

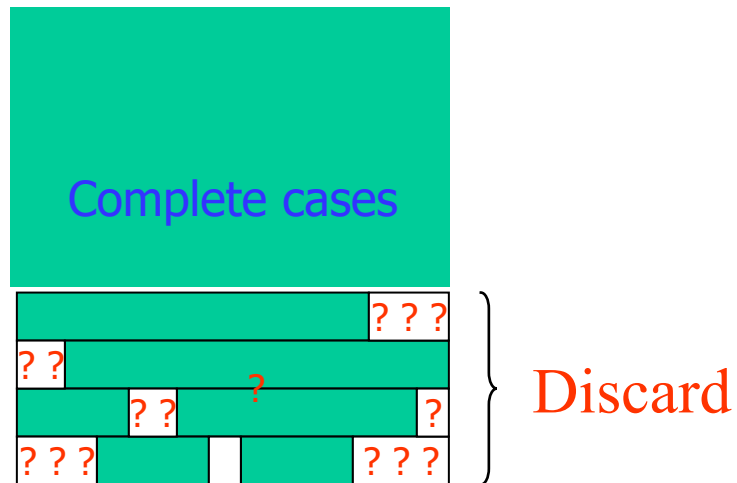
Statistical Analysis with Missing Data

Module 2

Available-Case Complete-Case Analysis and Weighting



Complete Case Analysis



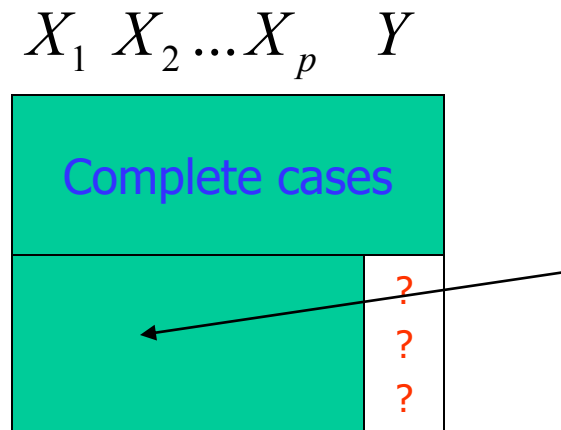
Unweighted CC Analysis

- Easy (but missing values must be flagged!)
- Does not invent data
- Simple and may be good enough with small amounts of missing data
 - but defining “small” is problematic; depends on
 - fraction of incomplete cases
 - recorded information in these cases
 - parameter being estimated

Limitations of CC Analysis

- Loss of information in incomplete cases has two aspects:
 - Increased variance of estimates
 - Bias when complete cases differ systematically from incomplete cases
 - restriction to complete cases requires that the complete cases are representative of all the cases for the analysis in question
 - For some (not all) analyses this implies MCAR, which is often questionable!

Validity of CC analysis depends on how much information in missing data is relevant to parameters of interest



- Y has missing values, $X=(X_1, \dots, X_p)$ are completely observed. The regression of Y on X has a multiple R^2 of 0.4.

How much information is there in an incomplete case for inference about the mean of Y ?

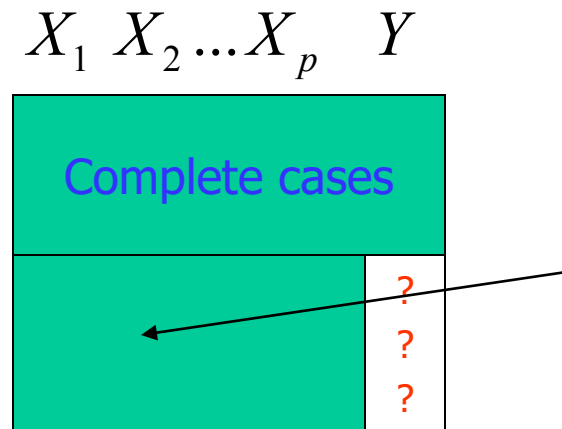
A lot

A moderate amount

A little

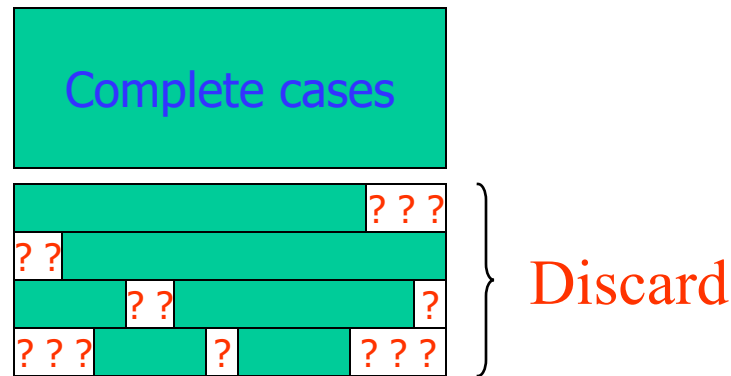
Depends on the form
of the regression

What if X is highly predictive of Y ?



- The regression of Y on X has a multiple R^2 of 0.9.
-
- What about the mean of Y , $E[Y]$?
-
- What about the regression coefficient of X if regressing Y on X ?

