Multiple Imputation

VGSM, Chapter 11

Dave Glidden

17 October 2017

Last Week

- Discussed developing models for imputing
- Look at Stata's capabilities
- Implemented imputations in Stata

UNOS Dataset

- Data on 9775 pediatric kidney transplants from 1990-2002
- Followed for transplant outcomes
  465 deaths
- Predictors: age, age of donor, year, HLA loci, txtype
- Cold ischemia time: missing in 23%

Multiple Imputation

1. Assume data MAR
2. Develop predictive model for missing data
3. Sample from predictive model
4. Make multiple complete datasets
5. Analyze datasets
6. Synthesize results

Chained Syntax

- mi register imputed hlamat age_don age cold_isc prevtx death txtype
  register everything
- mi impute chained (pmm) cold_isc (ologit) hlamat (regress) age_don (regress) age (logit) prevtx = death txtype, add(50) rseed(897)
- *(method) variable* = *nonmissing*
- Very flexible syntax

Chained Output

```
. mi impute chained (pmm) cold_isc (ologit) hlamat (regress) age_don (regress) age (logit)
prevtx = death txtype, add(20) rseed(897) dots

Conditional models:
               age: regress age i.prevtx age_don i.hlamat cold_isc death txtype
            prevtx: logit prevtx age age_don i.hlamat cold_isc death txtype
           age_don: regress age_don age i.prevtx i.hlamat cold_isc death txtype
            hlamat: ologit hlamat age i.prevtx age_don cold_isc death txtype
          cold_isc: pmm cold_isc age i.prevtx age_don i.hlamat death txtype

Performing chained iterations:
  imputing m=1 through m=20 .........10.........20 done
```

mi impute reply

```
Multivariate imputation                     Imputations =       20
Chained equations                                 added =       20
Imputed: m=1 through m=20                       updated =        0

Initialization: monotone                     Iterations =      200
                                                burn-in =       10

          cold_isc: predictive mean matching
            hlamat: ordered logistic regression
           age_don: linear regression
               age: linear regression
            prevtx: logistic regression

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
          cold_isc |       7525         2250      2250 |      9775
            hlamat |       9541          234       234 |      9775
           age_don |       9662          113       113 |      9775
               age |       9766            9         9 |      9775
            prevtx |       9702           73        73 |      9775
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

MI Dataset

```
tabulate _mi_m
```

_mi_m	Freq.	Percent	Cum.
0	9,367	18.82	18.82
1	2,020	4.06	22.88
2	2,020	4.06	26.94
3	2,020	4.06	31.00
4	2,020	4.06	35.06
5	2,020	4.06	39.12
6	2,020	4.06	43.18
7	2,020	4.06	47.23
8	2,020	4.06	51.29
9	2,020	4.06	55.35
10	2,020	4.06	59.41
11	2,020	4.06	63.47
12	2,020	4.06	67.53
13	2,020	4.06	71.59
14	2,020	4.06	75.65
15	2,020	4.06	79.71
16	2,020	4.06	83.76
17	2,020	4.06	87.82
18	2,020	4.06	91.88
19	2,020	4.06	95.94
20	2,020	4.06	100.00
Total	49,767	100.00	

original data

imputation “10”

9,347 record dataset becomes 49,767 records

Multiple Imputation

1. Assume data MAR
2. Develop predictive model for missing data
3. Sample from predictive model
4. Make multiple complete datasets
5. Analyze datasets
6. Synthesize results

Source of Uncertainty

- Predictive distn of cold_isc
  - $\text{cold\_isc}_i = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + .. \beta_p * x_p + \varepsilon_i$
- Sources of uncertainty
  1. prediction: ε_i
  2. parameters: $\beta_1, \beta_2 ..., \beta_p$
  3. form of model: *predictors, transformation*
- Imputation deals with 1 & 2

Result of 1st Imputation

```
Logistic regression                               Number of obs   =       9367
                                                  LR chi2(12)     =     168.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -1633.2037                       Pseudo R2       =     0.0491

