# 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

May 7, 2013

# Outline

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

# Motivation: Study of Factors Related to Breast Cancer Care

- Data were systematically missing and correlated with factors such as
  - year of diagnosis
  - patient affiliated institution
  - severity of disease
- We were particularly interested in synergistic associations
- Application of multiple imputation using different software in the presence of interaction gave different answers

# What should we do about missing data?

- Prevent it!
- Likelihood-based methods that specify the joint distribution of both observed and missing data are efficient
    - Little and Rubin. Statistical analysis with missing data. Wiley-Interscience. 1987.

However,

- Likelihood-based methods are complicated to impliment
- MI-based methods provide estimates with similarly desirable statistical properties
- MI is easily accessible through mainstream software

# 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

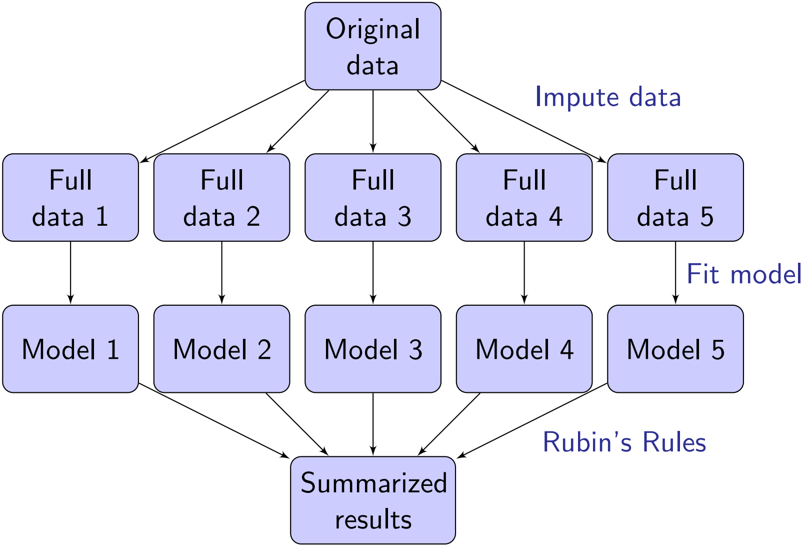
# Some background: Multiple Imputation (MI)

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

- Imputation method to employ
- 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

- Imputation method to employ
- 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 |

→

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. Round the values (rounding at 0.5 is not recommended for binary variables, Horton 2003)
2. Use these values (not an option for Race)
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 | | | | | | Imputed Data 1 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | Age | Male | Black | Asian | | ID | Age | Male | Black | Asian |
| 1 | 60 | 1 | 1 | 0 | | 1 | 60 | 1 | 1 | 0 |
| 2 | 50 | . | 0 | 1 | | 2 | 50 | 1.2 | 0 | 1 |
| 3 | . | 0 | . | . | → | 3 | 44.4 | 0 | 0.1 | 0.2 |
| 4 | 40 | 0 | . | . | | 4 | 40 | 0 | 0.3 | 1.5 |
| 5 | 55 | 1 | 0 | 0 | | 5 | 55 | 1 | 0 | 0 |

Options

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

# 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 |

→

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
- Transform, impute, then transform again

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

| 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 | 40 |
| 3 | 44.4 | 0 | 2 |
| 4 | 40 | 0 | 0 |
| 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

- van Buuren recommends FCS over JM, but he did not include interaction terms in his simulations
- von Hippel claims passive imputation is biased, but his simulations only included interactions between two continuous variables, or interactions between a binary and a continuous variable

# Categorical Interaction Variables

## JM

1. Convert interaction and other categorical variables 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

# Categorical Interaction Variables

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

Scientific Model:

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

# Imputation Model for JM and FCS

Scientific Model:

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

Imputation Model in JM:

$$Y, X_W, X_B, X_{DA}, X_{DB}, X_W X_{DA}, X_W X_{DB}, X_B X_{DA}, X_B X_{DB}, Z$$

Imputation Model in FCS:

$$Y, X_{Race}, X_{Drug}, I_{Race \times Drug}, Z$$

*$I_{Race \times Drug}$ has 5 categories

# 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

Stata also has 'mi' in which 'mi impute mvn' performs JM

# 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`, the imputed values of `int` will only have 4 levels (0, 1, 2, 4)

