Missing Values

VGSM, Chapter 11

Dave Glidden

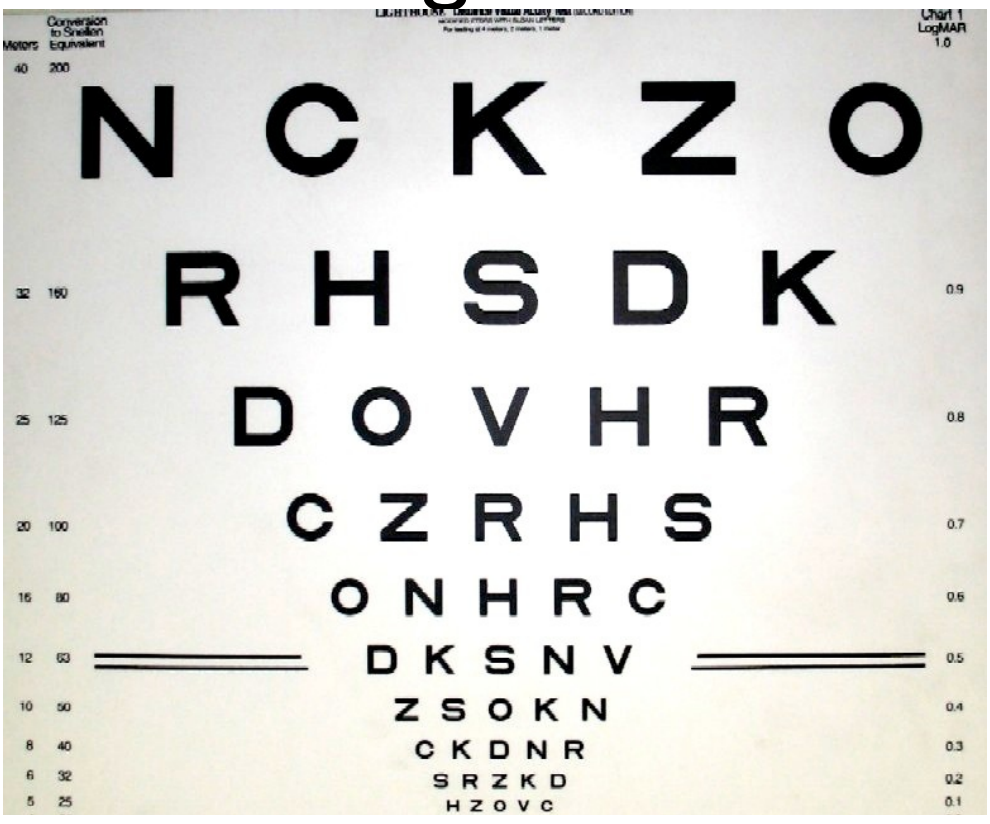
Outline

- Missing data problem
- Sampling and missing data
- Taxonomy of missing data
  MAR assumption
- Multiple Imputation

Example: Eye Study

- Eyes infection recovery
- Repeated measures: visual acuity
  0, 3, 12, and 52 weeks post-treatment
- What is recovery trajectory?
- Covariate effects would require interaction

logMAR



Available Data

```
xtset id vaspoint
xtdescribe
```

Freq.	Percent	Cum.	Pattern
392	78.40	78.40	1111
58	11.60	90.00	111.
19	3.80	93.80	11..
15	3.00	96.80	1...
9	1.80	98.60	11.1
4	0.80	99.40	1.11
3	0.60	100.00	1.1.
500	100.00		XXXX

Mean logMAR

smaller logMAR => better vision

Complete Data

vaspoint	mean(logmar)
Baseline	0.909
Week 3	0.590
Month 3	0.442
Month 12	0.400

Available Data

vaspoint	mean(logmar)
Baseline	0.909
Week 3	0.417
Month 3	0.295
Month 12	0.153

```
table vaspoint, contents(mean logmar) format(%5.3f)
```

Basic Analysis

- Calculating means at each timepoints in the setting of missing data
- Gives biased estimate of the mean
- Why? Data is no longer representative
- Attrition of people who have poorer FU vision
- Data that's missing different from data from data observed

Non-Representative

- Available data misleading
- Correlation between missingness and the value of the missing data
- How this issue arises is important
- Leads to the classifications of missing data

