Missing Data: Theory & Methods

Chapter 1
Introduction

Lu Mao lmao@biostat.wisc.edu

Introduction

- Statistical analysis with missing data is a rich and important field owing to the following two facts:
  1. Missing data are present in almost all practical situations as a result of incomplete measurement, subject loss to follow-up, survey non-response, etc..
  2. Many statistical problems with unobserved "latent variables", e.g., random effects models, causal inference under the counter-factual framework, etc., can be formulated into missing data problems.
- Naive approaches such as ignoring observations with missing elements may lead to invalid inference and loss of statistical efficiency.
- Need specialized methods for missing data.

Contents

Introduction

Example 1. Six-Cities data

- Consider the data from the Six Cities longitudinal study of the health effects of respiratory function in children (Ware et al., 1984). This is a well known environmental dataset that has been analyzed extensively in the literature.
- The binary response is the wheezing status (no wheeze, wheeze) of a child at age 11.
- The wheezing status is modeled as a function of the city of residence (x_1) and smoking status of the mother (x_2).
- The covariate x_1 is a binary covariate which equals 1 if the child lived in Kingston-Harriman, Tennessee, the more polluted city, and 0 if the child lived in Portage, Wisconsin.

Introduction

Example 1. Six-Cities data

- The covariate x_2 is maternal cigarette smoking measured in number of cigarettes per day.
- There are $n = 2394$ subjects in the dataset. The covariate x_1 is missing for 32.8% of the cases, and x_2 is missing for 3.3% of the cases.

Table 1.1: Summary of the Six-Cities Data

y		x_1		x_2	
0	$N = 1827$(76.3%)	0	$N = 862$(36.0%)	Obs'ved	mean 7.2 (s.d. 11.3)
1	$N = 567$(23.7%)	1	$N = 747$(31.2%)	NA	$N = 79$(3.3%)
		NA	$N = 785(32.8\%)$		

Introduction

Example 2. Liver cancer data

- Consider data on $n = 191$ patients from two Eastern Cooperative Oncology Group clinical trials, EST 2282 (Falkson et al., 1990) and EST 1286 (Falkson et al., 1994).
- Here, we are primarily interested in the patient's status as he/she enters the trials.
- In particular, we are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease in weeks (x_1), two biochemical markers (each classified as normal or abnormal): Alpha fetoprotein (x_2), and Anti Hepatitis B antigen (x_3); associated jaundice (yes, no) (x_4), body mass index (x_5) (defined as weight in kilograms divided by the square of height in meters), and age in years (x_6).

Example 2. Liver cancer data

- Table 1.2 shows that 28.8% of the patients have at least one covariate missing. The biochemical marker Anti-hepatitis B antigen, which is not easy to obtain, has the highest proportion missing.

Table 1.2: Missingness summary of the liver cancer data

Variable	Missing $N(\%)$
Time Since Diagnosis	17 (8.9%)
Alpha Fetoprotein	11 (5.8%)
Anti Hepatitis B	35 (18.3%)
Overall	55 (28.8%)

Example 3. Missing Quality of Life Data in Longitudinal Studies

- E1694 was a two arm phase III clinical trial comparing IFN to vaccine (GMK) in high-risk melanoma patients.
- QOL data was collected in this study. There were a total of 364 patients who participated in the QOL portion of this study.
- 54 cases were removed due to death before four QOL measurements could be taken.
- It is highly likely that patients who die within one year of starting treatment have significantly different QOL than patients who survive beyond one year. Therefore, none of the missingness is due to death.

Example 3. Missing Quality of Life Data in Longitudinal Studies

- We also removed 33 cases that had all four QOL measurements missing, so there are 277 observations in the data set, and the total QOL score is missing at least once for 118 of them (42.6%). The total fraction of missing QOL data is 19.0%.

Table 1.3: E1694 Patterns of Missingness

Number of missing QOL measurements	$N(\%)$
0	159 (57.4%)
1	59 (21.3%)
2	25 (9.0%)
3	34 (12.3%)

Example 3. Missing Quality of Life Data in Longitudinal Studies

- The covariates of interest include an indicator variable for treatment (HDI vs GMK), sex (0 for female and 1 for male), age, ulceration of the primary tumor (0 for no and 1 for yes), and a dichotomous variable for Breslow thickness of the primary tumor (0 for $<$ 3.00 mm and 1 for $\geq$ 3.00 mm).
- Ulceration is missing for 56 cases (20.2%) and Breslow thickness is missing for 49 cases (17.7%). Overall, 163 cases (58.8%) have either a missing longitudinal outcome and/or a missing baseline covariate.