# 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

# Simulation Scenarios

| *Scenario* | *Outcome* | $X_1$ | $X_2$ | *Missing proportion* |
|---|---|---|---|---|
| 1*a* | Binary | Binary | Binary | 20% |
| 1*b* | Binary | Binary | Binary | 40% |
| 1*c* | Continuous | Binary | Binary | 20% |
| 1*d* | Continuous | Binary | Binary | 40% |
| 2*a* | Binary | Binary | 3-level categorical | 20% |
| 2*b* | Binary | Binary | 3-level categorical | 40% |
| 2*c* | Continuous | Binary | 3-level categorical | 20% |
| 2*d* | Continuous | Binary | 3-level categorical | 40% |
| 3*a* | Binary | 3-level categorical | 3-level categorical | 20% |
| 3*b* | Binary | 3-level categorical | 3-level categorical | 40% |
| 3*c* | Continuous | 3-level categorical | 3-level categorical | 20% |
| 3*d* | Continuous | 3-level categorical | 3-level categorical | 40% |

‡ m=10 imputations

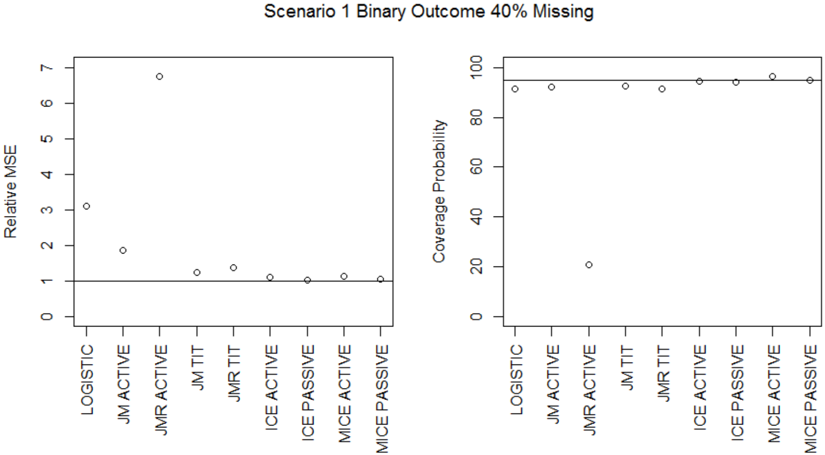
# MI Choices

| *Software* | *JM* | | *FCS* | |
|---|---|---|---|---|
| | *No round* | *Round** | *Active* | *Passive* |
| PROC MI | JM ACTIVE | JMR ACTIVE | LOGISTIC | |
| (SAS) | JM TIT | JMR TIT | DISCRIM | |
| ice (Stata) | | | ICE ACTIVE | ICE PASSIVE |
| mice (R) | | | MICE ACTIVE | MICE PASSIVE |

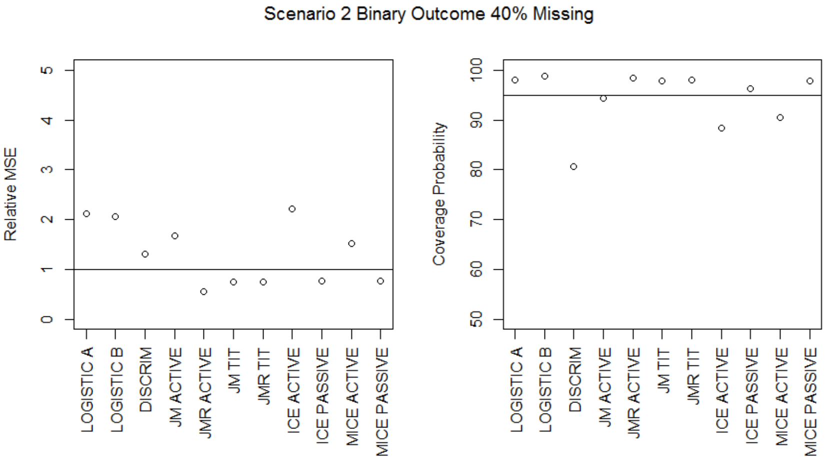
*If binary, round at 0.5 → Shown to be biased (Horton, et al.)