Notation

$(Y_{i0}, \ldots, Y_{i3})$ log MAR at the follow up times

R_{i2} $R_{i2} = 1$ data available at 2nd follow-up

$R_{i2} = 0$ data missing at 2nd follow-up

Sampling and Missing Data

- Close analogy
- Missing data: incompletely sampled
- Missing data can be handled
- Even if it is not "completely random"
- But need to think about how it arises
  leads to important classifications

Eye Data: MCAR

- Missing values: transportation/distance
- No association between transport/distance and severity at baseline or recovery
- Acuity independent of missingness
- Can treat data as simple random sample
- Missing completely at random

$R_{i2} \perp Y_{i2}$ *value of data is not associated with whether we observe it*

Consequence...

- Visual acuity at each visit same if data missing or measured
- Data behaves like a random sample
- Can ignore missingness in analysis
- Missing data: no bias, only loss of efficiency
- Simple analysis is possible

Eye Data: NMAR

- Data behaves like a completely biased sample
- People who have good vision come back
- $R_{i2} = 1 \text{ if } Y_{i2} - Y_{i0} > c$
- Not predictable from current data
- Get a biased sample from data

Not missing at random

Eye Data: MAR

- Distance associated with severity and missingness at month 3
- Severity effect goes through baseline
- Baseline is always observed

$$R_{i2} \perp Y_{i2} | Y_{i0}$$

This implies

- Distance and severity unobserved
- Severity effect on logMAR mediated by baseline
- Month 3 value/Missingness associated
- Association broken by baseline logMAR
  Missing values behave like sample
  controlling for baseline acuity

So-called missing at random

Missing at Random

- Missing data behaves like over-sampled data
- A conditional independence assumption
- Data and missingness associated through some variables: C_{MAR}
- C_{MAR} needs to be available in all
  can relax this later
- Much like assumptions about confounding
- Many analogies with confounding

Taxonomy of Missing Data

- MCAR: Data and Missing Independent
  Ignoring missingness is unbiased
  but loose efficiency
- MAR: Data and Missing Associated
  Sensitivity analyses are all that's possible
- MAR: Data and Missing Associated
  But independent given C_{MAR}
  Some analyses biased
  Principled Analyses are not

Continuum of Assumptions

- Can't verify if data is MAR, MCAR, NMAR
- Unless you recover some missing data
- Requires unverifiable assumptions
- MCAR: likely unrealistic
- NMAR: limited options
- MAR is where modeling is possible

MAR Modeling

- MAR is wrt a series of variables (C_{MAR})
  those must be 100% measured or MCAR
- Generally want to choose more rather than fewer: to make sure bias is eliminated
- Modeling requires insight into missing data
- Always a level of unverifiable assumption

Gotta be considered better than assuming MCAR

MAR v. NMAR

- Superficially similar: worse people overrepresented
- Difference between oversampling and biased "sample"
- MAR: oversampling from worse vision / organism strata
- NMAR: not like a sample

Formally

$(Y_{i0}, \ldots, Y_{i3})$ log MAR at the follow up times

R_{i2} $R_{i2} = 1$ data available at 2nd follow-up

$R_{i2} = 0$ data missing at 2nd follow-up

$R_{i2} \perp Y_{i2} | Y_{i0}$

$f(Y_{i2}|Y_{i0}, R_{i2} = 1) = f(Y_{i2}|Y_{i0})$ Implication #1

$\text{pr}(R_{i2} = 1|Y_{i2}, Y_{i0}) = \text{pr}(R_{i2} = 1|Y_{i0})$ Implication #2

Powerful Implications

Implication #1

- Directly appeals to sampling
- The observed data is a stratified sample of the potential data
- Can we analyzed through sampling weights
- Weights estimated from the data

Powerful Implications

Implication #2

- First: Association between Y_2, Y_0 from observed data is unbiased
- Association in the unobserved
- Can infer values of Y_2 for people who have it missing by their Y_0 values
- Drives 2 methods:
  Maximum Likelihood & Multiple Imputation

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%
  suppose C_{MAR} = txtype and HLA loci

