Missing Data

- Paul L. Hebert, PhD
- Research Associated Professor
- Department of Health Services
- UW School of Public Health

Outline

- Types of missing data
- Missing data mechanisms
- Solutions to the missing data problem
- Paper by Engels

Types of missing data

- Survey sampling
- Item non-response from a survey
- Loss to follow-up in a longitudinal survey
- Censoring
Survey sampling

- Draw a sample of potential survey respondents from the population
- Non-response means the respondents don’t look like the population anymore
  - Generalizability problem
- Solution: survey weights
- Not the type of missing data we are concerned with today.

Item non-response

Missing data on individual items (i.e., questions) in a survey

- Patients refuse to answer the question
  - Income is notorious for this
  - Reshaping the question (right) may help
    - 15% missing if ask for “exact” income
    - 15% missing for income in categories
      - 5-9% for <$50k
- Patients get tired of answering questions and hang-up halfway through the survey
  - Ask important questions first (?)

Loss-to-follow-up

- This pertains to longitudinal data or panel data
  - Follow the same person over time
  - Take measurements either at pre-arranged times (like in an RCT) or at random times (like in observational health care use data)
- Patient is there at baseline but is lost or misses a follow-up appointment
  - Withdraws from a trial or panel survey
  - Gets too sick to respond to the survey
  - With Medicare: Enters an HMO.
  - With Medicaid: loses eligibility
- Drop out/attrition: Patient misses a follow-up visit and then never returns
- Missing observations: patient misses one follow-up visit but returns for the next
Censoring

- Say you had a big trial that was enrolling people over several years and following them for a minimum of 1 year and a maximum of 3 years.
- At the end of the trial you have five cohorts of patients
  - Patients who gave complete responses for all three years
  - Patients who incomplete responses for all three years—i.e., they missed some scheduled follow-up visits in the middle
  - Patients who died before they gave three years of follow-up information
  - Patients who were lost to follow-up before giving 3 years of follow-up information
  - Patients who were administratively censored after <3 years of follow-up—i.e., they may have given more information but the study stopped and no one asked them

Censoring due to death

- Note: death is not generally considered loss-to-follow up
- If your only missing data is because patients died, use the data as if it were MCAR (see below for definition of MCAR).
- If you are studying functioning data (e.g., SF-12, etc), you could use Diehr method:
  - Categorize SF-36 scores into “good” health or “bad” health, based on some a priori categorization.
  - Code observations from patients who died into the lowest functioning category.
  - Estimate a logit model
    $$\text{Prob}(\text{Good Health}) = \text{logit}(\beta_0 + \beta_1 X_0 + \beta_3 T)$$
  - A significant $\beta$, means treatment leads to better health, after taking into account death.


Data used for today’s presentation

Systolic blood pressure from 416 patients with uncontrolled hypertension in a hypertension disease management trial in East and Central Harlem

<table>
<thead>
<tr>
<th>Time</th>
<th>Mean</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>153mmHg</td>
<td></td>
</tr>
<tr>
<td>9-month</td>
<td>140mmHg</td>
<td></td>
</tr>
<tr>
<td>18-month</td>
<td>138mmHg</td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Types of missing data
• **Missing data mechanisms**
• Solutions to the missing data problem
• Paper by Engels

Taxonomy of missing data mechanisms

• Missing completely at random (MCAR)
• Missing at random (MAR)
• Missing not at random (MNAR)

Missing completely at random (MCAR)

• The probability that an observation is missing is a completely random event that is uncorrelated with any patient clinical or demographic factors (X)

\[ \text{Prob}(Y_{ij} = 1 | X_i, Y_{ij}) = u \]

– E.g., the probability that BP (Y) at follow-up time=1 is missing (.) for patient i is a function only of the underlying (and completely random) probability of being missing (u)
Missing at random (MAR)

- There are two flavors of MAR
  - MAR(1): Covariate dependent missing: The probability that an observation is missing depends on the observed covariates, but no other factor relating to the unobserved observation.
    \[
    \Pr(Y_{i2} | X_{i1}, Z_{i0}) = f(X_{i1})
    \]
  - MAR(2): The probability an observation is missing is a function of observed covariates and past or future observed values of the dependent variables, but not on other factor relating to the unobserved observation.
    \[
    \Pr(Y_{i2} = Y_{i0}) = f(X_{i1}, Z_{i0})
    \]
- E.g., The probability that BP is missing for patient i at follow-up might be a function of the baseline BP reading \(Y_{i0}\) for that patient.