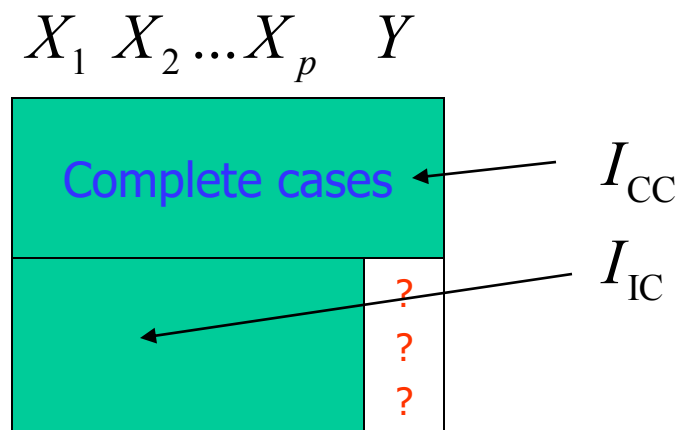
Increased variance from CC analysis



Inefficiency of CC Analysis depends on how much information is contained in the discarded incomplete cases (I_{IC}) relative to the information in the complete cases (I_{CC})

In a likelihood analysis, I_{IC} and I_{CC} are measured by the information matrices for these cases (more later)

Example: univariate nonresponse



Suppose X_1, \dots, X_p are strong predictors of Y

I_{IC} is substantial for unconditional mean of Y

$I_{IC} = 0$ for conditional mean of Y given X_1, \dots, X_p !

Bias of estimate of mean from complete cases

- Unit nonrespondents may differ from respondents, leading to
 - **nonignorable** missing data
 - biased estimates. A simple formula for means:

$$\mu_{CC} - \mu = (1 - \pi_{CC}) \times (\mu_{CC} - \mu_{IC})$$

Bias = NR rate x difference in CC and IC means

CC = respondent, IC = nonrespondent

We don't observe the IC means of missing variables, but comparing CC and IC means or distributions of observed variables is useful

Bias of CC analysis: Simulation Study

- True model:

$$\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$$

$$\text{Logit}[\Pr(\mathbf{E}=\mathbf{1}|\mathbf{X})]=\mathbf{0.5}+\mathbf{X}$$

$$\text{logit}[\Pr(\mathbf{D}=\mathbf{1}|\mathbf{E},\mathbf{X})]=\mathbf{0.25}+\mathbf{0.5X}+\mathbf{1.1E}$$

- Sample size: 500
- Number of Replicates: 5000
- Before Deletion Data Sets

Missing-Data Mechanism

- D and E : completely observed
- X : sometimes missing
- Values of X in each cell are set to missing with the following underlying probabilities:

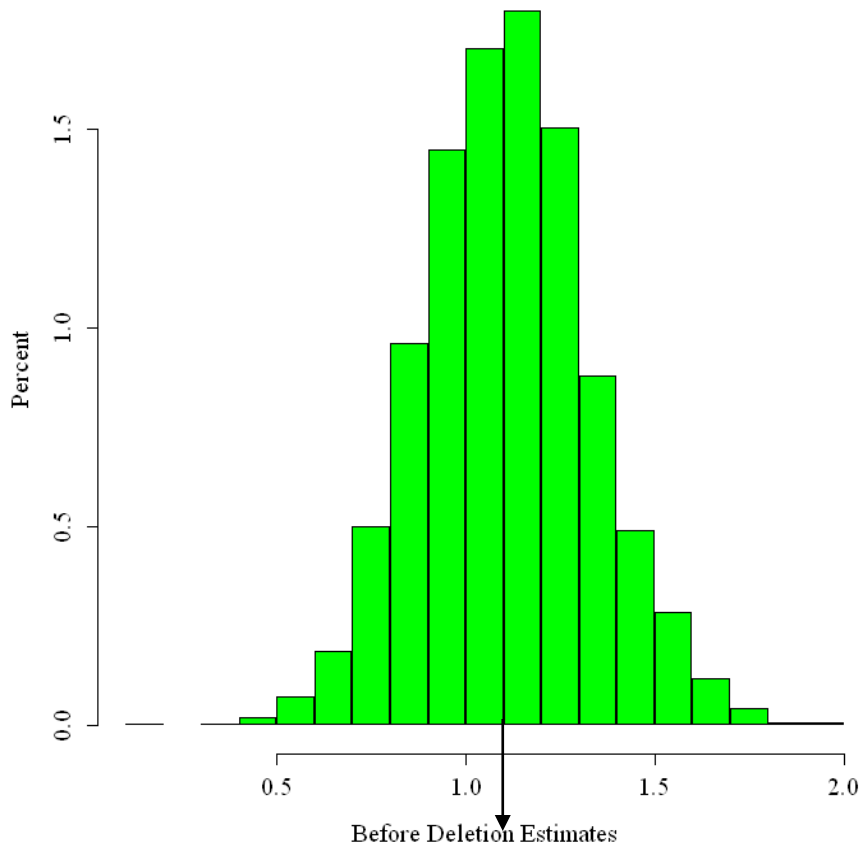
D	E	X
		?
		?
		?

$$D=0, E=0: p_{00}=0.19$$
$$D=0, E=1: p_{01}=0.09$$
$$D=1, E=0: p_{10}=0.015$$
$$D=1, E=1: p_{11}=0.055$$

