Missing Values

VGSM, Chapter 11

Dave Glidden
8 October 2013

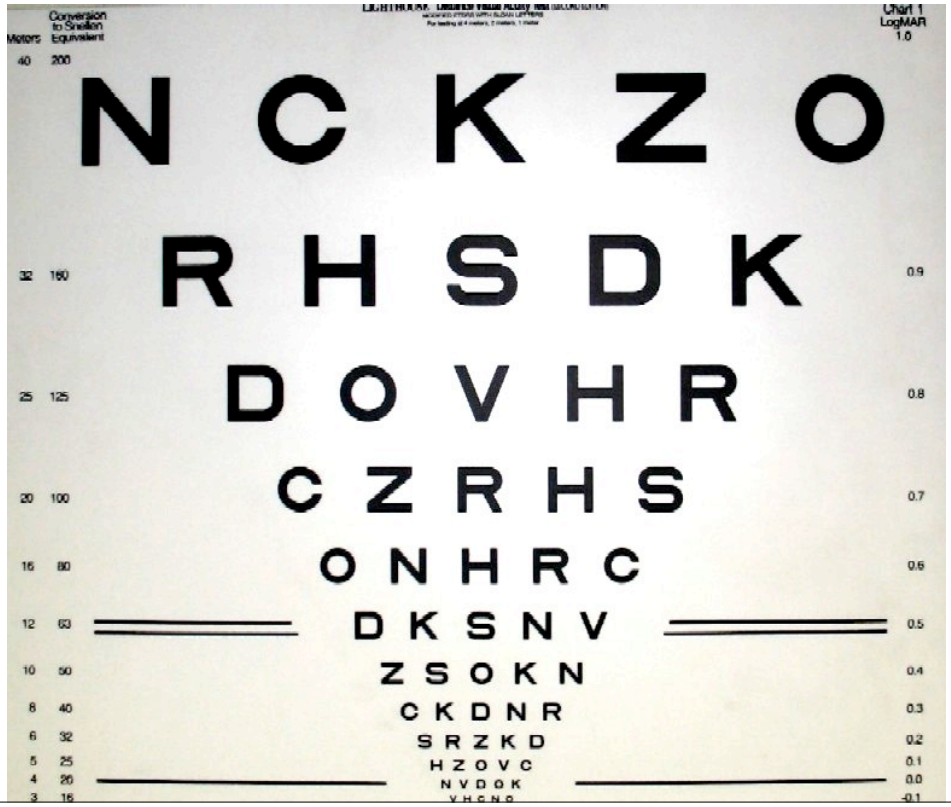
Outline

- Missing data problem
- Sampling and missing data
- Taxonomy of missing data
- Modeling missing values

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

Quick Illustration

- Compare GEE and mixed models
- With "complete data"
  better word -- balanced data
  every person has a value at every timepoint
- With missing values that depend on baseline value
- People with bad baseline vision don't return

Mean logMAR

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)
```

GEE based method

complete data

```
xtgee logmar i.vaspoint, corr(indep) robust i(id) link(identity) family(normal)

                                (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semirobust
      logmar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vaspoint |
           2 |  -.3191327   .0204863   -15.58   0.000    -.3592851   -.2789802
           3 |   -.467602   .0240345   -19.46   0.000    -.5147088   -.4204953
           4 |  -.5094388   .0260931   -19.52   0.000    -.5605803   -.4582973
             |
       _cons |   .9094388   .0321733    28.27   0.000     .8463803    .9724973
------------------------------------------------------------------------------
```

Mixed Models

complete data