Missing at not at random (MNAR)

- Also known as informative missing
  - MNAR(1) The probability that an observation is missing depends on the value of the missing observation.
    \[
    \Pr(Y_{i2} | X_{i1}, Y_{i0}) = f(X_{i1}, Y_{i0})
    \]
  - E.g., a patient’s blood pressure was so high s/he was admitted to the hospital and missed the follow-up BP reading.
  - MNAR(2) The probability an observation is missing is a function of some factor \(Z\) that might be correlated with the value of the missing data.
    \[
    \Pr(Y_{i2} = Y_{i0}) = f(X_{i1}, Z_{i0})
    \]
  - E.g., The probability that BP is missing for patient i at follow-up might be a function of the fact the patient left for vacation \(Z\).

Examples of missing data mechanisms

<table>
<thead>
<tr>
<th>Missing Data Mechanism</th>
<th>MCAR, MAR, or MNAR?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey personnel gets mugged coming back from interview with the patient and the laptop is stolen</td>
<td></td>
</tr>
<tr>
<td>Patient misses a follow-up BP reading because he was in prison</td>
<td>MCAR or MAR</td>
</tr>
<tr>
<td>Patients recruited from one site tend to miss follow-up appointment more than others</td>
<td></td>
</tr>
<tr>
<td>Patients with lower education levels tend to miss appointments more often</td>
<td></td>
</tr>
<tr>
<td>A patient misses the follow-up BP reading because he was hospitalized</td>
<td></td>
</tr>
<tr>
<td>A patient is lost to follow-up because he moved</td>
<td></td>
</tr>
<tr>
<td>A patient's follow-up BP observation is missed because the scheduled follow-up visit fell between Christmas and New Years</td>
<td></td>
</tr>
</tbody>
</table>
Notes on missing data mechanisms

- MCAR, MAR, and MNAR are fairly standard acronyms. MAR(1/2) and MNAR (1/2) are just my takes on individual author’s interpretations of these mechanisms
- MCAR<MAR<MNAR
  - If you can show that the data are MCAR then any statistical procedure that is robust to data that are MAR are also valid for you data

Testing for MCAR versus MAR

- Code the missing observation as 1 if missing and 0 if observed.
- Run a logistic regression of: 
  \[ \text{Prob}(Y_{it} = 1) = \logit(\beta_0 + \beta_1X_i + \beta_2T_i) \]
  where T is treatment assignment
- If nothing is significant than the data are MCAR and any statistical procedure you use is valid
- In particular, if \( \beta_2 \) is not significant then missingness does not differ by treatment assignment, which is especially important