Contents

Taxonomy of Missing Data

- Terminology
  - **Full data**: $Y = (Y_{\text{obs}}, Y_{\text{mis}})$
  - **Observed data**: Y_{obs}
  - **Missing data**: Y_{mis}
- Denote R as the missing data indicator ($R = 1$ if Y is observed and $R = 0$ if Y_{obs} is observed)
- Denote $p_Y(y; \theta)$ as the density of Y parameterized by θ. Suppose θ is the target of inference.
- With full data, inference on θ can be based on the likelihood $p_Y(y; \theta)$; with missing data, this is not possible.

Taxonomy of Missing Data

- Denote
$$\pi(y) = \Pr(R = 1|Y = y),$$
which describes the **missing data mechanism** given the full data, and let $\overline{\pi}(y) = 1 - \pi(y)$.
- If Y_{mis} is missing, that means $R = 0$ and Y_{obs} is observed. So the likelihood for the observed data $(R = 0, Y_{\text{obs}})$ is
$$\int \overline{\pi}(y) p_Y(y; \theta) d\nu(y_{\text{mis}}) \tag{1.1}$$
where $y = (y_{\text{obs}}, y_{\text{mis}})$ and ν is some dominating measure for Y_{mis}.
- Proper inference on θ hinges on the missing data mechanism π.

Taxonomy of Missing Data

- Classifications of missing data based on missing mechanisms
  - **Missing Completely At Random (MCAR)**: $R \perp\!\!\!\perp Y$; failure to observe a value does not depend on any data, either observed or missing.
  - **Missing At Random (MAR)**: $R \perp\!\!\!\perp Y_{\text{mis}}|Y_{\text{obs}}$; failure to observe a value does not depend on the unobserved value given the observed ones.
  - **Not Missing At Random (NMAR)**: failure to observe a value depends on the value that could have been observed (even given the observed ones).

Taxonomy of Missing Data

Missing Completely At Random (MCAR): $R \perp\!\!\!\perp Y$

- $\pi(y) = \pi_0$
- Observed data likelihood

$$(1.1) = (1 - \pi_0) p_{\text{obs}}(y_{\text{obs}}; \theta) \propto p_{\text{obs}}(y_{\text{obs}}; \theta),$$

  where $p_{\text{obs}}(y_{\text{obs}}; \theta) = \int p_Y(y; \theta) d\nu(y_{\text{mis}})$ is the marginal density for Y_{obs}.
- Examples include lost data, patient moves away, laboratory instrument accidentally breaks, or data management error. It is like flipping a coin to determine the probability of missingness.

Taxonomy of Missing Data

Missing Completely At Random (MCAR): $R \perp\!\!\!\perp Y$

- Suppose we have a random sample of n subjects. Call those with fully observed Y_i ($R_i = 1$) the **complete cases**.
- A complete-case (CC) analysis amounts to analyzing the complete-cases as if they are random sample of full data, i.e., discarding the cases with incomplete data ($R_i = 0$).
- For instance, a CC analysis using the maximum likelihood estimation (MLE) is based on the CC log-likelihood

$$\sum_{i=1}^{n} I(R_i = 1) \log p_Y(Y_i; \theta).$$

Taxonomy of Missing Data

Missing Completely At Random (MCAR): $R \perp\!\!\!\perp Y$

- CC analysis is perhaps the easiest thing to do with missing data, especially when the missing proportions are small. It is the default implementation in most statistical packages.
- Under MCAR, the complete cases are indeed a random sample. Therefore, the CC analysis is valid, though statistically inefficient as a result of tossing the information contained in the incomplete cases.
- To gather all information contained in the observed sample, we can use MLE based on all observed data with log-likelihood

$$\sum_{i=1}^{n} I(R_i = 1) \log p_Y(Y_i; \theta) + I(R_i = 0) \log p_{\text{obs}}(Y_{\text{obs},i}; \theta).$$

- But $p_{\text{obs}}(y_{\text{obs}}; \theta)$ may not have a simple or closed form.