*If categorical, round using Allison's method for nominal categorical variables

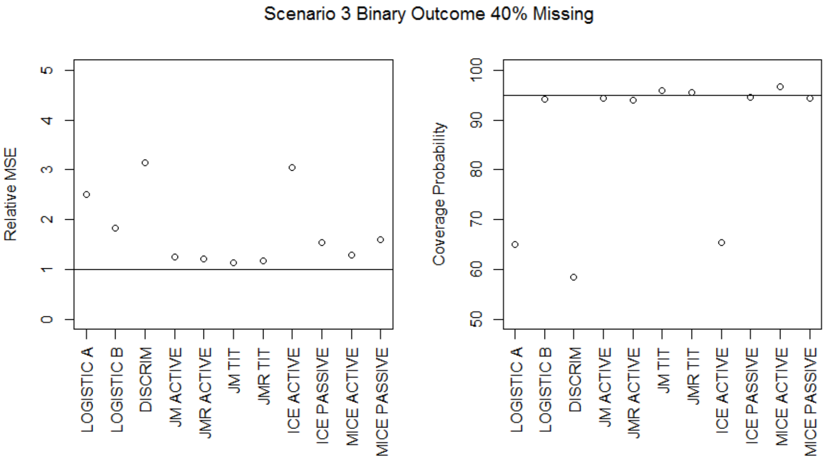
# Simulation Results



# Simulation Results



# Simulation Results



# 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 × Stage

# Comparison Across Software

| Odds Ratios | | SAS JM | | Stata ice | | R mice | |
|---|---|---|---|---|---|---|---|
| | CC | Active | TIT | 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 |

# Recommendation

- SAS Proc MI
  - Use JM – FCS option is still experimental and results are biased
  - No need to round
  - Use active imputation
- Stata ice
  - 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
- R mice
  - Study the imputation object, especially if using passive imputation (Method vector, Predictor matrix)
  - Use active or passive imputation
  - May be worthwhile to compare two approaches

# 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.

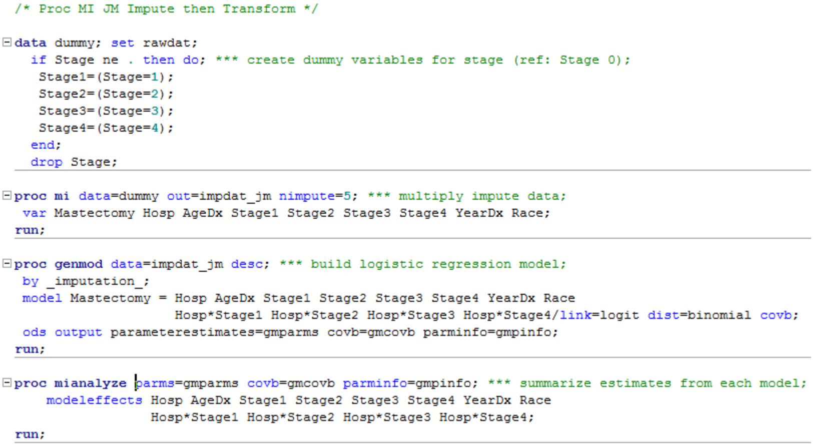
Yucel RM, He Y, Zaslavsky AM.
Gaussian-based routines to impute categorical variables in health surveys.
*Statistics in Medicine*, 30(29):3447–3460, 2011.

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.

