

Missing Data, Part I

Steyerberg, Chapter 7

Dave Glidden
10 October 2009

Outline

- Examples
- Inferior methods
- Sampling and missing data
- Taxonomy of missing data
- Modeling missing values

Transplant Data

- Outcome: 1 yr survival post-transplant
- Variables: transplant type, age at transplant, previous transplant, cold ischemia time
- Some of these predictors missing
- Cold ischemia time frequently missing

Data Completeness

Variable	Obs	Mean	Std. Dev.	Min	Max
prevtx	9702	1.1443	.3514119	1	2
txtype	9775	.4733504	.4993148	0	1
age	9766	11.64653	5.291385	0	18
cold_isc	7525	10.85967	11.51735	0	72
hlatat	9541	2.57363	1.378919	0	6

prevtx: 1%
txtype: 0%
age: >1%
hla loci: 3%
cold_isc: 23%

Example: Eye Study

- Eyes infection recovery
- Repeated measures: visual acuity
0, 3, 12, and 52 weeks post-treatment
- Does sex affect recovery trajectory?
- Main question is an interaction

Available Data

vaspoint	Freq.
Baseline	138 100%
Week 3	129 93%
Month 3	126 91%
Month 12	117 85%

81 men, 57 women

xtdescribe

Freq.	Percent	Cum.	Pattern
114	82.61	82.61	1111
10	7.25	89.86	111.
7	5.07	94.93	1...
4	2.90	97.83	11..
2	1.45	99.28	1.11
1	0.72	100.00	11.1
138	100.00		XXXX

Inferior Options

LOCF

- Last observation carried forward
- Fills missing data with last available value
- Biased: ignores time trend
- Ignores variability: single imputation
treats missing values as predicted without error
- Test with size 0.05 in RCT
estimate of treatment effect is biased
- Always do better with multiple imputation

Complete Case

- Use especially in regression
employed by Stata
- Can be biased or inefficient
- In benign case, just inefficient
- If missingness depends on other predictors
not outcome: unbiased but inefficient

Missing Data

- Incredibly broad topic: whole course
- Subject is as broad as sampling
- Missing data closely connected to sampling
- Sampling: general all or no data on people
- Missing data: partial sampled data

Sampling

- Statistics: infer about popn from sample
- US election 130,000,000 votes
- Polls required less than 1% samples
- All possible because of “random sampling”
sample not biased

Hypothetical Study

- Study of HIV risk factors/prevalance in STI clinic
- Requires consent, counseling
- Want to choose representative group

Simple Random Sample

- People completely at random
- Equal chance of being sampled
- Ideal -- rarely ever achieved
who really assembles sampling frame
- Sampling doesn't depend on outcome
- Can analyze data in straightforward way

Oversampling

- Want adequate numbers of men/women
or racial groups
- Can “oversample”
deliberately more likely to approach women
- No longer ‘representative’ of clinic
- Can still be analyzed:
need to take account of sampling

Other Oversampling

- Could sample on sexual behaviour
- Or results of other STD tests
- Sampling is very flexible

Nested Case Control

- Cohort of people at risk for disease (HIV)
- Store and freeze samples at each visit
- Select HIV+ and subset HIV-
thaw & analyze drug levels
- Oversampled HIV+ in cohort
- Drug levels unknown on 80%-90% sample
but can still analyze

Not a Sample

- Taking only people who seek HIV test
- Selects people most at risk?
or does it?
- Not clear what can be learned by this

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

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
- Not predictable from current data
- Get a biased sample from data

Not missing at random

Eye Data: MAR

- Distance is the reason
- Distant patients have more severe infections
- Different organisms, worse baseline acuity
- But doesn't affect recover beyond baseline acuity and organism

Not MCAR: acuity better among those measured

Consequence...

- Mean acuity among measured higher than among unmeasured
- Worse ulcers are over-represented
- Can ignore missing mechanism
- Missing values behave like measure *controlling for baseline and organism*

So-called missing at random

Missing at Random

- Missing data behaves like over-sampled data
- Get a fair peak at the data assuming sampled based on certain data
- In this case, baseline and organism
- Requires a model: missingness only *associated with acuity at FU only thru baseline and organism*

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

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
in this case baseline and organism
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

Use of MAR Assumption

- Consider the transplant example
- Cold ischemia time missing in 23%
- Unknown in those people
- Assumption: *association between cold_ischemia and other predictors same if data is missing or not*