Cold Ischemia Time

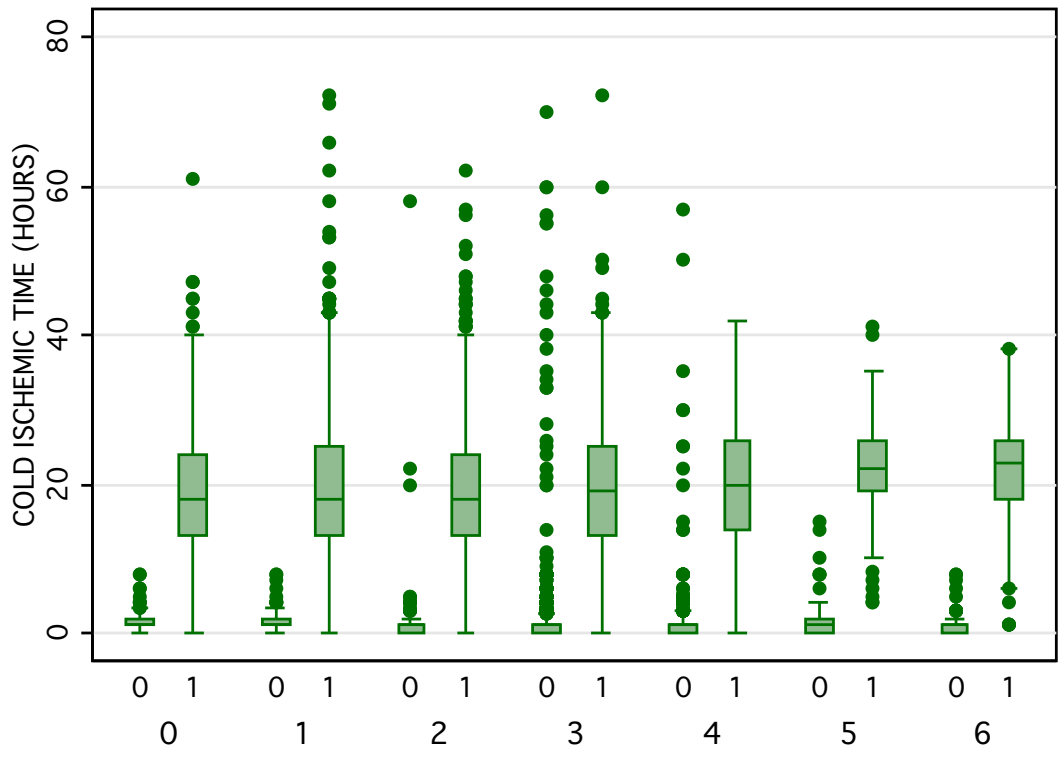
Living Donor

```
summ cold_isc if txt==0, detail

                COLD ISCHEMIC TIME (HOURS)
-------------------------------------------------------------
      Percentiles      Smallest
 1%             0             0
 5%             0             0
10%             0             0       Obs                3658
25%             0             0       Sum of Wgt.        3658

50%             1                     Mean           1.567272
                      Largest         Std. Dev.      4.359038
75%             1            60
90%             3            60       Variance       19.00122
95%             4            64       Skewness       9.501109
99%            20            70       Kurtosis       110.3521
```