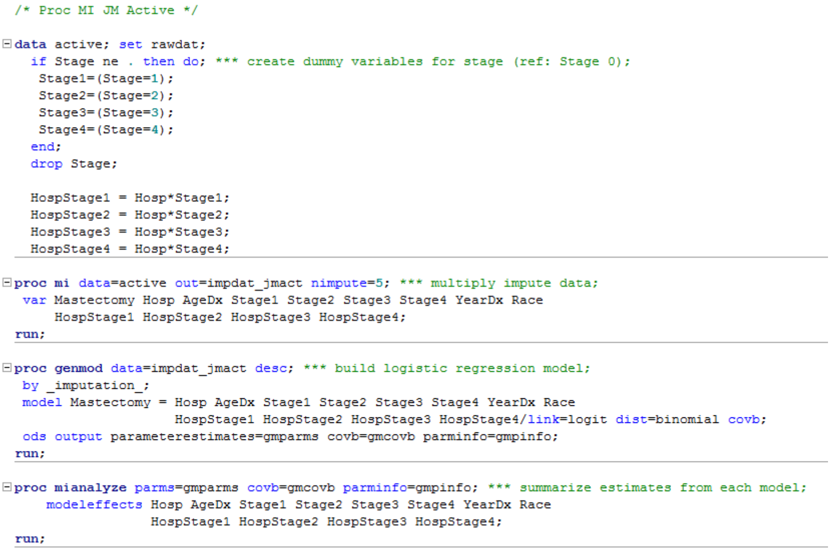
Horton NJ, Lipsitz SR, Parzen M.
A potential for bias when rounding in multiple imputation.
*American Statistician*, 57(4):229–232, 2003.

# JM with *Impute then Transform* Approach in SAS



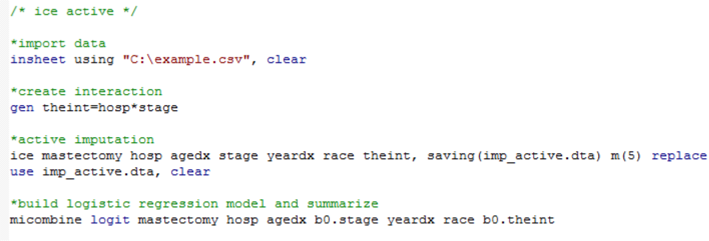
```
/* Proc MI JM Impute then Transform */

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 Using *Active Imputation* in SAS



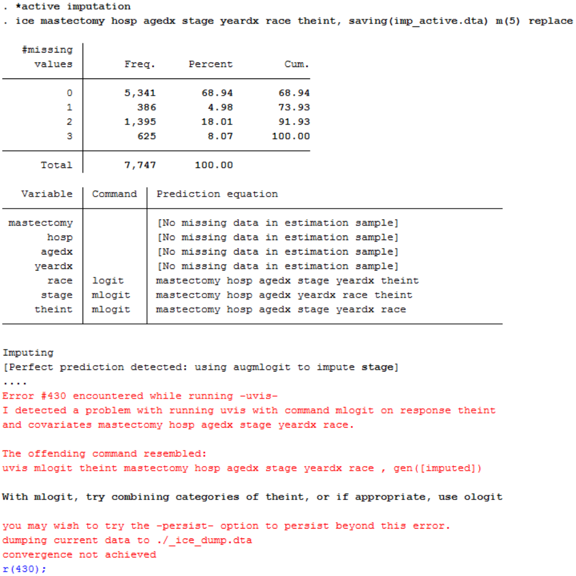
```
/* Proc MI JM Active */

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;
```

# FCS Using *Active Imputation* in Stata



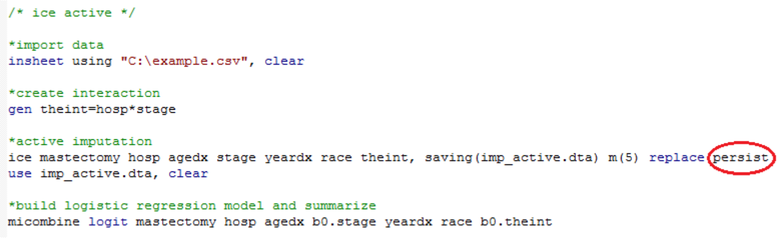
```
/* ice active */