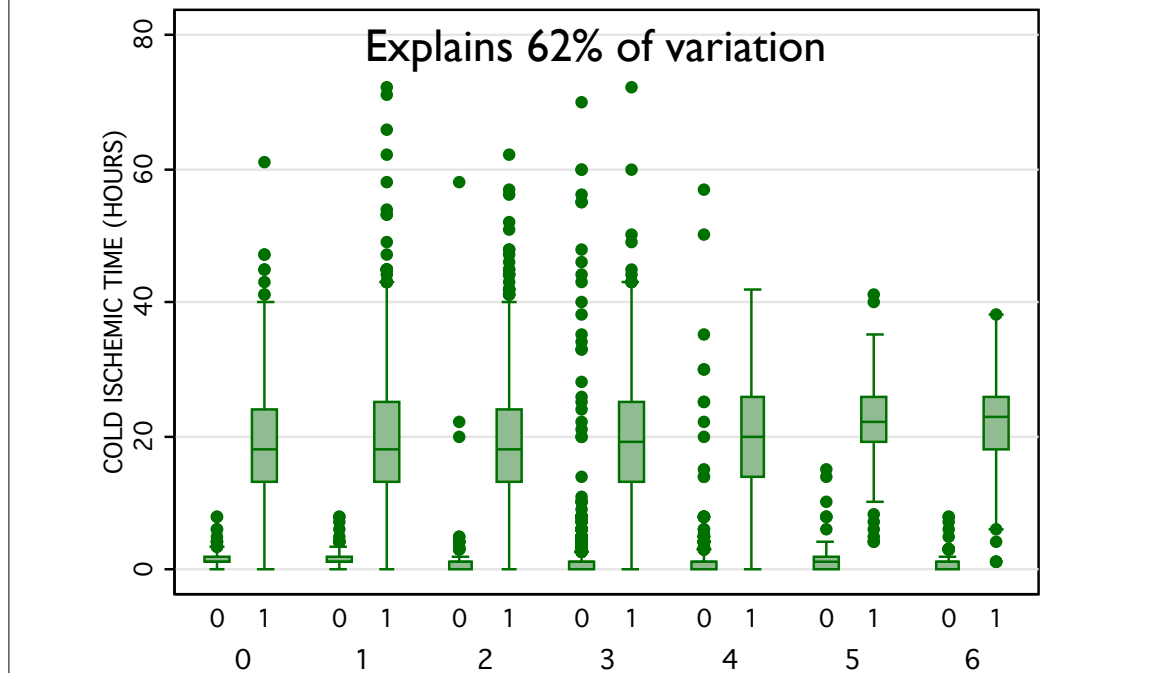
Cold Ischemia Time Living Donor

```
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

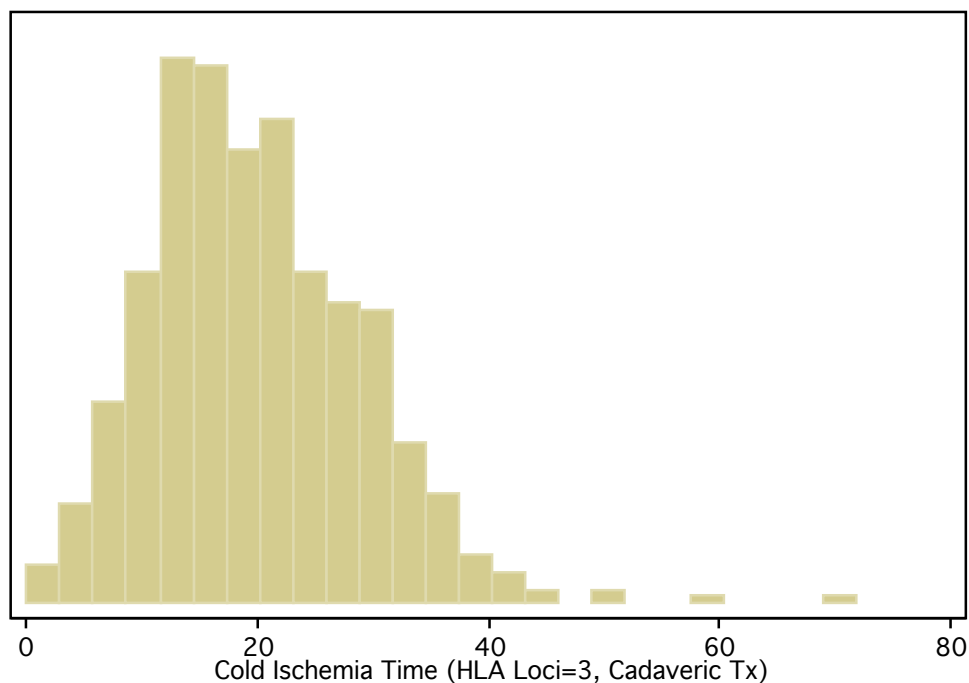
Person	HLA Loci	Txtype	Cold_Isc
1	3	Cadaveric	15h
2	3	Cadaveric	?

This is equivalent to saying that the dist'n 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



MAR

missing cold_isc: 67% living donors
available cold_isc: 50% living donors

missing times are more likely to be shorter

Missingness can be associated with the values -- as long as they are only associated thru HLA and txtype

MAR?

- Not recorded because it is actually 0
- Random 20% recorded at 1 center with bad outcomes
- Not associated with value after knowing observed data

A way forward

- Start with a MAR assumption
- Model missing values
- Perhaps based on regression or matching
- Key is using random variation

Major Methods

- Multiple Imputation:
sample from possible values for missing
- Maximum likelihood:
model missing data
- Both involve assuming MAR
and `averaging' over distn of unknown values
- *Will discuss these next week*

Missing Data Advice

- Avoid missing data if possible
ex: accept mis-timed over a missing visit
- Describe missing data
- Attempt to understand reasons for missing data
- Avoid assuming MCAR
unlikely unless reasons are extremely benign