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.000987   .0065701     0.15   0.881     .9881921    1.013947
     age_don |   .9979897   .0039913    -0.50   0.615     .9901974    1.005843
         age |   .9584817   .0090968    -4.47   0.000     .9408171    .9764779
        year |   .8574911   .0130516   -10.10   0.000     .8322883    .8834571
      prevtx |   1.109361     .14694     0.78   0.433     .8557115    1.438197
      txtype |   1.248548   .2388952     1.16   0.246     .8580995    1.816656
             |
      hlamat |
          1  |   1.004009    .199318     0.02   0.984     .6803856    1.481562
          2  |    .818982   .1662035    -0.98   0.325     .5502147    1.219036
          3  |   .6860487   .1422434    -1.82   0.069     .4569506    1.030008
          4  |   .6296867   .1537688    -1.89   0.058     .3901773    1.016218
          5  |   .8229399   .2596461    -0.62   0.537     .4434097    1.527323
          6  |   .8010403   .2694053    -0.66   0.509      .414361    1.548566
```

Imputations

Imputation #	OR txtype	SE(OR)
1	1.25	0.24
2	1.13	0.22
3	1.21	0.23
4	1.29	0.25
5	1.24	0.24
mean	1.22	0.24
SD	0.06	

Combining Imputations

K = Number of Imputations = 5

MI Estimate = $(OR_1 + OR_2 + \ldots + OR_K)/K$

$= 1.22$

MI Variance = $(1 + 1/K)^*\{ SE(OR_k) \}^2 +$

$\{ \text{mean}(\{SE_k(OR)\}^2) \}$

$= (1+1/5)^*\{ 0.06 \}^2 +$

$(0.24)^2$

$= .062$

MI SE = sqrt(MI Variance) = 0.25

Imputed Regression

the analysis step

- mi estimate, or : logistic death cold_isc txty i.hlam age_don age year
- Fits a logistic regression with imputations based on the last commands
- After ":" usual commands
- Need "or" to get odds ratios

Results

```
mi estimate, or: logistic death cold_isc age_don age year prevtx txtype i.hlamat
```

Multiple-imputation estimates — Imputations = 5
Logistic regression — Number of obs = 9367

death	exp(b)	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cold_isc	1.002313	.0071502	0.32	0.746	.9883722	1.01645
age_don	.9979795	.0039896	-0.51	0.613	.9901907	1.00583
age	.9584372	.0090968	-4.47	0.000	.9407726	.9764335
year	.8578479	.0130779	-10.06	0.000	.8325948	.8838669
prevtx	1.108886	.1468931	0.78	0.435	.855321	1.437622
txtype	1.216516	.2422243	0.98	0.325	.8233093	1.797515
hlamat						
1	1.002488	.1990982	0.01	0.990	.679246	1.479555
2	.8181683	.1660688	-0.99	0.323	.549628	1.217914
3	.6849142	.1420574	-1.82	0.068	.4561307	1.02845
4	.6284968	.1535163	-1.90	0.057	.3893938	1.014419
5	.8206398	.2589599	-0.63	0.531	.4421287	1.523199
6	.7986734	.26867	-0.67	0.504	.4130751	1.544221
_cons	6.3e+131	1.9e+133	9.98	0.000	8.0e+105	4.9e+157

PMM Results

```
 mi estimate, or: logistic death cold_isc age_don age year prevtx txtype i.hlamat

Multiple-imputation estimates                     Imputations     =          5
Logistic regression                               Number of obs   =       9367