Example: missing follow-up BP as a function of baseline characteristics

Logistic regression

| Variable  | Odds Ratio | Std. Err. | z   | P>|z|  | [95% Conf. Interval] |
|-----------|------------|-----------|-----|-----|----------------------------|
| age       | .9930458   | .0078345  | -0.88 | 0.376 | .9778086 - 1.008521 |
| female    | .5942291   | .1032213  | -3.00 | 0.003 | .422761 - .835243 |
| Hispanic  | 1.051846   | .2883253  | 0.18 | 0.854 | .704427 - 1.600023 |
| Black     | 1.017396   | .4299146  | 0.04 | 0.970 | .4444295 - 2.329041 |
| Diabetes  | 1.270181   | .2183412  | 1.39 | 0.164 | .9068719 - 1.779043 |
| Spanish   | 1.246429   | .3647774  | 0.75 | 0.452 | .7023552 - 2.216967 |
| Education | 1.037029   | .2022647  | 0.19 | 0.852 | .7075713 - 1.519887 |
| Education | 0.6992848  | .1689949  | -1.48 | 0.139 | .4354575 - 1.122955 |
| BMI       | .9907371   | .0102292  | -0.90 | 0.367 | .9708898 - 1.010991 |
| Smoker    | 1.047112   | .1988888  | 0.24 | 0.808 | .7215246 - 1.519395 |
| CAD       | 1.045762   | .2044002  | 0.23 | 0.819 | .7129528 - 1.533929 |
| Alcohol   | 0.6463659  | .22381    | -1.26 | 0.208 | .3279006 - 1.274132 |
| Diabetes  | 1.075092   | .1684416  | 0.46 | 0.644 | .7908291 - 1.461534 |
| Depression| 1.069269   | .2233887  | 0.32 | 0.749 | .7099964 - 1.61034 |
| Psychosis | 1.139513   | .2618339  | 0.57 | 0.570 | .7263276 - 1.787746 |
| Renal     | 0.6854961  | .1540675  | -1.68 | 0.093 | .4412601 - 1.064916 |
| Nurse     | 1.065933   | .1971786  | 0.35 | 0.730 | .7417763 - 1.531747 |
| BP         | 1.007935   | .1940874  | 0.04 | 0.967 | .6947743 - 1.470753 |
Testing for MAR(1) versus MAR(2)

- Code the missing observation as 1 if missing and 0 if observed.
- Run a logistic regression of:
  \[ \text{Prob}(Y_i = 1) = \logit(\beta_0 + \beta_1 X_i + \beta_2 Y_0) \]

- If \( \beta_1 \) is significant than the data are MAR(2)
- This is technically no worse than MAR(1) but if \( \beta_1 \) is not significant you can make a stronger case that the data are not MNAR

Example: missing follow-up BP as a function of baseline characteristics and baseline blood pressure

Logistic regression  Number of obs = 1167
LR chi2(19) = 38.52
Prob > chi2 = 0.0051
Log likelihood = -542.71435 Pseudo R2 = 0.0343

|                | Odds Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----------------|------------|-----------|------|------|----------------------|
| age_c          | 0.9930776  | 0.0078579 | -0.88 | 0.380 | 0.9777951 - 1.008599 |
| female         | 0.5675809  | 0.0992118 | -3.24 | 0.001 | 0.4029394 - 0.799495 |
| hispanic       | 1.035645   | 0.2849481 | 0.13  | 0.899 | 0.6039634 - 1.77587  |
| blackhisp      | 1.002144   | 0.4226545 | 0.01  | 0.996 | 0.4384658 - 2.29047  |
| stofhla_inad   | 1.259742   | 0.2178222 | 1.34  | 0.182 | 0.8976362 - 1.767922 |
| spanish        | 1.236306   | 0.3626248 | 0.72  | 0.470 | 0.6957564 - 2.19682  |
| edhs           | 1.022394   | 0.2009586 | 0.11  | 0.910 | 0.6955184 - 1.502892 |
| edgths         | 0.7022775  | 0.1710049 | -1.45 | 0.147 | 0.4357534 - 1.131818 |
| bmi_c          | 0.9946638  | 0.0104722 | -0.51 | 0.611 | 0.974349 - 1.015402  |
| smoker         | 1.033435   | 0.1968824 | 0.17  | 0.863 | 0.7114082 - 1.501232 |
| cad            | 1.023812   | 0.2014712 | 0.12  | 0.905 | 0.6961711 - 1.50565  |
| etoh           | 0.6062512  | 0.2107168 | -1.44 | 0.150 | 0.3067592 - 1.19814  |
| diabete        | 1.098418   | 0.1725486 | 0.60  | 0.550 | 0.8073351 - 1.49445  |
| depress        | 1.148184   | 0.2412788 | 0.66  | 0.511 | 0.7605719 - 1.733335 |
| psych          | 1.159661   | 0.2674898 | 0.64  | 0.521 | 0.7378891 - 1.822514 |
| renald         | 0.6797639  | 0.1536475 | -1.71 | 0.088 | 0.4364761 - 1.058658 |
| nurse          | 1.115677   | 0.2081084 | 0.59  | 0.557 | 0.7740393 - 1.608104 |
| bpm            | 0.9825922  | 0.1903834 | -0.09 | 0.928 | 0.6721213 - 1.436478 |
| bpsys0         | 1.014676   | 0.0046269 | 3.20  | 0.001 | 1.005648 - 1.023785  |

