp_active,
  glm(Mastectomy ~ Hosp + AgeDx + stage + YearDx + race + theint, family="binomial")))
summary(result)
```

FCS Using *Active Imputation* in R

```
> summary(imp_active)
Multiply imputed data set
Call:
mice(data = toimp, m = 5, method = c("", "", "", "polyreg", "",
    "logreg", "polyreg"))
Number of multiple imputations:  5
Missing cells per column:
      Hosp Mastectomy      AgeDx      stage     YearDx       race     theint
         0          0          0       2020          0       1011       2020
Imputation methods:
      Hosp Mastectomy      AgeDx      stage     YearDx       race     theint
        ""         ""         ""  "polyreg"         ""   "logreg"  "polyreg"
VisitSequence:
 stage   race theint
     4      6      7
PredictorMatrix:
           Hosp Mastectomy AgeDx stage YearDx race theint
Hosp          0          0     0     0      0    0      0
Mastectomy    0          0     0     0      0    0      0
AgeDx         0          0     0     0      0    0      0
stage         1          1     1     0      1    1      1
YearDx        0          0     0     0      0    0      0
race          1          1     1     1      1    0      1
theint        1          1     1     1      1    1      0
Random generator seed value:  NA
```

FCS Using *Active Imputation* in R

```
> table(complete(imp_active)$stage, complete(imp_active)$theint)

       0    1    2    3    4
  0 1966    0    1    1    0
  1 1462 1173    0    0    1
  2 1019    0 1136    0    0
  3  259    0    0  465    0
  4   78    0    0    0  186
```

FCS Using *Passive Imputation* in R

```
## Passive imputation

# Make dummy variables for interaction
toimp$theint1 <- ifelse(toimp$theint == 1, 1, 0)
toimp$theint2 <- ifelse(toimp$theint == 2, 1, 0)
toimp$theint3 <- ifelse(toimp$theint == 3, 1, 0)
toimp$theint4 <- ifelse(toimp$theint == 4, 1, 0)

# Select variables in the imputation model
toimp2 <- toimp[c("Hosp","Mastectomy", "AgeDx", "stage", "YearDx", "race"
                 , "theint1", "theint2", "theint3", "theint4")]

# Dry run to get meth and pred
ini <- mice(toimp2, max=0, print=FALSE)

# Save the methods and specify to passively impute the interactions
meth <- ini$meth
meth["theint1"] <- "~I(Hosp*stage.1)"
meth["theint2"] <- "~I(Hosp*stage.2)"
meth["theint3"] <- "~I(Hosp*stage.3)"
meth["theint4"] <- "~I(Hosp*stage.4)"

# Remove interactions from predicting main effecs
pred <- ini$pred
pred[c("stage"), c("theint1", "theint2", "theint3", "theint4")] <- 0

# Impute
imp_passive <- mice(toimp2, m=5, method=meth, pred=pred, print=FALSE)

# Summarize scientific model
result <- pool(with(data=imp_passive,
  glm(Mastectomy ~ Hosp + AgeDx + stage + YearDx + race
      + theint1 + theint2 + theint3 + theint4, family="binomial")))
summary(result)
```

FCS Using *Passive Imputation* in R

```
> ini$meth
     Hosp Mastectomy      AgeDx      stage     YearDx       race    theint1    theint2    theint3    theint4
       ""         ""         "" "polyreg"         ""  "logreg"     "pmm"      "pmm"      "pmm"      "pmm"
> ini$pred
           Hosp Mastectomy AgeDx stage YearDx race theint1 theint2 theint3 theint4
Hosp          0          0     0     0      0    0       0       0       0       0
Mastectomy    0          0     0     0      0    0       0       0       0       0
AgeDx         0          0     0     0      0    0       0       0       0       0
stage         1          1     1     0      1    1       1       1       1       1
YearDx        0          0     0     0      0    0       0       0       0       0
race          1          1     1     1      1    0       1       1       1       1
theint1       1          1     1     1      1    1       0       1       1       1
theint2       1          1     1     1      1    1       1       0       1       1
theint3       1          1     1     1      1    1       1       1       0       1
theint4       1          1     1     1      1    1       1       1       1       0
```

FCS Using *Passive Imputation* in R