HLA and Donor Source



Essence of Assumption

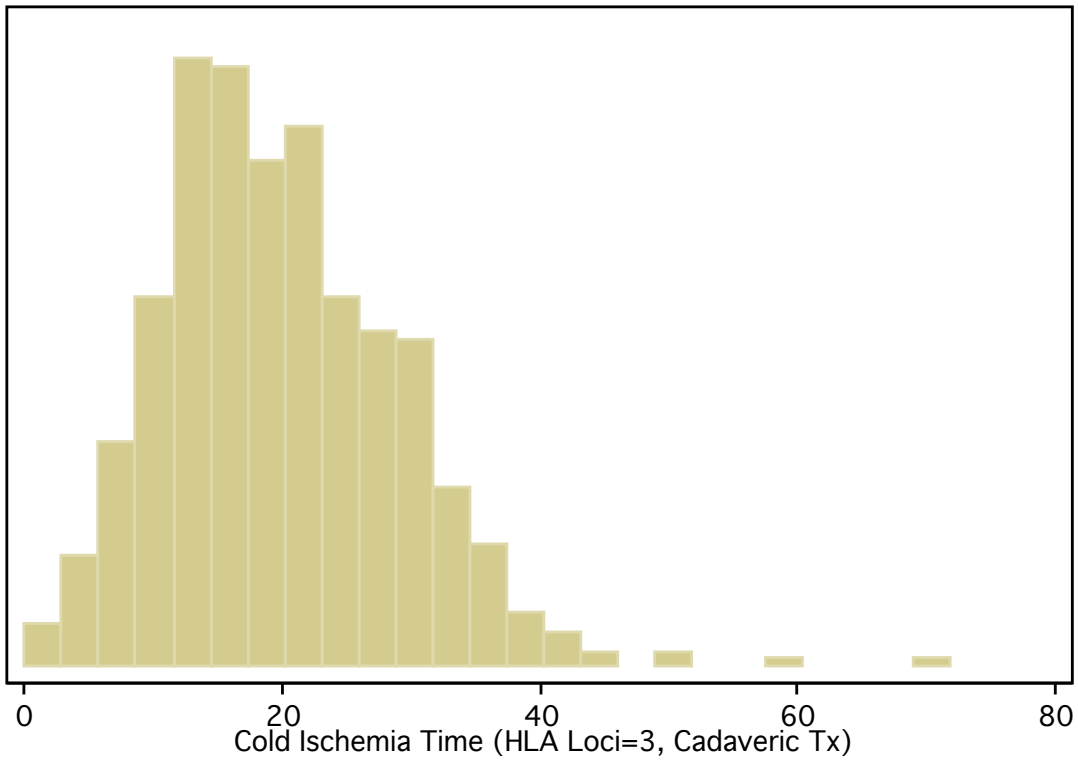
Person	HLA Loci	Txtype	Cold_Isc
1	3	Cadaveric	15h
2	3	Cadaveric	?

This is equivalent to saying that the distn of cold_isc among people who share *observed values* on HLA and txtype have the same distribution whether observed or not.

Uncertainty

- Don't know the cold_ischemia times
- But if willing to make the assumption....
- We have some reasonable guesses
  based on txtype and # HLA loci
- Fill missing with "mean" from regression
  not good -- treats values as known
- Represent uncertainty with prob distn

3 HLA, Cadaveric Tx



Multiple Imputation

- Don't know the cold_ischemia times
- Using MAR and implication #1
- We have some reasonable guesses
  based on txtype and # HLA loci
- Represent uncertainty with prob distn