Taxonomy of Missing Data

Missing At Random (MAR): $R \perp\!\!\!\perp Y_{\text{mis}} | Y_{\text{obs}}$

- $\pi(y) = \pi(y_{\text{obs}})$ (abusing the notation).
- Observed data likelihood

$$\begin{aligned}(1.1) &= \int \overline{\pi}(y_{\text{obs}}) p_Y(y; \theta) d\nu(y_{\text{mis}}) \\ &= \overline{\pi}(y_{\text{obs}}) p_{\text{obs}}(y_{\text{obs}}; \theta) \\ &\propto p_{\text{obs}}(y_{\text{obs}}; \theta),\end{aligned}$$

where $\overline{\pi}(y_{\text{obs}}) = 1 - \pi(y_{\text{obs}})$ and the proportionality holds when $\pi(y_{\text{obs}})$ is not a function of θ (missing data mechanism is uninformative of the parameter of interest).

Taxonomy of Missing Data

Missing At Random (MAR): $R \perp\!\!\!\perp Y_{\text{mis}} | Y_{\text{obs}}$

- A popular interpretation of MAR is that missingness is allowed to depend on the observed data but not on the missing data. More precisely, MAR allows missingness to depend on the missing value *only through* the observed ones.
- In both MCAR and MAR, the missing data mechanism can be ignored in making inferences about the parameters of the sampling model. They are hence called ignorably missing.
- MAR is a more realistic assumption than MCAR. But under MAR, a CC analysis is generally not valid, since the complete cases need not be a representative sample from the population.

Taxonomy of Missing Data

An example of CC analysis invalid under MAR

- Let $Y_1 \sim N(\mu + \theta, 1)$, $Y_2 \sim N(\mu, 1)$, and $Y_1 \perp\!\!\!\perp Y_2$. Suppose we want to make inference on θ.
- Suppose Y_2 is always observed and Y_1 is possibly missing (indicated by $R = 0$), with probabilities that depend on Y_2. This is a case of MAR.
- With a random sample of $(R_i, Y_{1i}, R_i Y_{2i})$ $(i = 1, \cdots, n)$, the CC analysis based on the MLE is

$$\widehat{\theta}_n^{CC} = \frac{\sum_{i=1}^n R_i Y_{1i}}{\sum_{i=1}^n R_i} - \frac{\sum_{i=1}^n R_i Y_{2i}}{\sum_{i=1}^n R_i}.$$

- Show that $\widehat{\theta}_n^{CC}$ is consistent for $\mu + \theta - E(Y_2|R = 1)$
- $E(Y_2|R = 1)$ need not be equal to μ.

Taxonomy of Missing Data

Missing At Random (MAR): $R \perp\!\!\!\perp Y_{\text{mis}} | Y_{\text{obs}}$

- These are special cases when CC analysis for certain parameters is valid even under MAR.
- In a regression model, for instance, let $p(y|x;\theta)$ be the conditional density of Y given X, where θ is the regression parameter.
- The full data likelihood is
$$p(y|x;\theta)\eta(x),$$
where η is the density of X.
- When the response Y is missing, with probabilities possibly dependent on X, the observed data likelihood is
$$\int p(y|x;\theta)\eta(x)dy = \eta(x),$$
which has nothing to do with the regression parameter.
- All information about θ is contained in the complete cases.

Taxonomy of Missing Data

Not Missing At Random (NMAR):

- $\pi(y)$ is a function of both y_{obs} and y_{mis}. So, the observed data likelihood (1.1) cannot be further reduced.
- Under NMAR, the failure to observe a value depends on the value that would have been observed.
- NMAR is the most general situation. Examples of NMAR include longitudinal studies measuring QOL, where study dropout often depends on how sick the patient is. Also, in survey studies, non-response may arise from the respondent's reluctance to disclose a particular choice or characteristic of his/hers due to, e.g., fear of social stigma.

Taxonomy of Missing Data

Not Missing At Random (NMAR):

- Valid inferences generally require specifying the correct model for the missing data mechanism. The resulting estimators and tests are typically sensitive to these (unverifiable) assumptions.
- Because the missingness mechanism under NMAR cannot be ignored, it is also called **non-ignorable missingness**.
- The difficulty with NMAR data is inherently associated with the issue of identifiability.

