Multiple Imputation and other MAR Approaches

Dave Glidden

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

How Many Imputations

- Feasibly many:
  "50% missingness 5 draws is highly efficient"
- True but gives poor standard errors
  better guideline 20-40
- Chained Imputation on UNOS data
  - 20 imputations in 2 min.
  - 50 imputations in 5 min.
  - 500 imputations in 50 min.

50 Imputations

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

Multiple-imputation estimates                     Imputations     =         50
Logistic regression                               Number of obs   =       9775
                                                  Average RVI     =     0.0371
                                                  Largest FMI     =     0.1946
DF adjustment:   Large sample                     DF:     min     =    1310.53
                                                          avg     =  894264.52
                                                          max     = 5923449.56
Model F test:       Equal FMI                     F(  11,351411.5)=       9.02
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
       death |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.009856   .0066926     1.48   0.139     .9968112    1.023071
     age_don |   .9905272   .0038156    -2.47   0.013      .983077    .9980339
         age |   .9482437   .0084732    -5.95   0.000     .9317811    .9649971
      prevtx |   1.376873   .1752271     2.51   0.012     1.072907    1.766954
      txtype |   1.237584   .2299089     1.15   0.251     .8598254    1.781308
             |
      hlamat |
           1 |   1.048506   .1886219     0.26   0.792     .7369546    1.491768
           2 |   .8519965   .1592486    -0.86   0.391     .5906521    1.228977
           3 |   .8045399   .1525117    -1.15   0.251     .5548634    1.166565
           4 |   .6962644   .1608806    -1.57   0.117     .4426748    1.095125
           5 |   .8605673   .2608053    -0.50   0.620     .4751321    1.558674
           6 |   .9041763   .2940166    -0.31   0.757     .4780357    1.710196
             |
```

500 Imputations

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

Multiple-imputation estimates                     Imputations     =        500
Logistic regression                               Number of obs   =       9775
                                                  Average RVI     =     0.0381
                                                  Largest FMI     =     0.1853
DF adjustment:   Large sample                     DF:     min     =   14556.70
                                                          avg     =   1.19e+07
                                                          max     =   7.27e+07
Model F test:       Equal FMI                     F(  11, 3.4e+06)=       9.02
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
       death |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cold_isc |   1.010071   .0066465     1.52   0.128     .9971269    1.023184
     age_don |   .9905008   .0038143    -2.48   0.013     .9830531    .9980049
         age |   .9482384   .0084729    -5.95   0.000     .9317765    .9649913
      prevtx |   1.373944   .1765813     2.47   0.013     1.067999    1.767533
      txtype |     1.2328   .2287713     1.13   0.259      .856902    1.773593
             |
      hlamat |
           1 |   1.049783   .1892282     0.27   0.788     .7373369    1.494627
           2 |   .8522425   .1586011    -0.86   0.390     .5917734    1.227357
           3 |   .8053358   .1526196    -1.14   0.253     .5554756    1.167587
           4 |   .6966721    .159863    -1.58   0.115       .44433    1.092323
           5 |   .8545295   .2592478    -0.52   0.604     .4715075    1.548694
           6 |   .9090107   .2962998    -0.29   0.770     .4798573    1.721971
```

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 a 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
- $\text{logit}(Y) = \beta_0+\beta_{age}*age + \beta_{age\_don}*age\_don + \beta_{year}*year + \beta_{txtype}*txtype + \beta_{cold\_isc}*cold\_isc + \beta_{hlamat}*hlamat$
- 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
  $logit\{\pi(C_{MAR})\} = \beta^T C_{MAR}$
- 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 $\pi_i = 1$, full represented, $w_i = 1$
  given no additional weight in analysis

Equivalent Models

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

Counterfactuals and Potential Outcomes Estimation

Dave Glidden
Biostatistics 210

BMD and Fracture Risk

- Cohort of older adults
- Have bone mineral density measured
- Followed for onset of fracture risk
- Risk of fracture associated level with BMD=x

Causal Question

- Many thing contribute to outcomes
- Want to manipulate the exposure
- How would outcome change?
- Observed differences in exposure don't approximate change in exposure

difference between "see" and "do"

Some Questions

- Kidney transplant
  better survival with living organ donation?
- Exercise benefit
  regular exercise control glucose in older women?
- Objective: causal inference or as close as well can come

What is a confounder?