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

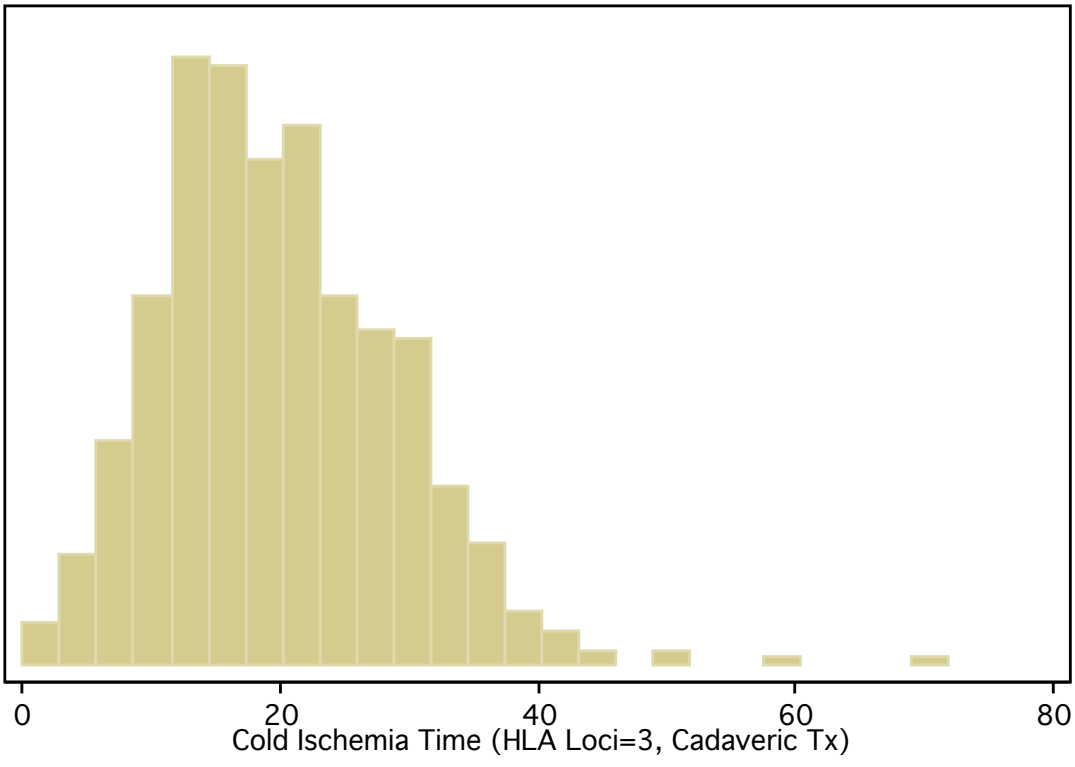
1. MAR Assumption

MAR Assumption

step 1

- Assume that cold_ischemia is MAR
- That values MAR conditional on HLA and txtype
- Txtype and # HLA loci can inform
- Predictive model might simply take cold_isc from dist of same loci and txtype
  so-called *"hot deck imputation"*

3 HLA, Cadaveric Tx



Selecting MAR Variables

- Bias is the biggest threat
  - More predictors make MAR plausible
  - Include the outcome
- Unimportant covariates: little cost
  - probably not much gained
- Any predictor in model -- include impute
  - possible some that aren't

MAR Assumption

- Assume that cold_ischemia is MAR
- That values MAR conditional on available data: txtype, HLA loci, age of recipient, age of donor, year of transplant, and survival
- Suddenly not easy to find matches
- Use a regression model as the basis for predictions

2. Develop predictive model for missing data

Congeniality

- Two modeling steps
  imputation model, analysis model
- Imputation model needs to be consistent with analysis model
- Imputation more general than analysis or parameters estimates inconsistent
- More flexible
  small efficiency costs
- This is termed "congeniality"

Variable for Imputations

Include more than just "MAR variables"

- Any variable in analysis model should be used to generate imputations
- Auxiliary: predictive variables can increase efficiency
- Congeniality of variables
  - interaction in model => also imputation
  - transform in model => transform imputations

Include the Outcome

when imputing a variable which is a predictor in the analysis model

- Yes
  Seems awkward even circular
- MAR: no distinction between pred/ outcomes
- Complete case analysis:
  unbiased if MAR wrt predictors only
  adding outcome extends MAR assumptions
- Makes full use of MI capabilities

Modeling Choice

- Specify a parametric model for imputations
- cold_ischemia: continuous variable
  but also has many "0" s
- Many missing values like to be "0"
- Difficult distribution to model
- Try Normal distribution

Predictive Model

```
. . xi: reg cold_isc txty death i.hla age_don age year
i.hlamat          _Ihlamat_0-6        (naturally coded; _Ihlamat_0 omitted)

      Source |       SS       df       MS              Number of obs =    7347
-------------+------------------------------           F( 11,  7335) = 1126.79
       Model |  611905.778    11   55627.798           Prob > F      =  0.0000
    Residual |  362116.493  7335  49.3683017           R-squared     =  0.6282
-------------+------------------------------           Adj R-squared =  0.6277
       Total |  974022.271  7346  132.592196           Root MSE      =  7.0263

------------------------------------------------------------------------------
    cold_isc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |   18.49092   .2235349    82.72   0.000     18.05273    18.92911
       death |    .243987   .3871391     0.63   0.529    -.5149169    1.002891
  _Ihlamat_1 |   .6966437   .3575722     1.95   0.051    -.0043006    1.397588
  _Ihlamat_2 |   .4241767   .3550638     1.19   0.232    -.2718505    1.120204
  _Ihlamat_3 |   .7843253   .3517806     2.23   0.026     .0947341    1.473916
  _Ihlamat_4 |   .9951523    .394421     2.52   0.012     .2219737    1.768331
  _Ihlamat_5 |   1.636238   .5308462     3.08   0.002     .5956272    2.676849
  _Ihlamat_6 |   1.722591    .560876     3.07   0.002     .6231127    2.822069
     age_don |   .0153432   .0067599     2.27   0.023     .0020918    .0285946
         age |   .0286415   .0161681     1.77   0.077    -.0030526    .0603356
        year |  -.2486191   .0229721   -10.82   0.000    -.2936511   -.2035872
       _cons |   495.8731   45.87233    10.81   0.000     405.9501     585.796
------------------------------------------------------------------------------
```

Summary

- Missing data grouped by relationship between missingness and it's value
- MAR is the assumption where most action is
- Motivates multiple imputation
- Model-based guesses at missing data