MAR mechanism

Before Deletion Estimates

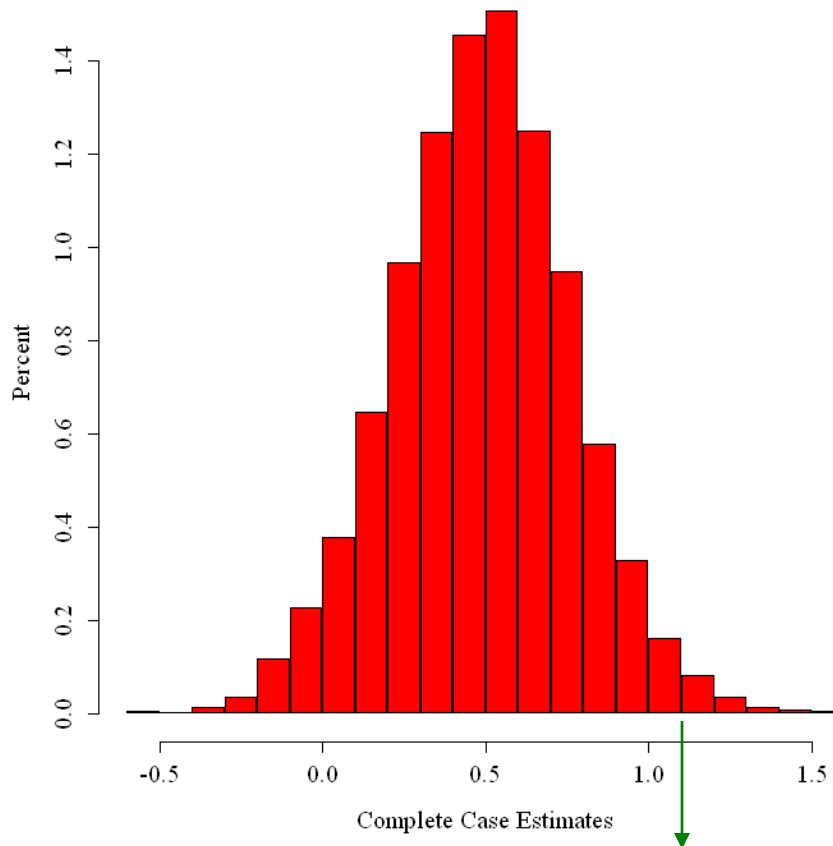
Histogram of 5000 Point Estimates



- Histogram of 5000 estimates before deleting values of X
- logistic model
 $\text{logit } \Pr(D=1|E,X)$
 $=\beta_0+\beta_1 E+\beta_2 X$

Complete-Case Estimates

Histogram of 5000 Point Estimates

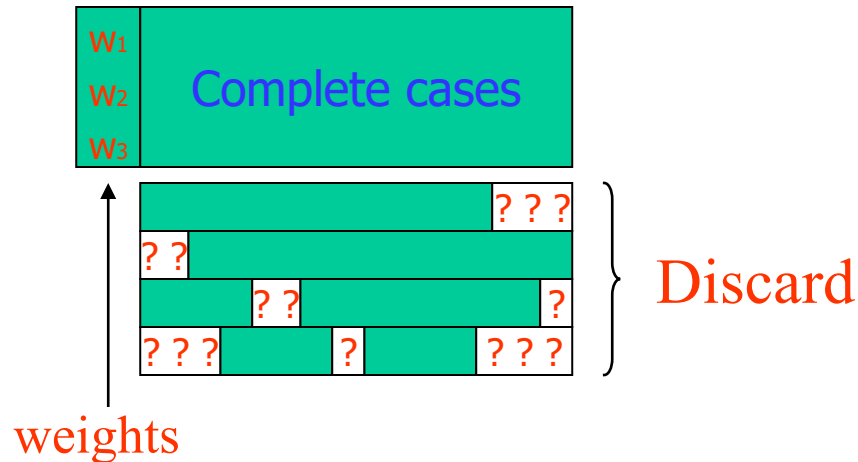


Histogram of
complete- case
analysis estimates

Delete subjects with
missing X values

True value = 1.1,
serious negative bias

Weighted Analysis








- One way to reduce the potential bias of CC analysis is to **weight** complete cases differentially e.g. a CC mean becomes a weighted mean
- Common for unit nonresponse in surveys
- “Quasi-randomization inference” : extends ideas of randomization inference in surveys

Sample survey example: Unit nonresponse

- Suppose we have:
- Design variables Z observed for whole population
- Fully observed survey variables X , measured for respondents and nonrespondents
- Survey variables Y measured only for respondents (see diagram)
- Goal of weighting is to use information in X and Z to weight the respondents, improve estimates

Sample				Pop
Z	X	Y	M	Z

			0
			1



Sampling weights

- In a probability survey, each sampled unit i “represents” w_i units of the population, where

$$w_i = \frac{1}{\text{Pr}(\text{unit } i \text{ sampled})}$$

w_i is determined by the sample design and hence *known*

- Extend this idea to unit nonresponse ...

Unit Nonresponse Weights

- If probability of response was known, could obtain weight for units that are sampled and respond:

$$\begin{aligned}w_i &= \frac{1}{\text{Pr}(\text{unit } i \text{ is sampled and responds})} \\&= \frac{1}{\text{Pr}(i \text{ sampled})} \times \frac{1}{\text{Pr}(i \text{ responds} | \text{sampled})} \\&= (\text{sampling weight}) \times (\text{response weight})\end{aligned}$$

Since prob of response is not known, we need to estimate it.