*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



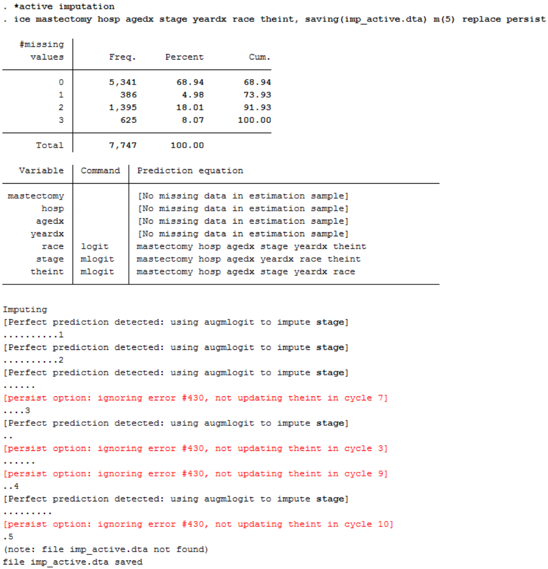
```
. *active imputation
. ice mastectomy hosp agedx stage yeardx race theint, saving(imp_active.dta) m(5) replace

  #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



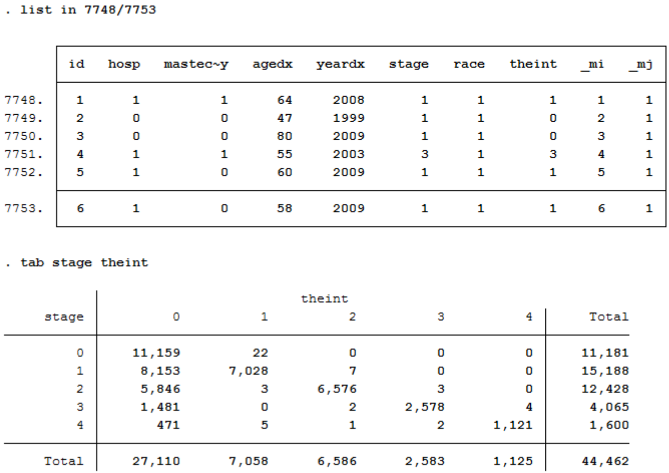
```
/* ice active */

*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



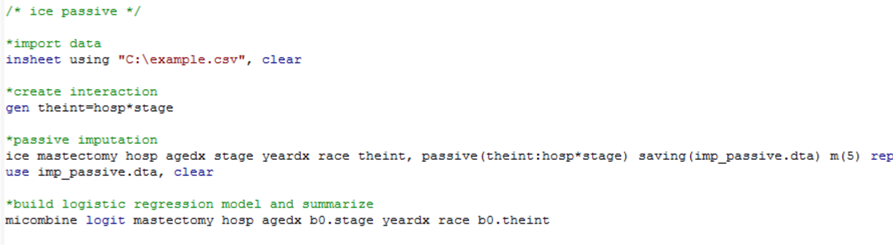
```
. *active imputation
. ice mastectomy hosp agedx stage yeardx race theint, saving(imp_active.dta) m(5) replace persist

  #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



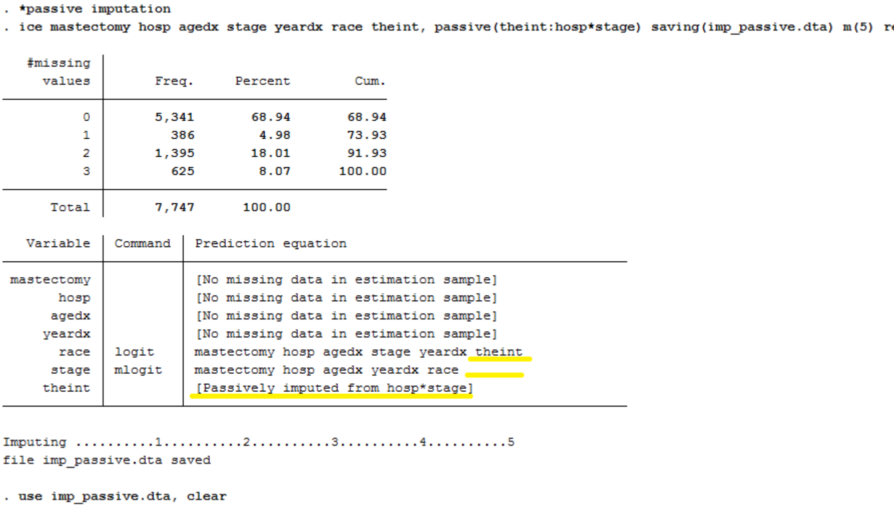
```
. list in 7748/7753