```
> summary(imp_passive)
Multiply imputed data set
Call:
mice(data = toimp2, m = 5, method = meth, predictorMatrix = pred,
    printFlag = FALSE)
Number of multiple imputations:  5
Missing cells per column:
      Hosp Mastectomy      AgeDx      stage     YearDx       race    theint1    theint2    theint3    theint4
         0          0          0       2020          0       1011       2020       2020       2020       2020
Imputation methods:
                Hosp           Mastectomy                AgeDx                stage               YearDx                 race              theint1
                  ""                   ""                   ""            "polyreg"                   ""             "logreg" "~I(Hosp*stage.1)"
           theint2              theint3              theint4
"~I(Hosp*stage.2)" "~I(Hosp*stage.3)" "~I(Hosp*stage.4)"
VisitSequence:
  stage     race theint1 theint2 theint3 theint4
      4        6       7       8       9      10
PredictorMatrix:
           Hosp Mastectomy AgeDx stage YearDx race theint1 theint2 theint3 theint4
Hosp          0          0     0     0      0    0       0       0       0       0
Mastectomy    0          0     0     0      0    0       0       0       0       0
AgeDx         0          0     0     0      0    0       0       0       0       0
stage         1          1     1     0      1    1       0       0       0       0
YearDx        0          0     0     0      0    0       0       0       0       0
race          1          1     1     1      1    0       1       1       1       1
theint1       1          1     1     1      1    1       0       1       1       1
theint2       1          1     1     1      1    1       1       0       1       1
theint3       1          1     1     1      1    1       1       1       0       1
theint4       1          1     1     1      1    1       1       1       1       0
Random generator seed value:  NA
```

FCS Using *Passive Imputation* in R

```
> head(complete(imp_passive))
  Hosp Mastectomy AgeDx stage YearDx race theint1 theint2 theint3 theint4
1    1          1    64     1   2008    1       1       0       0       0
2    0          0    47     0   1999    0       0       0       0       0
3    0          0    80     1   2009    1       0       0       0       0
4    1          1    55     3   2003    1       0       0       1       0
5    1          0    60     2   2009    1       0       1       0       0
6    1          0    58     1   2009    1       1       0       0       0
> table(complete(imp_passive)$theint1, complete(imp_passive)$stage)

       0    1    2    3    4
  0 1713 1415 2300  717  274
  1    0 1328    0    0    0
> table(complete(imp_passive)$theint2, complete(imp_passive)$stage)

       0    1    2    3    4
  0 1713 2743 1045  717  274
  1    0    0 1255    0    0
> table(complete(imp_passive)$theint3, complete(imp_passive)$stage)

       0    1    2    3    4
  0 1713 2743 2300  259  274
  1    0    0    0  458    0
> table(complete(imp_passive)$theint4, complete(imp_passive)$stage)

       0    1    2    3    4
  0 1713 2743 2300  717   73
  1    0    0    0    0  201
```

Results from the Two Methods in Mice