Weights calculation using adjustment cells

- Group respondents and nonrespondents into adjustment cells with similar values on variables recorded for both:
- e.g. white females aged 25-35 living in SW

100 in sample $\begin{matrix} < \\ \nearrow & \searrow \end{matrix}$ $\begin{matrix} 80 \text{ respondents} \\ 20 \text{ nonrespondents} \end{matrix}$

$\text{pr}(\text{response in cell}) = 0.8$

response weight = 1.25

Back to simulation study

- True model:

$$\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$$

$$\text{Logit}[\Pr(\mathbf{E}=\mathbf{1}|\mathbf{X})]=\mathbf{0.5}+\mathbf{X}$$

$$\text{logit}[\Pr(\mathbf{D}=\mathbf{1}|\mathbf{E},\mathbf{X})]=\mathbf{0.25}+\mathbf{0.5X}+\mathbf{1.1E}$$

- Sample size: 500
- Number of Replicates: 5000
- Before Deletion Data Sets

Missing-Data Mechanism

- D and E : completely observed
- X : sometimes missing
- Values of X in each cell are set to missing with the following underlying probabilities:

D	E	X
		?
		?
		?

$$D=0, E=0: p_{00}=0.19$$

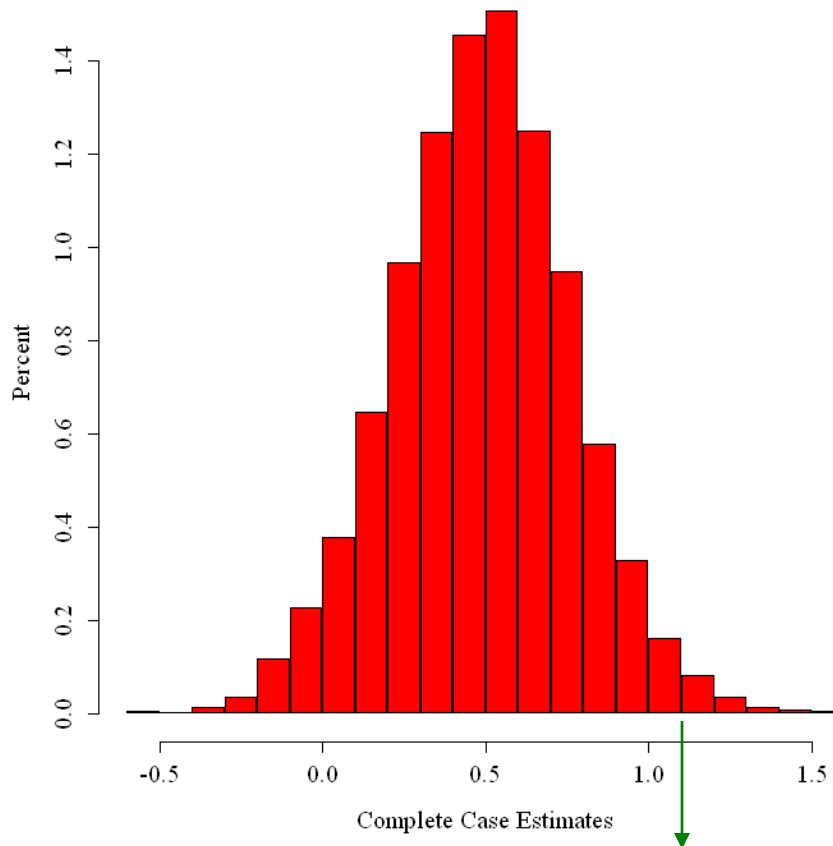
$D=0, E=1: p_{01}=0.09$

$$D=1, E=0: p_{10}=0.015$$
$$D=1, E=1: p_{11}=0.055$$

MAR mechanism

Complete-Case Estimates biased

Histogram of 5000 Point Estimates



Histogram of
complete- case
analysis estimates

Delete subjects with
missing X values

True value = 1.1,
serious negative bias

Weighting the respondents

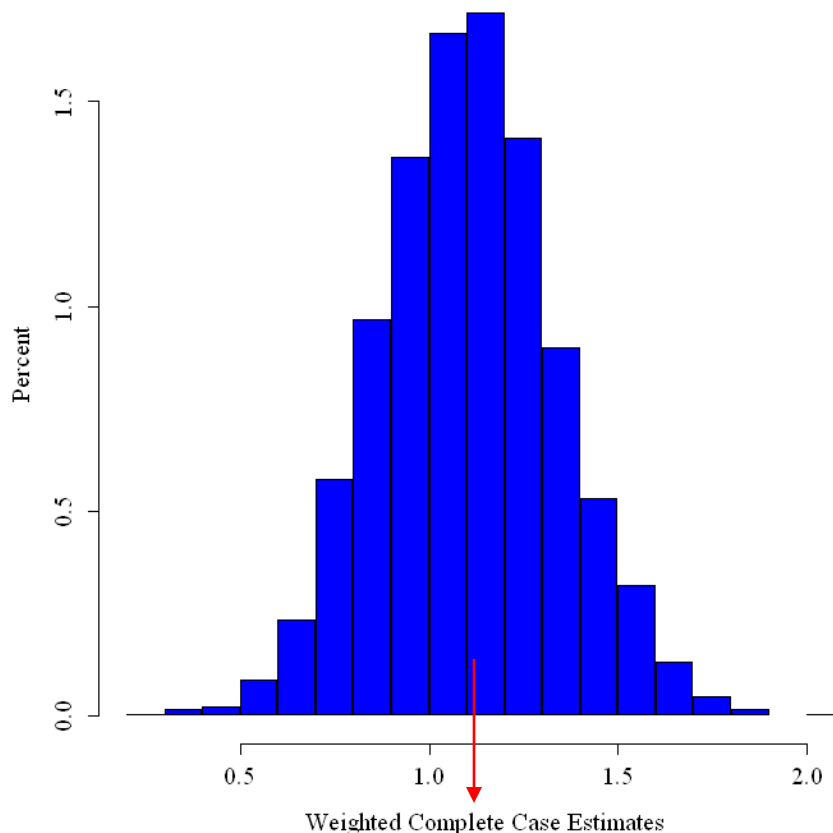
- r_{ij} = response rate in cell $D = i$ and $E = j$
- weight respondents in cell $D = i$ and $E = j$ by

$$w_{ij} = 1 / r_{ij}$$

Weighted Estimates

- Weighted Logistic Regression Model

Histogram of 5000 Point Estimates



Weighted by inverse
of cell-specific
response rates

True Value: 1.1

Bias is removed

Choice of Adjustment Cells

- With extensive covariate information, can't cross-classify on all of them
- How do we choose which variables to use?
- Consider problem of estimate the mean of an outcome Y

X_1	X_2	...	X_p	Y	M
Complete cases					0
					0
					0
				?	1
				?	1
				?	1

Impact of weighting for nonresponse

$$\text{corr}^2(X, Y)$$

		Low	High
$\text{corr}^2(X, M)$	Low	---	var \Downarrow
	High	var \Uparrow	var \Downarrow bias \Downarrow

Too often adjustments do this?

- Standard “rule of thumb” $\text{Var}(\bar{y}_w) = \text{Var}(\bar{y}_u)(1 + \text{cv}(w))$
fails to reflect that nonresponse weighting can reduce variance
- Little & Vartivarian (2005) propose refinements

Choice of Adjustment Cells

- To reduce bias, adjustment cell variable needs to be related to nonresponse M and outcome Y
- To reduce variance, adjustment cell variables needs to be related to outcome (otherwise weighting *increases* variance)
- Two methods for creating adjustment cells with multiple X 's are
 - Response propensity stratification, based on regression of M on X