      +----------------------------------------------------------------------------------+
      | 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



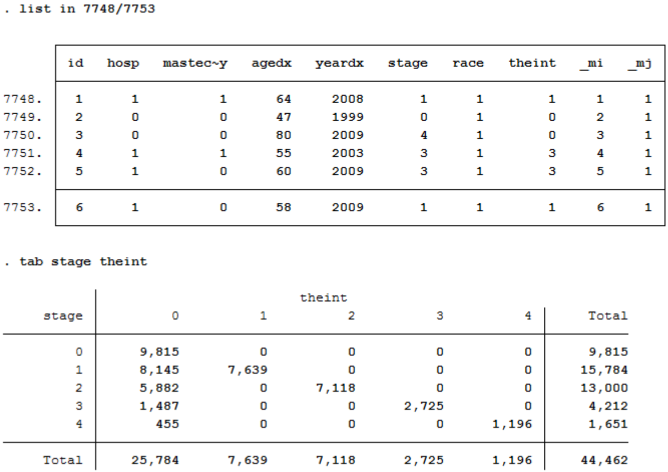
```
/* ice passive */

*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



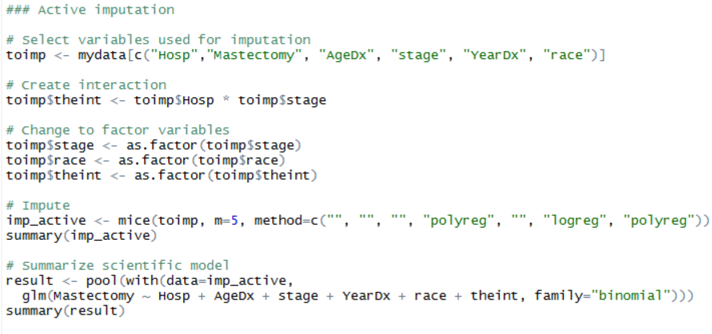
```
. *passive imputation
. ice mastectomy hosp agedx stage yeardx race theint, passive(theint:hosp*stage) saving(imp_passive.dta) m(5) re

   #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



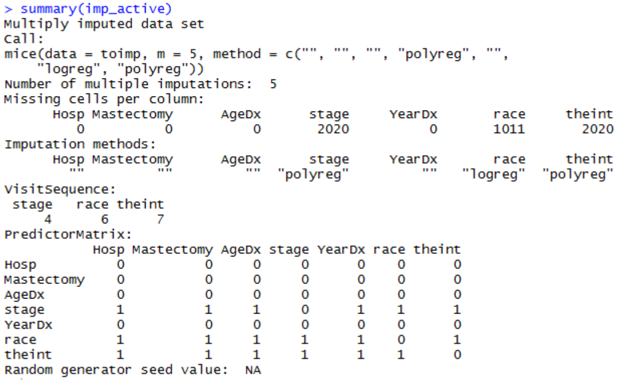
```
. list in 7748/7753

      +--------------------------------------------------------------------------+
      | 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 
```