What I learned….

A variable associated with the outcome and the exposure. One that leads to bias if it is not "controlled" for.

Not a rigorous definition

Now we use DAGs to identify adjustment sets

New Approaches

- Rigorous definitions
- Causal Inference
- Causal Estimands (targets)
- How to estimate targets
  which variables need to be controlled for?
- Interesting interplay with missing data

Some Notation

- $Y_{i:}$: observed outcome for the ith person
- E_i : observed exposure for the ith person
  Exercise: binary, SOF: continuous
- $C_i=(C_{1i}\ldots.C_{pi})$ predictors associated Y_i, E_i

Potential Outcomes

- $Y_i(1)$: Glucose ith person who exercises
- $Y_i(0)$: Glucose ith person no exercise
- $Y_i(1)$-$Y_i(0)$: causal effect in ith person
  would be rigorous to observe this!
- Only get to see one realization
  even in a randomized trial
  possible exception: crossover trials

Average Causal Effect

- Effect of exposure in population
- A.C.E. $E[Y(1)] - E[Y(0)]$
  $= 1/n \sum Y_i(1) - Y_i(0)$
- Average causal effect
- Very elegant "target"
- Why is it causal?
  compares exposure in all $i=1,\ldots,n$
  the identical population

Table 9.1. Potential Outcomes Framework for the Effect of Exercise

Sub-Population	Person	$Y(1)$	$Y(0)$	$Y(1)$-$Y(0)$
C = 0	1	94	99	-5
	2	98	99	-1
	3	99	98	1
	4	103	105	-2
	5	96	99	-3
C = 1	6	95	94	1
	7	92	95	-3
	8	94	95	-1
	9	90	94	-5
	10	99	101	-2
Overall	Mean	96	98	-2

Fundamental Problem

- Only get to see actual exposure
  not potential outcome
- E_i: exposure in the ith person
- E_i: $_i$: associated w/ factor(s) C related to outcome
- E_i: -- confounder. Associated with exposure and outcome
- Distorts the observed difference in exposure groups

Table 9.2. Effect of Exercise Using Observed Data Only

Sub-Population	Person	$Y(1)$	$Y(0)$	Estimated Difference
$C = 0$	1		99	
	2		99	
	3		98	
	4		105	
	5	96		
$C = 1$	6	95		
	7		95	
	8	94		
	9	90		
	10	99		
Overall	Mean	94.8	99.2	-4.4

C=1: family history of diabetes, C=0: no family history of diabetes

Missing Data Problem

- What is mean of Y(1)?
- Missing in some people
  they were exposed to $E_i=0$
- Missingness associated with C
- C is associated with latent Y(1)
- Classic missing data problem

"Fundamental problem" is a missing data problem

Randomization Assumption

- E_i is independent of Y(1) and Y(0)
- If true can identify E[Y(1)] and E[Y(0)] *estimate them rigorously*
- Observed difference identifies causal effect
- Causal inference: without having counterfactuals

Why we love the RCT

Potential outcome is "missing completely at random"

No Confounders

- E_i is independent of Y(1) and Y(0)
- If true can identify E[Y(1)] and E[Y(0)] *estimate them rigorously*
- Observed difference identifies causal effect
- Causal inference: without having counterfactuals

Why we love the RCT

Potential outcome is "missing completely at random"

Conditional Independence

- Relaxes this assumption
- If C is a confounder...
  - suppose within levels of C, E and Y(1),Y(0) are not associated
  - exposure random within levels of the confounder
- Adjustment identifies effect within C
- Assumption behind "adjustment" for confounders

Conditional Independence

- Within levels of family hx diabetes
  - unconfounded effect of exercise
  - within levels exercise not associated with lipids if exercise or no exercise
- Then "see" reveals "do"
- Principle extends to complex confounders *continuous or multiple confounders*

Conditional Independence

- Core assumption of causal inference
- Synonymous with "no unmeasured confounders" beyond C
- Estimate effect of exposure with C
  and combine in some sense
- Basis of several methods
  regression, inverse weighting, propensity scores, potential outcomes estimation