 - Predictive mean stratification, based on regression of Y on X

Weight the response rates?

- Nonresponse weights are often computed using units weighted by their sampling weights $\{w_{1i}\}$

$$w_{2j}^{-1} = \left(\sum_{r_i=1, x_i=j} w_{1i} \right) / \left(\sum_{r_i=1, x_i=j} w_{1i} + \sum_{r_i=0, x_i=j} w_{1i} \right)$$

- Gives unbiased estimate of response rate in each adjustment cell defined by X
- Not correct from a prediction perspective

Arguments against weighting response weights

- Unnecessary if cells are created properly!
 - Adjustment cells should be homogeneous with respect to response propensity
 - In that case, weighting the nonresponse rates is unnecessary and adds variance to estimates
- Doesn't work if cells are created improperly!
 - If adjustment cells are not homogeneous with respect to response propensity, then weighting RR's does not yield unbiased estimates survey estimands.
- The right approach is to create adjustment cells based on classification of the observed variables and the survey design variables.
 - Then weighting is unnecessary

Simulation Study

- Simulations in Little & Vartivarian (2003 Statistics in Medicine) consider the variance and bias of estimators of weighted and unweighted rates and alternative estimators, under a variety of population structures and nonresponse mechanisms.
- Binary outcome Y , stratum Z , adjustment cell X to avoid distributional assumptions such as normality.
- 25 populations to cover the factor space

Simulation Results

- ML for the model used to generate the data is always best or close to best.
- Unweighted-RR(x, z) is best overall: form cells based on X and Z
- Additive model theoretically biased when the data-generating model includes XZ interaction, but in these simulations the bias is modest.
- Unweighted-RR (x) is biased when both Y and R depend on Z .
- Weighted RR (x) does not generally correct the bias in these situations: similar to Unweighted-RR(x) overall

Remarks

- Don't weight response rates! Rather
 - Condition on design variables when creating adjustment cells
- Too many strata? Response propensity stratification (see next slide)
- To improve efficiency: weight shrinkage by multilevel modeling

Stratified Analysis: an alternative to weighting

Response propensity stratification

- X = covariates observed for respondents and nonrespondents, Y missing
- M = missing-data indicator
 - nonrespondent = 1, respondent = 0
- (A) Regress M on X (probit or logistic), using respondent and nonrespondent data $\hat{p}(M = 0 | X) = \text{propensity score}$
- (B1) Weight respondents by inverse of propensity score from (A), $1 / \hat{p}(M = 0 | X)$, or:
- (B2) form adjustments cells by categorizing $1 / \hat{p}(M = 0 | X)$
- Note that this method is only effective if propensity is also related to the outcome