# FCS Using *Active Imputation* in R



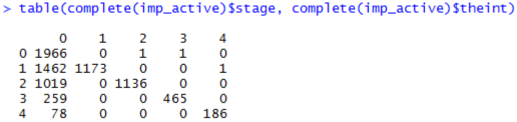
```r
### Active imputation

# 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



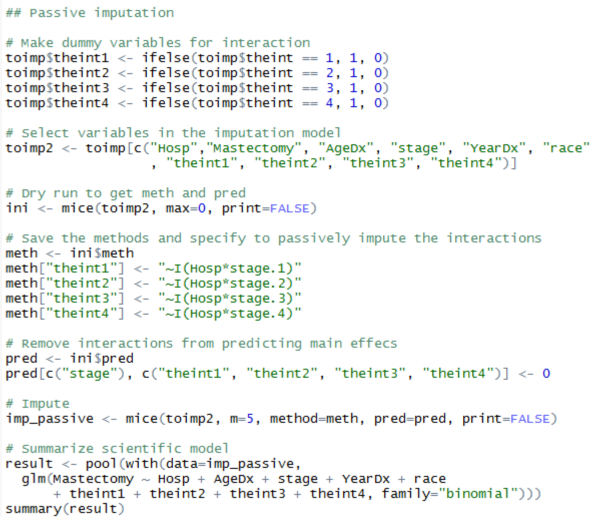
```
> summary(imp_active)
Multiply imputed data set
Call:
mice(data = toimp, m = 5, method = c("", "", "", "polyreg", "",
    "logreg", "polyreg"))
Number of multiple imputations:  5
Missing cells per column:
      Hosp Mastectomy      AgeDx      stage     YearDx       race     theint
         0          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



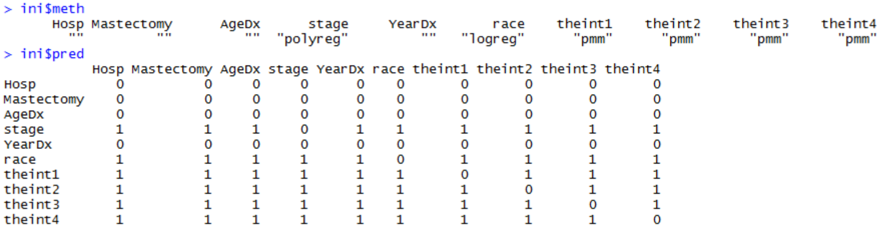
```
> table(complete(imp_active)$stage, complete(imp_active)$theint)

       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



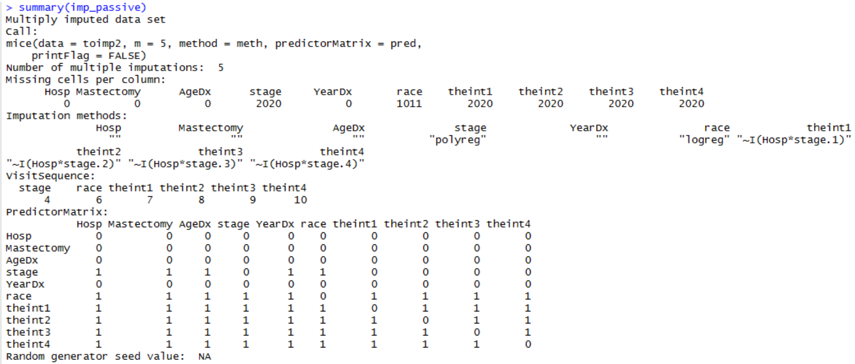
```
## Passive imputation

# 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



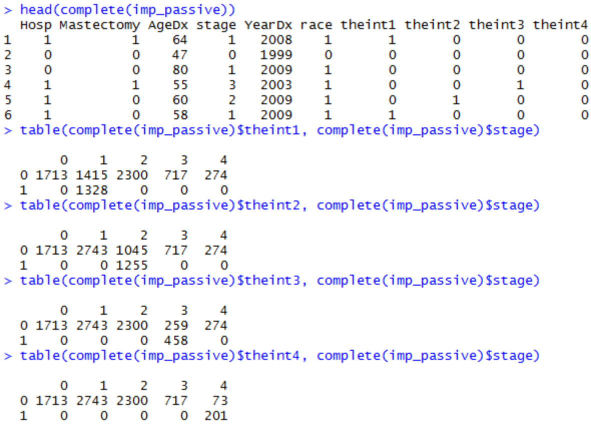
```
> ini$meth
     Hosp Mastectomy      AgeDx      stage     YearDx       race    theint1    theint2    theint3    theint4
       ""         ""         ""  "polyreg"         ""   "logreg"      "pmm"      "pmm"      "pmm"      "pmm"
> ini$pred
           Hosp Mastectomy AgeDx stage YearDx race theint1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