------------------------------------------------------------------------------
       death |     exp(b)   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.002321   .0070415     0.33   0.741     .9886026     1.01623
     age_don |   .9979781   .0039895    -0.51   0.613     .9901895    1.005828
         age |   .9584332   .0090967    -4.47   0.000     .9407689    .9764293
        year |   .8577968   .0130501   -10.08   0.000     .8325966    .8837597
      prevtx |   1.108817   .1468895     0.78   0.436     .8552597    1.437547
      txtype |   1.216574   .2400089     0.99   0.320     .8263761    1.791016
             |
      hlamat |
          1  |   1.002359   .1990838     0.01   0.991     .6791438    1.479398
          2  |   .8180324   .1660531    -0.99   0.322     .5495211    1.217746
          3  |   .6847243   .1420257    -1.83   0.068     .4559942    1.028187
          4  |   .6284921   .1535049    -1.90   0.057     .3894033    1.014379
          5  |   .8208394   .2589845    -0.63   0.531     .4422767    1.523429
          6  |    .798598   .2686421    -0.67   0.504     .4130386    1.544066
             |
       _cons |   7.1e+131   2.1e+133    10.00   0.000     1.0e+106    4.9e+157
------------------------------------------------------------------------------
```

Data Management

- Imputing creates phantom observations
- Data is only good for imputations
- Useful command -- "mi extract" creates a dataset based on one "imputation"
- mi extract 0, clear
  back to regular incomplete data
- mi extract 10, clear
  complete dataset using imputation #10
  allows you to check imputation set

Imputation 10

```
.  logistic death cold_isc age_don age year prevtx txtype i.hlamat

Logistic regression                               Number of obs   =       9367
                                                  LR chi2(12)     =     168.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -1633.2127                       Pseudo R2       =     0.0491

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.000447   .0065814     0.07   0.946     .9876305     1.01343
     age_don |    .997997   .0039921    -0.50   0.616     .9902032    1.005852
         age |   .9584903   .0090981    -4.47   0.000     .9408232    .9764891
        year |   .8573443   .0130516   -10.11   0.000     .8321415    .8833103
      prevtx |   1.109332   .1469424     0.78   0.433      .855679    1.438175
      txtype |   1.261713   .2416046     1.21   0.225     .8668913    1.836355
             |
      hlamat |
          1  |   1.004259   .1993966     0.02   0.983     .6805166    1.482015
          2  |   .8191515   .1662461    -0.98   0.326     .5503179    1.219312
          3  |   .6863007   .1423103    -1.82   0.069     .4570993     1.03043
          4  |    .630004   .1538591    -1.89   0.059     .3903584    1.016771
          5  |   .8238943    .259907    -0.61   0.539     .4439664    1.528949
          6  |   .8018133   .2697156    -0.66   0.511     .4147099    1.550251
             |
       _cons |   2.0e+132   6.1e+133    10.03   0.000     2.8e+106    1.4e+158
------------------------------------------------------------------------------
```

but this dropped all other imputations

mi xeq

simpler approach

- mi xeq 10: summ cold_isc
  summarize cold_isc from imputation 10

```
m=10 data:
-> summ cold_isc

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    cold_isc |       9367    10.04335    11.53007  -20.96038         72
```

mi xeq

examining results

```
. mi xeq 10: logistic death cold_isc age_don age year prevtx txtype i.hlamat

m=10 data:
-> logistic death cold_isc age_don age year prevtx txtype i.hlamat

Logistic regression                               Number of obs   =       9367
                                                  LR chi2(12)     =     168.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -1633.2127                       Pseudo R2       =     0.0491