```
> summary(result_active)
                est       se      t     df Pr(>|t|)   lo 95   hi 95
(Intercept) 57.6903 16.87418  3.419 5225.3 0.000634 24.6098 90.7707
Hosp        -0.1326  0.10195 -1.301 5666.8 0.193426 -0.3325  0.0673
AgeDx       -0.0177  0.00194 -9.084 5502.8 0.000000 -0.0215 -0.0138
stage2       0.1418  0.10243  1.384  216.3 0.167702 -0.0601  0.3437
stage3       1.0763  0.10915  9.861  113.6 0.000000  0.8601  1.2925
stage4       2.1375  0.18380 11.630  101.6 0.000000  1.7729  2.5021
stage5      -0.4399  0.32853 -1.339   91.7 0.183880 -1.0924  0.2126
YearDx      -0.0287  0.00842 -3.405 5266.6 0.000667 -0.0452 -0.0122
race2       -0.2028  0.06612 -3.067  362.3 0.002321 -0.3328 -0.0728
theint2      0.0771  0.13658  0.565 1235.1 0.572418 -0.1908  0.3451
theint3     -0.3142  0.13615 -2.308 2425.0 0.021084 -0.5812 -0.0473
theint4     -0.7338  0.23555 -3.115   57.9 0.002857 -1.2054 -0.2623
theint5      0.6219  0.40605  1.532   42.6 0.133020 -0.1972  1.4410
> summary(result_passive)
                est       se      t     df Pr(>|t|)   lo 95   hi 95
(Intercept) 55.2286 17.04365  3.240 1939.1  0.00121 21.8027 88.6544
Hosp        -0.0334  0.11867 -0.282  113.0  0.77870 -0.2685  0.2017
AgeDx       -0.0165  0.00195 -8.468 3961.4  0.00000 -0.0203 -0.0127
stage2       0.0862  0.12369  0.696   23.2  0.49303 -0.1696  0.3419
stage3       0.9823  0.10518  9.339  204.1  0.00000  0.7749  1.1897
stage4       2.0390  0.18385 11.090  109.6  0.00000  1.6746  2.4033
stage5      -0.4783  0.34063 -1.404   54.4  0.16595 -1.1611  0.2045
YearDx      -0.0274  0.00850 -3.228 1917.9  0.00127 -0.0441 -0.0108
race2       -0.1730  0.06825 -2.535  160.4  0.01221 -0.3078 -0.0382
theint1     -0.1674  0.16427 -1.019   41.3  0.31403 -0.4991  0.1642
theint2     -0.4433  0.14526 -3.052  311.5  0.00247 -0.7291 -0.1575
theint3     -0.6575  0.22347 -2.942  179.5  0.00369 -1.0985 -0.2165
theint4      0.3676  0.37828  0.972  100.1  0.33350 -0.3829  1.1181
```

Recommendation in R

- Convert all categorical variables to factor variables using 'as.factor()'
- Study the imputation object, especially if using passive imputation
  - Method vector
  - Predictor matrix
- Use active or passive imputation
  - However, code for active imputation is far more simple
- May be worthwhile to compare two approaches

Comparison Across Software

Odds Ratios		SAS JM		Stata ice		R mice	
	CC	Active	Imp $\rightarrow$ Trf	Active	Passive	Active	Passive
Hosp A Stage 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hosp A Stage 1	1.17	1.12	1.25	1.14	1.11	1.15	1.09
Hosp A Stage 2	2.99	2.65	3.18	3.02	2.65	2.93	2.67
Hosp A Stage 3	8.73	6.62	8.24	8.07	7.24	8.48	7.68
Hosp A Stage 4	0.69	0.82	0.69	0.63	0.73	0.64	0.62
Hosp B Stage 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hosp B Stage 1	1.00	0.92	0.90	1.18	0.95	1.24	0.92
Hosp B Stage 2	1.85	1.74	1.67	2.10	1.74	2.14	1.71
Hosp B Stage 3	4.13	3.71	3.68	4.52	3.87	4.07	3.98
Hosp B Stage 4	0.33	0.91	0.90	1.11	0.94	1.20	0.90
Age	0.99	0.98	0.98	0.98	0.98	0.98	0.98
White vs Other Race	0.79	0.84	0.82	0.83	0.82	0.82	0.84
Year	0.97	0.98	0.97	0.97	0.97	0.97	0.97

Conclusion

- Do not treat imputed values as real – Summarized estimates and inference are the most important
    - e.g. Do not create Table 1 from the imputed data
- Make sure you know what method/option is used
    - JM or FCS?
    - Active or Passive?
    - Interactions?
- Perform sensitivity analysis
    - Try a different approach
    - Increase the number of imputations
- Keep up with the literature
- Study the software manual

References

Rubin DB.
Multiple imputation for nonresponse in surveys.
J. Wiley &Sons, New York, 1987.

Bernaards CA.
Robustness of a multivariate normal approximation for imputation of incomplete binary data.
Statistics in Medicine, 26(6):1368–1382, 2007.

Yucel RM, He Y, Zaslavsky AM.
Using calibration to improve rounding in imputation.
American Statistician, 62(2):125–129, 2008.

Demirtas H.
Rounding Strategies for Multiply Imputed Binary Data.
Biometrical Journal, 51(4):677–688, 2009.

Allison PD.
Missing data.
Thousand Oaks, CA: Sage Publications, 2001.

References cont.

Song R, Harrison KM, Hanson DKL, Hall HI.
Correction of Bias in Imputing Missing Values of Categorical Variables.
Communications in Statistics-Theory and Methods, 39(2):350–362, 2010.

Yucel RM, He Y, Zaslavsky AM.
Gaussian-based routines to impute categorical variables in health surveys.
Statistics in Medicine, 30(29):3447–3460, 2011.

van Buuren S.
Multiple imputation of discrete and continuous data by fully conditional specification.
Statistical Methods in Medical Research, 16(3):219–242, 2007.

von Hippel PT.
How to impute interactions, squares and other transformed variables.
Sociological Methodology, 39:265–291, 2009.

Horton NJ, Lipsitz SR, Parzen M.
A potential for bias when rounding in multiple imputation.
American Statistician, 57(4):229–232, 2003.