X_1	X_2	...	X_p	Y	M
Complete cases					0
					0
					0
					?
					1
					1
					?
					1
					1

Predictive mean stratification

- X = covariates observed for respondents and nonrespondents
- Y = outcome with missing data
- (A) Regress Y on X (linear, other as appropriate) using respondent data only
- (B) Form adjustment cells with similar values of predictions from the regression $\hat{Y}(X)$
- This method has potential to reduce both bias and variance
- Note that the adjustment cells depend on outcome, so this method yields different cells for each outcome, hence is more complex with many outcomes with missing values

X_1	X_2	...	X_p	Y	M
Complete cases					0
					0
					0
				?	1
				?	1
				?	1

Construction of stratification scores

- One should think about missing data at the design stage and collect information predictive of participation and key variables of interest
- Common sources for data
 - Administrative data from which the sample has been drawn
 - Medicare enrollment files
 - Driver license file
 - A large study may be used to sample subjects for a supplement or a smaller study.; Two-stage, nested case-control study etc.

- Data collected during the study conduct phase (“Paradata”)
 - Statements made by sampled subjects to the interviewer
 - Interviewer observations
- Data at Neighborhood level
 - Block or Block-group level characteristics
- Example
 - Assets and Health Dynamics (AHEAD) Study
 - Primary outcome variable: 5 point Self-reported health status
 - Interviewer noted the following four variables
 - Time delay statements
 - Negative statements about participation in study
 - Positive statements
 - Statements about being Old to participate etc

- Subject level variable
 - Age and Sex
- Neighborhood level variables
 - Large urban area (yes/no)
 - Barriers to contact (yes/no)
 - Block: Persons per sq-mile
 - Block : % persons 70 or older
 - Block: % Minority populations
 - Block: % Multi-unit structures (10+)
 - Block: % Occupied Housing Units
 - Block: % Single person Hus
 - Block: % vacant Hus
 - Block: Persons per occupied Hu
- Sample size: n=10,173, respondents=8,212 (80.7%)

Comparison of “Paradata” between respondents and nonrespondents

Variable	Respondent	Nonrespondent	Effect Size
Age	77.41 (6.81)	77.83 (6.12)	0.065
Sex (% Female)	63.3%	63.6%	0.006
Mention age or illness	10.2%	16.1%	0.175
Negative Statement	9.5%	37.7%	0.703
Positive Statement	11.6%	2%	0.388
Time delay statement	8.1%	13.8%	0.183

Effect size

$$D = \frac{|\bar{x}_R - \bar{x}_{NR}|}{\sqrt{(s_R^2 + s_{NR}^2) / 2}}$$

For proportions, $s^2 = p(1 - p)$

Small : $D \leq 0.25$

Medium : $0.25 < D \leq 0.5$

Large : $0.5 < D \leq 0.75$

Very Large : $D > 0.75$

Comparison of Neighborhood level data between respondents and nonrespondents

Variable	Respondents	Nonrespondents	Effect size
% from Large urban areas	15.4%	22.7%	0.19
% with Barriers to contact	12.7%	20.0%	0.20
Block: Persons per sq-mile	6162.20 (14318.64)	7750.92 (15086.18)	0.11
Block : % persons 70 or older	10.98 (8.18)	11.16 (7.96)	0.02
Block: % Minority populations	22.76 (30.16)	24.60 (31.94)	0.06
Block: % Multi-unit structures (10+)	10.96 (20.24)	12.65 (21.74)	0.08
Block: % Occupied Housing Units	33.50 (23.76)	35.43 (24.68)	0.08
Block: % Single person HUs	24.48 (11.54)	25.17 (11.75)	0.06
Block: % vacant HUs	9.05 (9.18)	8.82 (8.90)	0.03
Block: Persons per occupied Hu	2.64 (0.47)	2.62 (0.48)	0.03

Building a Response Propensity Model

- Goal is to obtain a well fitting (logistic or probit) model that balances the covariates between respondents and nonrespondents

$$e(x) = \text{Propensity score}$$

$$X \perp R \mid e(x)$$

- One useful measure of goodness of fit: Hosmer-Lemeshow test
 - Create deciles based on the estimated propensity score , compare the observed and expected frequencies /counts in these 10 classes
 - Chisquare statistics with 9 degrees of freedom as a measure of lack of fit

- Checking the balancing property
 - Once satisfied with the logistic or probit model create classes (or strata or groups) based on quartiles or quintiles or deciles (depending upon your sample size) of the estimated propensity score.
 - In each class compute the effect size as we did for the overall sample.
 - Good balancing is indicated by the effect size being very small in each class
- Developing the propensity score model and checking its balancing property is an iterative process

- One can start with the main effect model and check the Hosmer-Lemeshow chisquare statistic
- Add two factor interactions to reduce the value of chisquare statistic
- Use stepwise or other selection procedure to reduce the model complexity
- Add higher order interactions if necessary
- Transform the covariates or standardized the continuous covariates (i.e. subtract the overall mean and divide by the overall standard deviation)