Testing MAR vs MNAR

- Technically this cannot be done since to test for MNAR you would have to know the value of the missing observation. If you knew it, it wouldn’t be missing.
- You can look for hints of MNAR, like look at the trajectory of BPs leading up to the missing values
Testing MAR vs MNAR

- You can also try to incorporate external data.
  - If you have hospitalizations from administrative data, see if missing survey response is a function of whether a patient was hospitalized in some time vicinity around the follow-up measurement time.
  - Incorporate data from patient charts.

Potential solutions to missing data problems

- Ignore the missing data
  - Item deletion or substitution
  - Complete case analysis
  - Available data analysis
- Impute the missing data
  - Ad hoc methods
  - Model-based methods
  - Multiple imputation
- Inverse probability of missing weighting
- Special considerations for longitudinal data
  - Likelihood based estimation vs GEE

Ignore the missing data

- Item deletion: Remove from your analytic model variables that are often missing and not very important
  - Do you really need the marital status variable in the model if it is often missing?
  - Can you replace the income variable (which is often missing and frequently a lie even when it is not missing) with education?
- Case deletion or Complete Case Analysis
  - Toss any person from your sample that failed to respond to any survey item or survey
  - This is fine, if data are MCAR, but inefficient because you can throw out a ton of good data
  - If 1% of observations are missing on 20 different items you can end up throwing out 20% of your data
Ignore the missing data (2)

- Available data analysis (for longitudinal data)
  - Use all the data people reported and just ignore the fact that some people were lost to follow-up
- This is fine if data are MCAR
  - The data that you don’t see is just a random sample of the data you do see, so doing the analysis of the available data is just like doing an analysis of a random sample of people.
- Inappropriate if the data are MAR or MNAR
  - If data are MAR then certain types of people are more likely to have missing data. Your analysis plan should take this into consideration.

Impute the missing data: Ad hoc methods

- If the data are not that important to the analysis and can be legitimately put into categories, then put them in categories and create a “missing” category.
  - Make sure the missing is not the omitted category in any regression
  - This is often used with variables like race (i.e., have an “Other/Unknown” category, missing BMI)

<table>
<thead>
<tr>
<th>Education level</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; High sch</td>
<td>220</td>
<td>52.88</td>
<td>52.88</td>
</tr>
<tr>
<td>High sch</td>
<td>113</td>
<td>27.16</td>
<td>80.05</td>
</tr>
<tr>
<td>High sch</td>
<td>77</td>
<td>18.51</td>
<td>98.56</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1.44</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>416</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Impute missing data: ad hoc methods (2)

- Put in the sample mean value for the missing data
  - E.g., if mean BP in he sample is 135, then for every person who has a missing BP reading impute the number 135 for the missing value
- Issues
  - Doesn’t make good use of the data.
    - If you know men have higher BP on average then women, then wouldn’t you want to impute a higher BP if the missing data were from a man?
  - Significantly reduces the sample variance
Imputing missing BP data with the sample mean