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.000447   .0065814     0.07   0.946     .9876305     1.01343
     age_don |    .997997   .0039921    -0.50   0.616     .9902032    1.005852
         age |   .9584903   .0090981    -4.47   0.000     .9408232    .9764891
        year |   .8573443   .0130516   -10.11   0.000     .8321415    .8833103
      prevtx |   1.109332   .1469424     0.78   0.433      .855679    1.438175
      txtype |   1.261713   .2416046     1.21   0.225     .8668913    1.836355
             |
      hlamat |
          1  |   1.004259   .1993966     0.02   0.983     .6805166    1.482015
          2  |   .8191515   .1662461    -0.98   0.326     .5503179    1.219312
          3  |   .6863007   .1423103    -1.82   0.069     .4570993     1.03043
          4  |    .630004   .1538591    -1.89   0.059     .3903584    1.016771
          5  |   .8238943    .259907    -0.61   0.539     .4439664    1.528949
          6  |   .8018133   .2697156    -0.66   0.511     .4147099    1.550251
             |
       _cons |   2.0e+132   6.1e+133    10.03   0.000     2.8e+106    1.4e+158
------------------------------------------------------------------------------
```

mi extract

- Useful command
- Analysis not supported in mi estimate:
- Use “mi extract, `k’” looped for k
- Save results
- Analyze using published formulae
  not optimal, but possible

How Many Imputations

- Feasibly many:
  "50% missingness 5 draws is highly efficient"
- True but gives poor standard errors
  better guideline 20-40
- Stata on UNOS data
  - 20 imputations in 6 sec.
  - 50 imputations in 16 sec.
  - 500 imputations in 360 sec.

50 Imputations

```
. mi estimate, or: logistic death cold_isc age_don age year prevtx txtype i.hlamat

Multiple-imputation estimates                     Imputations     =         50
Logistic regression                               Number of obs   =       9367
                                                  Average RVI     =     0.0126
DF adjustment:   Large sample                     DF:     min     =    2492.02
                                                          avg     =   2.42e+10
                                                          max     =   2.69e+11
Model F test:       Equal FMI                     F(  12, 3.2e+06)=      13.04
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
       death |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.001093   .0071376     0.15   0.878     .9871925    1.015189
     age_don |   .9979925    .003991    -0.50   0.615     .9902007    1.005846
         age |   .9584724   .0090979    -4.47   0.000     .9408057    .9764708
        year |   .8575193   .0130689   -10.09   0.000     .8322835    .8835204
      prevtx |   1.109199     .14693     0.78   0.434     .8555691    1.438017
      txtype |   1.245769   .2475528     1.11   0.269     .8438537    1.839111
             |
      hlamat |
          1  |   1.003516   .1992915     0.02   0.986     .6799574    1.481042
          2  |   .8187808   .1661875    -0.99   0.325     .5500469    1.218809
          3  |   .6857152   .1422276    -1.82   0.069     .4566588    1.029664
          4  |   .6294577    .153747    -1.90   0.058      .389994    1.015957
          5  |    .822776   .2596321    -0.62   0.536     .4432816    1.527156
          6  |   .8006779   .2693564    -0.66   0.509     .4140995    1.548143
------------------------------------------------------------------------------
```

500 Imputations