AHEAD example

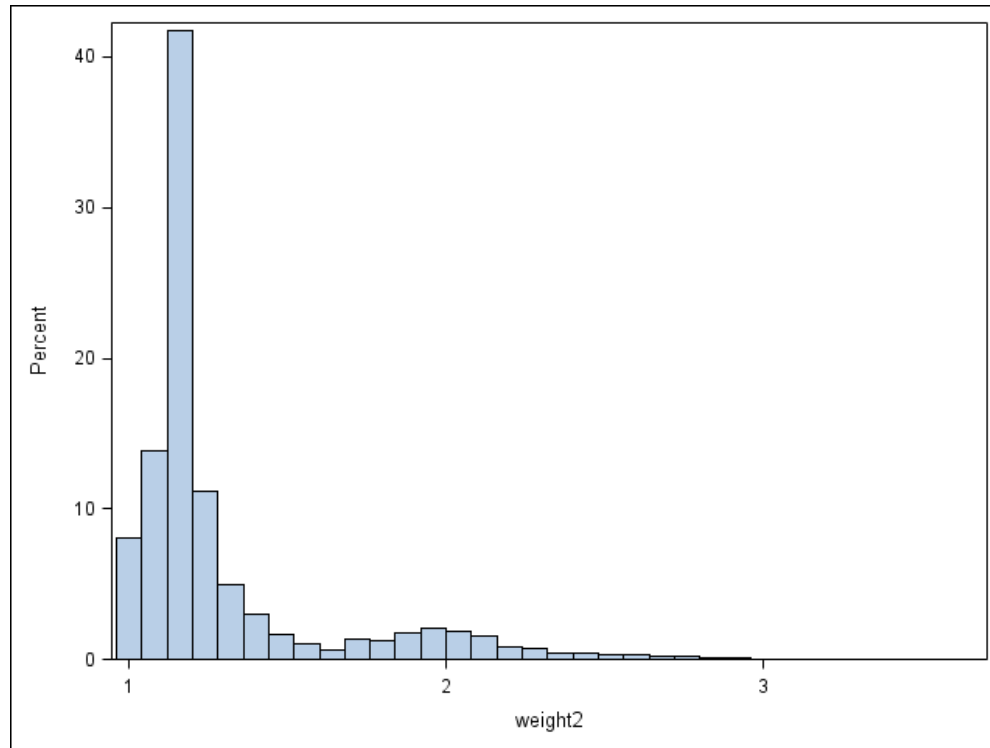
- First model with 16 main effects results in Hosmer-Lemeshow chisquare statistic of 28.68 with the p-value of 0.0004
- Next tried the model with all two factor interactions which reduced the chisquare statistic as 17.33 with the p-value of 0.03.
- As is typical, the full interaction model may be overmatching and result in poor fit.
- Next tried stepwise selection with different entry and exit probabilities. Finally with 0.5 as the probability for both entry and exit, obtained the model with Chisquare statistic of 4.75 with the p-value 0.784

Checking the Balance

- Created 4 groups (strata or classes) based on the quartiles of the estimated propensity squares.
- Computed the effect sizes for the 16 variables in each class. All were smaller than 0.05 indicating good balance.
- Propensity score stratification weights

Group	Respondents	Nonrespondents	Weight
1	1505	1037	1.6890
2	2100	445	1.2119
3	2237	306	1.1368
4	2370	173	1.0730

- Inverse probability weighting :
 $\text{weight} = 1 / \text{estimated propensity score}$
- Histogram of the nonresponse adjustment weight

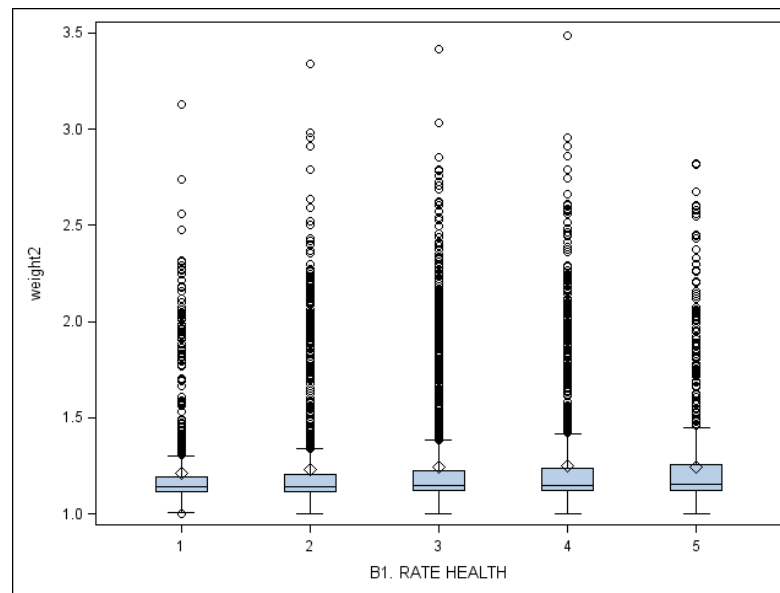


Weighted and Unweighted frequencies

Self-rated health	Unweighted	Propensity score Class weighted	inverse probability weighted
1	10.75	10.51 (0.34)	10.52 (0.34)
2	22.80	22.53 (0.47)	22.61 (0.47)
3	30.35	30.41 (0.52)	30.50 (0.52)
4	23.08	23.29 (0.48)	23.31 (0.48)
5	13.03	13.26 (0.38)	13.06 (0.38)

Weighted and Unweighted analysis

- The effect of weighting will depend upon the correlation between weight and the survey variable of interest



