Universität Konstanz

III. Nonresponse

Session 6: Empirical strategies for missing data

Overview

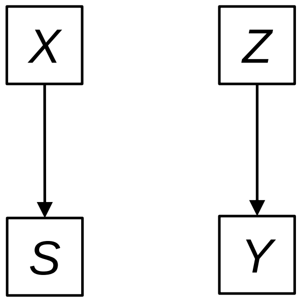
- Introduction
- Assumptions about missingness
- Conventional strategies
- Information available about those missing
- Reweighting
- Hot decking
- Single imputation
- Multiple imputation
- Final remarks

Introduction

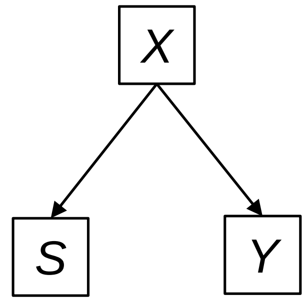
- Nonresponse introduces statistical inefficiency
- Nonresponse will also lead to biased estimates if the propensity to respond to the survey (or an item) is related to the attribute of interest
- Ex ante attempts to reduce unit and item nonreponse seldom are fully successful (e.g. increase and dispersion of contact attempts, incentives, conversion by specially trained interviewers, reduction of respondent burden / interview length via matrix sampling etc.)
- Nonresponse rate and nonresponse bias seem hardly related
- Importance of ex post strategies for missing data
- Different strategies are based on different assumptions about missingness and different kinds of information available on those missing

Assumptions about missingness

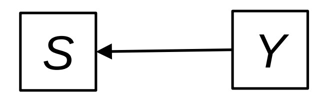
Missing completely at random (MCAR)



Missing at random (MCAR)



Missing not at random (MNAR)



Conventional strategies

Listwise deletion, complete case analysis

- Default in many statistical software packages
- Loss of ~1/3 of the cases in typical regression models based on survey data
- Decrease in statistical efficiency
- "Listwise deletion is evil" (King 1998)
- Potential bias if MCAR does not hold
- If MAR holds, analysis including the relevant Xs yield unbiased results

Pairwise deletion

- Uses different sets of sample units for different params
- Nonconstant n, therefore difficult to obtain estimates of standard error / CIs of analytic statistics

Conventional strategies

Mean imputation

- Descriptive statistics: leaves $\bar{y}$ unchanged, but decreases variances
- Overestimation of confidence
- Analytical statistics: decrease in correlations

Logical / best guess imputation

- E.g. due to filtering
- Eg. knowledge questions

Information about those missing

- Population information about the distribution of variables ***x*** presumedly responsible for missingness from censuses, other sources
- Individual measurements on nonrespondents from sampling frames, record linkages, screening interviews, interviewer observations, previous panels in a panel study, other items in the same survey (in instances of item nonresponses)
- Individual measurements on „unlikely“ respondents using contact histories, reports of intentions to respond to a later survey, multi-phase sampling strategies

Reweighting

Poststratification

- Ex post stratification of the sample along one or several categorical variables ***x*** …
  - … which are suspected to co-determine missingness and the attribute of interest *y* („common causes“ according to Groves, 2006)
  - … whose (joint) population distribution is known
- Conceive of the categories / cross-classifications of ***x*** as groups or strata *h*, use the stratified estimator
- Corrects for disproportionalities between $\frac{N_h}{N}$ and $\frac{n'_h}{N}$ that may be either due to sampling or due to nonresponse
- Assumes MAR

Reweighting

Poststratification

- Poststratified estimator of $\bar{Y}$:

$$\bar{y}_{PS} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h'$$

- The sampling variance of of the poststratified estimator is estimated by

$$V(\bar{y}_{PS}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sum_{i=1}^{n_h} (y_i' - \bar{y}_h')^2}{n_h' - 1} \frac{N_h - n_h'}{N_h}$$

Reweighting

Poststratification:

- Inclusion probability: $\pi_i = \frac{n'_h}{N_h} = \frac{n_h}{N_h}\frac{n'_h}{n_h}$
- Combined sampling and response weight: $\omega_i = \frac{1}{\pi_i}$
- Poststratified estimator:

$$\bar{y}_{PS} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}'_h = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n'_h} \frac{y'_i}{\pi_i}$$