```
Multiple-imputation estimates                     Imputations     =        500
Logistic regression                               Number of obs   =       9367
                                                  Average RVI     =     0.0130
DF adjustment:   Large sample                     DF:     min     =   23812.19
                                                          avg     =   1.54e+11
                                                          max     =   1.54e+12
Model F test:       Equal FMI                     F(  12, 3.1e+07)=      13.02
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
       death |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.002006   .0070717     0.28   0.776     .9882409    1.015964
     age_don |   .9979834   .0039899    -0.50   0.614     .9901938    1.005834
         age |   .9584474    .009097    -4.47   0.000     .9407824    .9764441
        year |   .8577653   .0130703   -10.07   0.000     .8325267    .8837691
      prevtx |   1.109026   .1469091     0.78   0.435     .8554327    1.437797
      txtype |   1.223848   .2423487     1.02   0.308     .8301736    1.804205
             |
      hlamat |
          1  |   1.002682   .1991348     0.01   0.989     .6793796    1.479836
          2  |   .8183211   .1660985    -0.99   0.323     .5497324    1.218137
          3  |   .6850761   .1420932    -1.82   0.068     .4562356    1.028699
          4  |   .6287796   .1535788    -1.90   0.057     .3895769    1.014854
          5  |   .8212404    .259145    -0.62   0.533     .4424569    1.524297
          6  |   .7992063   .2688505    -0.67   0.505     .4133495    1.545256
------------------------------------------------------------------------------
```

Approach to Imputation Model

- Inclusivity over parsimony
  more predictor: MAR more plausible
- Will remove biases -- primary objective
- Two things to include: outcome, predictor in your analytic model
- Functional form tends not to be critical

Making Up Data?

- No.
- Making a model for missingness
- Making guess at missing data
- Fully representing uncertainty
  - this is the key to it's validity

Inverse Weighting for Missing Data

UNOS Dataset

- Data on 9775 pediatric kidney transplants from 1990-2002
- Followed for transplant outcomes
  465 deaths
- Predictors: age, age of donor (age_don), year, # HLA loci (hlamat), living v. cadaveric donor (txtype) and cold ischemia time
- Cold ischemia time: missing in 23%
- Ignore missingness in other variables

MAR Assumption

- Cold Ischemia values MAR with respect to some covariates and outcome (C_{MAR}) *same assumption as in multiple imputation*
- R: (1=cold ischemia observed, 0=missing)
- Formally $pr(R=1|X_{age,..}X_{cold\_isc,\ldots,}X_{hlamat}, Y) = pr(R=1|\ X_{age,\ldots.,}X_{hlamat}, Y)$
- Probability of being observed doesn't dependent on X_{cold_isc}
- Depends on C_{MAR} $(X_{age},\ldots,X_{hlamat}, Y)$ *like a "stratified sample*

Implication #2

- Inverse weighting uses "Implication #2"
  $pr(R=1|X_{age,..}X_{cold\_isc,\ldots,}X_{hlamat},Y)$ =
  $pr(R=1|\ X_{age,\ldots.,}X_{hlamat},Y)$
  doesn't depend on X_{cold_isc}
- Denote $C_{MAR} = \{X_{age,\ldots.,}X_{hlamat},Y\}$
  the set of MAR variables
- $Pr(R_i=1|\ C_{MAR}) = \pi_i$
- Behave like a probability sample
  with unknown probability of selection
  But it is estimable

Steps :Inverse Weighting

1. Assume Data is Missing at Random (MAR)
2. Specify Model for Complete Data
3. Estimate/Calculate Weights
4. Analyze Available Data with Weights

MAR Assumption

Step 1

- Cold Ischemia values MAR with respect to other model covariates and outcome
  same assumption as in multiple imputation
- Denote $C_{MAR} = \{X_{age,\ldots,}X_{hlamat}, Y\}$
  as the set of MAR variables
- Want to include Y
- Should be inclusive: more variables
  similar consideration to imputation
- But "Congeniality" not a consideration

UNOS Analysis Model

Step 2

Model for the complete data

- Y: survival after transplant
- X: age, age_don, year, hlamat, txtype, cold_isc
- $\text{logit}(Y) = \beta_0 + \beta_{age}*\text{age} + \beta_{age\_don}*\text{age\_don} + \beta_{year}*\text{year} + \beta_{txtype}*\text{txtype} + \beta_{cold\_isc}*\text{cold\_isc} + \beta_{hlamat}*\text{hlamat}$
- Just a logistic regression

Model for Weights

Step 3

Model for probability of observed data

- Probability being observed not known *must be estimated. possibly through a model*
- $pr(R_i=1 \mid C_{MAR}) = \pi(C_{MAR})$
- Can specify as logistic regression $\mathrm{logit}\{\pi(C_{MAR})\} = \beta^T C_{MAR}$
- Could specify any model

Modeling Choices

Step 3

- MAR assumption dictates C_{MAR} variables
- But not the form of $pr(R_i=1 \mid C_{MAR}) = \pi(C_{MAR})$
- Include transformation, spline, interactions?
  lab last week needed an interaction
- Congeniality w/ analysis model not needed
  interaction in one doesn't require it the other
- Validity requires $\pi(C_{MAR})$ model is correct
  suggests non-parsimonious modeling

Calculate $\pi(C_{MAR})$

Step 3

- Estimable for anyone with C_{MAR} available
- Key to the unbiased analysis
- weight: $w_i = 1/\pi_i$
  - logit R i.hlamat age_don age death txtype year
  - predict pi, p
  - gen w = 1/p

Sampling Weights

Step 3

- Attempts to create the full sample
- "Re-weight" Data| R=1 to mimic it
- Large weight => increased importance
- When π_i small: person under-represented *leads to a large weight*
- When π_i =1, full represented, w_i = 1 *given no additional weight in analysis*

Equivalent Models

- Complete Data:
  $\text{logit}(Y) = \beta_0+\beta_{age}*\text{age} + \beta_{age\_don}*\text{age\_don} + \beta_{year}*\text{year} + \beta_{txtype}*\text{txtype} + \beta_{cold\_isc}*\text{cold\_isc} + \beta_{hlamat}*\text{hlamat}$

- Data with R=1
  $\text{logit}(Y) = \beta_0+\beta_{age}*\text{age} + \beta_{age\_don}*\text{age\_don} + \beta_{year}*\text{year} + \beta_{txtype}*\text{txtype} + \beta_{cold\_isc}*\text{cold\_isc} + \beta_{hlamat}*\text{hlamat}$
  weighted by w_i

Analysis Commands in Stata

- logit R i.hlamat age_don age death txtype year
- predict pi, p
- gen w = 1/pi
- logit death i.hlamat age_don age cold txtype year [pweight=w]

Stata Results