- Weighting is not that effective

Proper Inference from Weighted/ Stratified Data

- Role of weights in analytical inference (regression, factor analysis, ...) is controversial
- Can use packages for computing standard errors for complex sample designs -- but often these do not take into account sampling uncertainty in weights
- Bootstrap/Jackknife of weighting procedure propagates uncertainty in weights – but weights need to be recalculated on each BS/JK sample

Weight calibration using external information: Post-stratification

- Weight respondents to match distribution of a categorical variable with distribution known from external data (for example the census)

	Age post-stratum			
	20-29	30-39	40-49	>50
Respondents	m_1	m_2	m_3	m_4
Population	N_1	N_2	N_3	N_4

Assign **respondents** in **Age category j** the weight:

$$w_j = C \times N_j / m_j$$

C chosen so that weights sum to number of respondents

Post-stratification (continued)

- Often post-stratification is a final step after other weighting adjustments
- Corrects for bias from differential nonresponse across post-strata
- Useful post-stratification variables are predictive of both nonresponse and survey outcomes
- Statistician does not control the counts m_j
 - If these are too noisy the weights may need smoothing
- *Raking* extends method to cases where more than one margin is available

Post-stratification Example

- Behavior Risk Factor Surveillance System funded by CDC collects data from a large number of subjects from each State.
- The subjects are sampled through a Random Digit Dial
- Each state is responsible for collecting the data and report to CDC
- One of the question asks whether a respondent has any kind of private or public insurance.
- Data from Michigan 2011.

Raw Data on Respondents

Age Gender	18-24	25-34	35-44	45-54	55-64	65+
Male	170 (29%)	181 (27%)	371 (19%)	652 (19%)	871 (12%)	1130 (2%)
Female	144 (23%)	318 (20%)	610 (17%)	1012 (13%)	1342 (13%)	2062 (2%)

The numbers in the cells are sample sizes and proportion of people who reported not having any private or public insurance

Based on this raw data, the proportion of people reporting having no insurance in Michigan is 10.91%. This is potentially underestimate as there is severe under representation of younger people where large percent report not having insurance.

Population Size and Post-stratification Weights

Age Gender	18-24	25-34	35-44	45-54	55-64	65+
Male	494,025	580,833	633,321	743,803	608,637	587,184
Female	479,864	583,316	644,653	766,230	643,360	774,346

Age Gender	18-24	25-34	35-44	45-54	55-64	65+
Male	2906	3209	1707	1141	699	520
Female	3332	1834	1057	757	479	476

Weighted Estimate: 21.12%

Raking

- In the previous example, the population count was available for each age category by gender cell. What if only marginal distributions of age category and gender are available?
- Simple Example. The numbers in the cross classified table are the sample sizes

Sex Race	Male	Female	Population Total
White	10	10	1000
Non-White	10	10	300
Population total	1050	250	1300

- Rake on Race (rows)

Sex Race	Male	Female	Population Total
White	$10/20 * 1000 = 500$	$10/20 * 1000 = 500$	1000
Non-White	$10/20 * 300 = 150$	$10/20 * 300 = 150$	300
Population total	1050	250	1300

- Rake on Column (columns)

Sex Race	Male	Female	Population Total
White	$500/650 * 1050 = 807.7$	$500/650 * 250 = 192.3$	1000
Non-White	$150/650 * 1050 = 242.3$	$150/650 * 250 = 57.7$	300
Population total	1050	250	1300

Raking software

- Generally, the process will be iterated until the cell weights converge to a stable numbers
- CALMAR is a SAS macro (written by French statisticians and so the comments are in French)
- RAKING.SAS is a macro written by statisticians from Abt associates
- Both programs with documentation and examples is available on the internet

Summary of Weighting Methods

- Weighting is a relatively simple device for reducing bias from complete-case analysis
- Same weight for all variables -- simple, but better methods tune adjustment according to outcome
- No built in control of variance
 - ad-hoc trimming is common in surveys
- Less useful when:
 - Covariate information is extensive
 - Pattern of missing-data is non-monotone

Available-case (AC) analysis

- For univariate analyses, include all units where that variable is present
- Method uses all the available values
- Sample changes from variable to variable according to the pattern of missing data -- creates problems of comparability across variables if missingness is not MCAR.
- Likelihood approach is correct under MAR and correctly specified models.

Pairwise AC for covariance matrix

- Means and variances from set of cases observed for each variable observed
- Covariances from set of cases I_{jk} observed for both variables in the pair:

$$s_{jk}^{(jk)} = \sum_{i \in I_{jk}} (y_{ij} - \bar{y}_j^{(jk)})(y_{ik} - \bar{y}_k^{(jk)}) / (n^{(jk)} - 1)$$

- Issues:
 - ad-hoc: why not pairwise correlations rather than covariances?
 - can yield cov matrix that is not positive definite
 - Simulations and theory suggest the method works better when data are close to MCAR and correlations are low rather than high – but prefer methods that work in a variety of situations, e.g. ML for normal data

Summary

- CC analysis is simple, and works best when there is not much information in the incomplete cases
 - Weighting: can reduce bias for means, but potentially inefficient
- AC analysis: can recover more information than CC, but ad-hoc and does not work in all situations
- ML/Bayes: more principled, assumptions underlying the methods are explicit in the underlying model