```
xtmixed logmar i.vaspoint || id: , ml
```

logmar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
vaspoint						
2	-.3191327	.0200753	-15.90	0.000	-.3584795	-.2797858
3	-.467602	.0200753	-23.29	0.000	-.5069489	-.4282552
4	-.5094388	.0200753	-25.38	0.000	-.5487857	-.4700919
_cons	.9094388	.0291278	31.22	0.000	.8523493	.9665282

Mixed models

with missing data

logmar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
vaspoint						
2	-.3121872	.0294344	-10.61	0.000	-.3698776	-.2544968
3	-.4086854	.0352252	-11.60	0.000	-.4777254	-.3396453
4	-.4937575	.0519739	-9.50	0.000	-.5956245	-.3918906
_cons	.9094388	.0301893	30.12	0.000	.8502689	.9686086

GEE based method

with missing data

logmar	Coef.	Semirobust Std. Err.	z	P>|z|	[95% Conf.	Interval]
vaspoint						
2	-.4927721	.039634	-12.43	0.000	-.5704533	-.4150909
3	-.6148736	.0512371	-12.00	0.000	-.7152965	-.5144506
4	-.7569388	.0411782	-18.38	0.000	-.8376465	-.676231
_cons	.9094388	.0321733	28.27	0.000	.8463803	.9724973

Mirrors answer based on x sectional means

GEE v. Mixed Models

- GEE loses it's robustness with some kinds of missing data
- Hence, it can give misleading answers
- Mixed models can work under many missing data mechanisms
- But requires parametric assumptions *including those about correlation structure*

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 Indicator

- Does poorly in regression models
  even if data is MCAR
- Regression model for post-tx survival:
  pred: cadaveric organ, conf: cold ischemia
- Have to make cold ischemia categorial
  category for missing cold ischemia
- For 23% of data, no adj for cold ischemia
- txtype coef. is blend of crude & adjusted

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 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*

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 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

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

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
hlamat	9541	2.57363	1.378919	0	6

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

Use of MAR Assumption

- 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

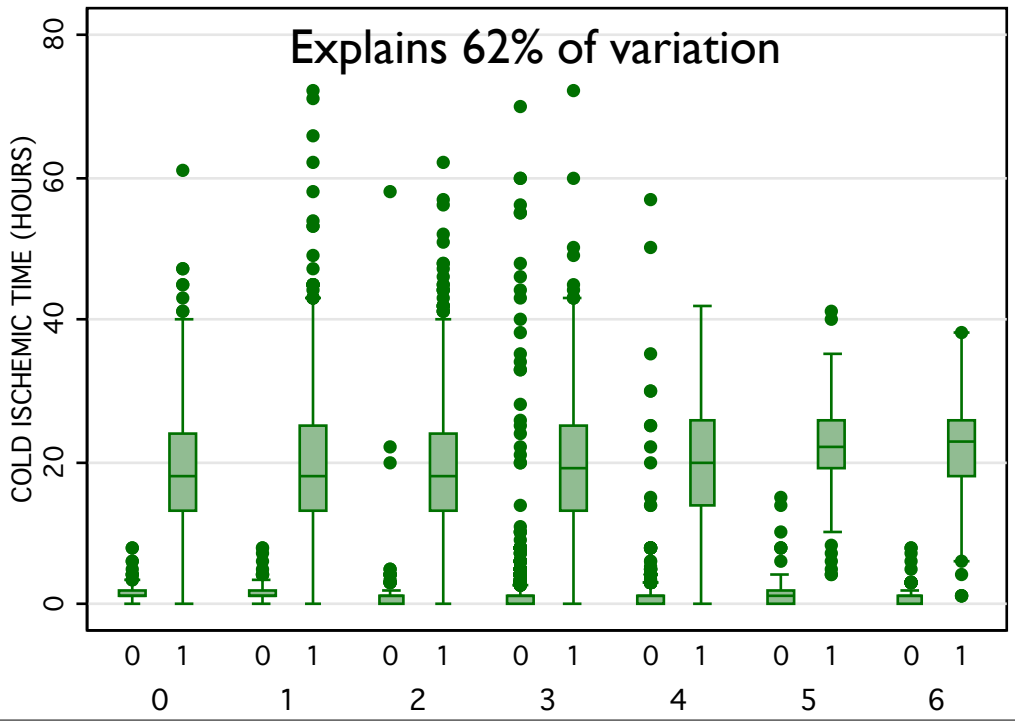
Living Donor

```
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

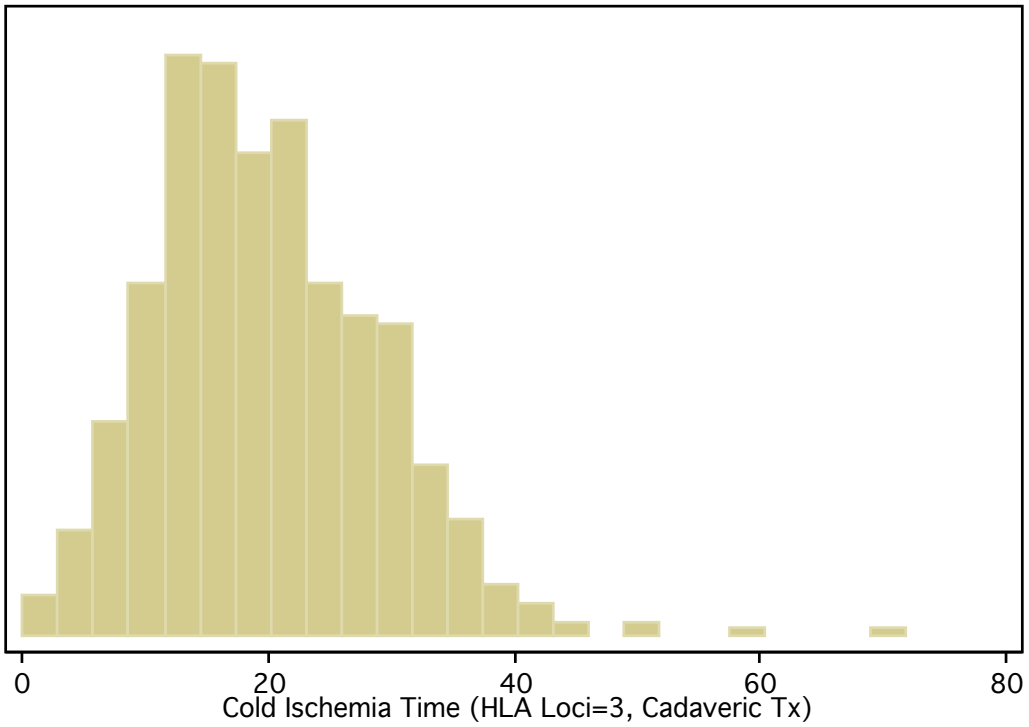
Person	HLA Loci	Txtype	Cold_Isc
1	3	Cadaveric	15h
2	3	Cadaveric	?