Reweighting

Raking

- Poststratification requires that N_h is known and that $n_h' \neq 0$ for all h
- This is rarely met for joint distributions of several $\boldsymbol{X}$
- Situation resembles an ecological inference problem where the margins of a contingency table are known and the inner cells are unknown:

	$X_2=1$	$X_2=0$	
$X_1=1$	?	?	P_1
$X_1=0$	?	?	$1-P_1$
	P_2	$1-P_2$	1

	$x_2=1$	$x_2=0$	
$x_1=1$	✓	✓	p_1
$x_1=0$	✓	✓	$1-p_1$
	p_2	$1-p_2$	1

- *Iterative proportional fitting* of the cell weights ω_h so that the marginal distributions in the sample approximate those in the population

Hot decking

- Requires individual measurements of $\mathbf{x}$ on nonrespondents
- Match individuals with missing *y*-values to one or more individuals for which *y* is observed based on $\boldsymbol{x}$; impute (average) *y*-value
- Nonparametric; considers response *patterns*

Potential problems

- Assumes MAR
- What if there are no identical individuals in the data set?
- How to incorporate matching variance?
- How to incorporate imputation uncertainty?
- What if $\boldsymbol{x}$ also contains missing values?

Hot decking

Potential solutions

- Helpful if there are no identical individuals in the data set
- Model the response indicator using $\boldsymbol{x}$ as covariates (e.g. logit, probit)
- Matching based on predicted response probabilities
- Alternatively, use inverse predicted response probabilities as weights

Single imputation

Deterministic regression-based (conditional mean) imputation

- (Linear) regression of *y* on ***x*** using complete cases:

$$y_i' = \alpha + \boldsymbol{\beta} \boldsymbol{x}_i' + e_i$$

$$e_i \sim Normal(0, \sigma_e^2)$$

- Impute prediced *y*-values for nonrespondents using known x-values and estimated regression parameters:

$$\hat{y}_i = \hat{\alpha} + \widehat{\boldsymbol{\beta}} \boldsymbol{x}_i$$

Single imputation

Potential problems

- Assumes MAR
- Regression model needs to be specified correctly, potentially different distributions of $\boldsymbol{x}$ among respondents and nonrespondents
- Decreases variance in y, increases correlations between imputed y and $\boldsymbol{x}$
- How to incorporate prediction uncertainty?
- How to incorporate imputation uncertainty?

Single imputation

Stochastic regression-based imputation

- Easy way to incorporate prediction uncertainty
- As previously, linear regression of *y* on ***x*** using complete cases
- Random draws of $\hat{e}_i$ from $Normal(0, \hat{\sigma}_e^2)$
- Impute predicted *y*-values for nonrespondents using known ***x***-values and estimated regression parameters plus $\hat{e}_i$: $\hat{y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}\mathbf{x}_i + \hat{e}_i$

Potential problems

- Assumes MAR
- Regression model needs to be specified correctly, potentially different distributions of $\boldsymbol{x}$ among respondents and nonrespondents
- How to incorporate imputation uncertainty?

Multiple imputation

- Way to also incorporate imputation uncertainty
- As previously, but multiple stochastic imputations *m=1,2,…, M*
- Combination of estimates:

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{im}$$

- Sample mean of observed and (average) imputed *y*-values, $\bar{y}_{MI}$, is an unbiased and consistent estimate of $\bar{Y}$ (assuming SRS)
- The combined sampling and imputation variance of this estimate is given by

$$V(\bar{y}_{MI}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\sum_{i=1}^{n}(y_i - \bar{y}_m)^2}{n-1} \frac{N-n}{N} + \left(1 + \frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}(\bar{y}_m - \bar{y})^2$$

Final remarks

- Before applying a strategy for missing data, we ought to ask whether an underlying „true“ value exists and, if so, whether the missing value is truly unknown
- If individual measurements for nonrespondents are available, multiple imputation (MI) is the current gold standard for handling nonresponse, as it incorporates both prediction and imputation uncertainty
- MI makes data analysis more complex (we have to handle *m* data sets)
- Model specification is still a major concern, also MI assumes MAR
- If the missingness process is suspected to be non-ignorable (NMAR), one has to refrain to *selection models* or *pattern-mixture models*
- Some NMAR problems could perhaps be converted into MAR problems by collecting additional data on response propensities