```
Logistic regression                             Number of obs     =      7,347
                                                Wald chi2(11)     =     131.68
                                                Prob > chi2       =     0.0000
Log pseudolikelihood =  -1717.388               Pseudo R2         =     0.0436

------------------------------------------------------------------------------
             |               Robust
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hlamat |
          1  |   .0962106   .2291147     0.42   0.675    -.3528461    .5452672
          2  |   .0084009   .2333768     0.04   0.971    -.4490093     .465811
          3  |  -.1779645   .2334702    -0.76   0.446    -.6355577    .2796286
          4  |   -.216194    .275066    -0.79   0.432    -.7553134    .3229255
          5  |  -.0910452   .3676077    -0.25   0.804     -.811543    .6294525
          6  |  -.1018275   .3715002    -0.27   0.784    -.8299546    .6262995
             |
     age_don |  -.0023699   .0044193    -0.54   0.592    -.0110315    .0062916
         age |  -.0329875   .0111418    -2.96   0.003    -.0548251     -.01115
    cold_isc |   .0007704    .006628     0.12   0.907    -.0122202    .0137609
      txtype |   .2828537   .1976179     1.43   0.152    -.1044703    .6701777
        year |  -.1467212   .0152131    -9.64   0.000    -.1765384    -.116904
       _cons |   290.0792   30.36358     9.55   0.000     230.5677    349.5908
------------------------------------------------------------------------------
```

Advice: Inverse Weights

- Use non-parsimonious model for R
- Use external information
  to guide the MAR assumption/model
- Use when only a single missing variables
  avoid if missing data in >1 variable/predictor
- Examine the probability weights
  large weights yield big influence
- Analysis so similar to complete data
  just with weights
- Can be inefficient

Maximum Likelihood (ML)

Based on implication #1

- Maximum likelihood valid for MAR data
  as long as CMAR variables incorporated
- Works especially well for missing outcome
  especially in longitudinal analysis
  where attrition depends on previous values
- ML often fit with the EM algorithm
  EM also iterative fills in missing data
- Imputation & analysis models consistent
  MI and maximum like very similar

Disadvantages of ML

- Cumbersome for missing predictors
  don’t usual model their distribution
- Awkward if CMAR variables are not a part of desired
- Lacks flexibility of MI

3 Methods

- Same MAR assumption
- Inverse weight:
  model missingness, data model flexible
- MI/ML:
  model data, missingness model unspecified