This is equivalent to saying that the 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



MAR

Living organs: cold_isc missing 28%
Cadaveric: cold_isc missing 15%

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

Multiple Imputation

- 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

Imputation Model

- 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

Source of Variation

- Predictive distn of cold_isc
  - $cold\_isc_i = \alpha+\beta_1*x_1+\beta_2*x_2+..\ \beta_p*x_p+\varepsilon_i$
- 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 = $(OR_1 + OR_2 + ... + OR_K)/K$
= 1.19

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

= $(1+1/5)*\{ 0.06 \}^2 +$
$(0.24)^2$
= .058

MI SE = sqrt(MI Variance) = 0.24

Imputation Steps

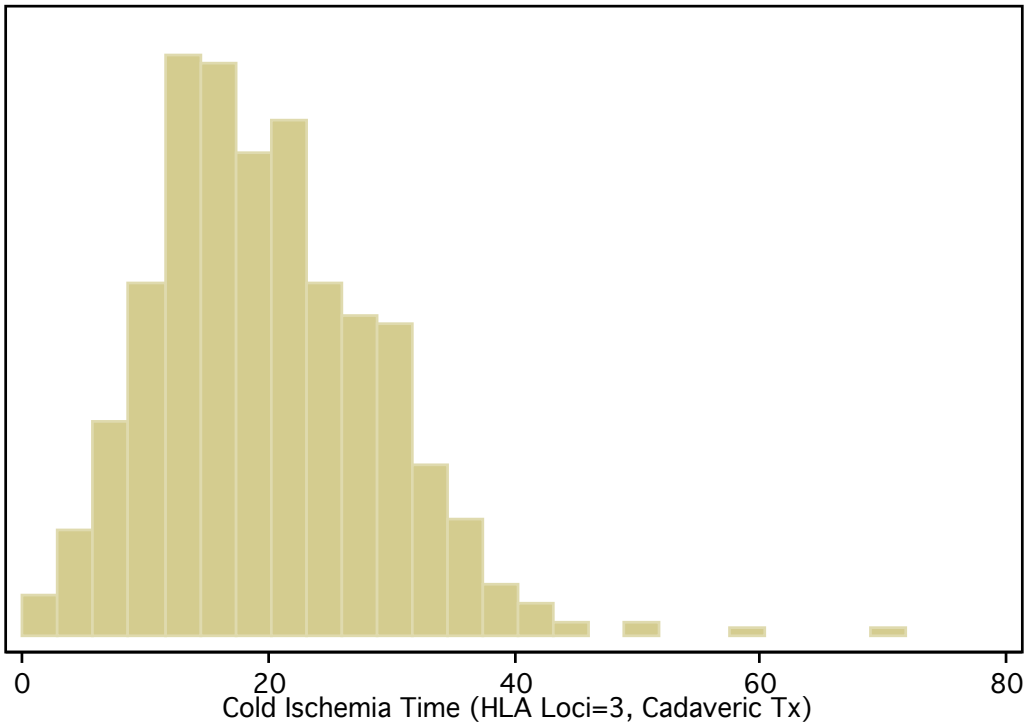
- Can be done manually
- Process would be tedious
- And error prone
- Stata has capability to do this for you
- And display nice output

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

3 HLA, Cadaveric Tx



Good for motivating

Not the best in practice

- How confident in MAR: HLA, txtype
  if true, then there isn't any bias
  so why *not* include outcomes?
  and other things?
- Suddenly not easy to find matches
- Better distribution: base it on regression
- outcome: cold_isc, predictors: everything

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