Ad hoc methods (3): Last data carry forward, Next and last

• This applies to longitudinal data
• Last data carry forward (LDCF): Impute missing value at time t for person i with the last non-missing value for person i
  • If baseline BP was 135 for an individual and he never returned for any follow-up measures (and he wasn't deceased at the end of the study), then replace all of his missing follow-up BP measures with 135
• Next and last (NAL): Impute missing observations with the average of the observation immediately preceding and immediately following the missing measure.
• ISSUES
  • LDCF is robust to data that are MAR
  • Can be highly effective (see Engle/Diehr article) if there is a strong individual-specific component to the data, and data are stable within individual over time.
  • Like mean imputation, this reduces the sample variance inappropriately—i.e., that person's blood pressure wasn't really 135 mm Hg each time. It varied but the imputed value didn't.

Imputing missing BP data with LDCF
Potential solutions to missing data problems

- Ignore the missing data
  - Item deletion or substitution
  - Complete case analysis
  - Available data analysis
- Impute the missing data
  - Ad hoc methods
  - Model-based methods
  - Multiple imputation
- Special considerations for longitudinal data
  - Likelihood based estimation vs GEE

Model based methods

Model predicted mean imputation without (left) and with (right) additional random variation
Potential solutions to missing data problems

- Ignore the missing data
  - Item deletion or substitution
  - Complete case analysis
  - Available data analysis
- Impute the missing data
  - Ad hoc methods
  - Model-based methods
  - Multiple imputation
- Special considerations for longitudinal data
  - Likelihood based estimation vs GEE

Multiple imputation (MI): Motivation

- The imputations you make should reflect your uncertainty about the imputations.
- Each of the β’s in the regression have a point estimate but also a standard error around them.
- If you aren’t sure what the β’s are, how can you be certain of the imputation you derived from the β’s?

Multiple imputation: approach

- The approach to MI involves three steps which is repeated “M” times
  1. Imputation step: Impute the missing data M times, each time estimating an imputation like:
     \[ \hat{Y}_k = f(Y_{-k}, \hat{\beta}_k, \hat{\sigma}_k) \]
     Each time draw a different value of \( \hat{\beta}_k \) based on the sampling distribution of the \( \hat{\beta}_k \).
  2. Estimation step: run whatever analysis you are doing “M” times, one for each imputation. That is, calculate the comparative effectiveness of the treatment “M” times.
  3. Pooling step: Combine the “M” estimates of the effect of treatment into a single estimate:
     1. Combine the point estimate of the treatment effect in just the results of the individual treatment effect.
     2. The estimate calculation is performed across all of the “M” studies, but is typically a weighted average of the weighted imputation estimates of the variable in the treatment effect, and this variance in the treatment effects between imputations.
Multiple Imputation: Issues

- **Issues**
  - Robust to data that are MAR
  - Gives confidence intervals that appropriately reflect your uncertainty about your imputations
  - Has better coverage properties than other methods
  - Sounds hard but is easy to do in Stata ("mi" command)
  - Easy to multiply impute all of your missing data at once in a single step
- **Problems:**
  - Implementing Rubin’s rules if you are using non-standard Stata commands is a mild pain.

Potential solutions to missing data problems

- Ignore the missing data
  - Item deletion or substitution
  - Complete case analysis
  - Available data analysis
- Impute the missing data
  - Ad hoc methods
  - Model-based methods
  - Multiple imputation
- Special considerations for longitudinal data
  - Likelihood based estimation vs GEE

Special considerations for longitudinal data

- If you have longitudinal data and data that (you think) are MAR, you the option of estimating models using likelihood based methods (xtmixed in Stata, Proc Mixed in SAS, HLM, etc).
  - If the likelihood of observing the data (i.e., the likelihood the data are not missing) can be separated from the likelihood of the data given that it is observed, then likelihood based methods are robust to data that are MAR.
  - This is not true for other methods, like generalized estimating equations (GEE)
    - GEE is only robust to data that are MCAR
Summary

• Types of missing data
• Missing data mechanisms
  • MCAR, MAR, MNAR
• Solutions to the missing data problem
• Paper by Engels

References & Resources