Contents

Identifiability with Missing Data

- Generally, a parameter is said to be identifiable with the data if it uniquely indexes their distribution, that is, one distribution of the data corresponds to one and only one value of the parameter.
- In other words, there do not exist two parameters that give rise to the same distribution.
- Mathematically, let $p_Y(y;\theta)$ be the model for the density function of Y indexed by θ. Then, θ is identifiable if

$$p_Y(y;\theta_1) = p_Y(y;\theta_2), \quad \forall y \ \ a.e.$$

implies $\theta_1 = \theta_2$.

Identifiability with Missing Data

- The importance of identifiability is evident. Since we can only infer about the distribution from the data, if the parameter is not uniquely linked to the distribution, the obtained information about the distribution cannot be carried on to the parameter. So the problem of making inference on the parameter would be ill posed.
- A simple example of unidentifiable parameters is the over-parameterized model

$$Y \sim N(\mu_1 + \mu_2, 1), \ \ \mu_1, \mu_2 \in \mathbb{R}.$$

All pairs of μ_1 and μ_2 would have given rise to the same distribution of Y as long as their sum is the same.

Identifiability with Missing Data

- Identifiability in the presence of missing data is a very importance problem, and sometimes a very hard one.
- The identifiability problem with missing data can be decomposed into two layers. One is the identifiability with the full data, i.e., the model $p_Y(y;\theta)$. The other is the question whether coarsening of the data (from Y to Y_{obs}) incurs extra non-identifiability issue.
- The first layer has nothing to with missing data *per se* and needs to be examined on a case-by-case basis.
- We are interested in the second layer. In particular, we want to know under what general circumstances there is no extra non-identifiability with the coarsened data.

Identifiability with Missing Data

- Under the assumption of MAR, if there is a positive probability of observing the full data given any values of the observed data, then identifiability with coarsened data is the same as that with full data.

Proposition 1.1 (Identifiability under MAR)

Under MAR, denote $\pi(Y_{obs}) = \Pr(R = 1|Y)$, *and suppose*

$$\pi(Y_{obs}) > 0 \quad a.s.. \tag{1.2}$$

If θ *is identifiable with the full data* Y *with density* $p_Y(y|\theta)$, *then it is identifiable with the coarsened data* (R, Y_{obs}, RY_{mis}).

Identifiability with Missing Data

- To see why the assumption of MAR and the positivity condition (1.2) would help preserve identifiability from full to coarsened data, it is useful to think of the whole population as divided into subpopulations based on the values of the observed data.
- Let $y_{\text{obs}}^{(1)}, y_{\text{obs}}^{(2)}, y_{\text{obs}}^{(3)}, \cdots$, be the possible values of Y_{obs}. The kth subpopulation consists of subjects with $Y_{\text{obs}} = y_{\text{obs}}^{(k)}$.
- Because of the MAR assumption, the missingness mechanism within each subpopulation is completely random. Because of the positivity condition (1.2), there is always a probabilistically (positive) fraction of complete cases in each subpopulation.
- The lost information contained in the incomplete cases can thus be inferred from complete cases, so the composition of each subpopulation can be reconstructed.

Identifiability with Missing Data

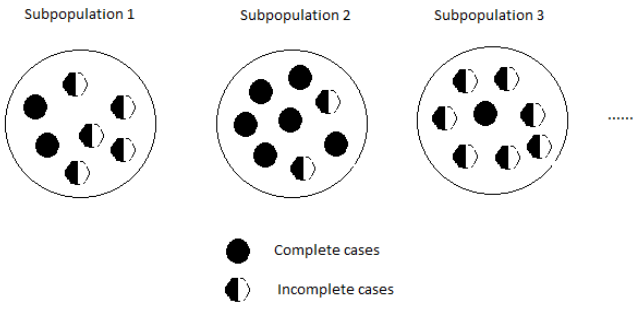

Figure 1.1: An illustration of identifiability under MAR. Black indicates observed; hollow indicates missing. The information contained in the right halves of the half-filled circles can be inferred based on the fully-filled circles, because the latter are a representative sample of the subpopulation.

Identifiability with Missing Data

- Because the composition of the whole population by each of the subpopulation is certainly observable (because the partition is based on the *observed* data), all aspects about the composition of the whole population are preserved.
- Therefore, the coarsened data are as good as the full data in terms of identifiability.

Identifiability with Missing Data

- A formal proof of Proposition 1.1 can go along the following lines.
- Under MAR, the density of $(R, Y_{\text{obs}}, RY_{\text{mis}})$ is

$$f(R, Y_{\text{obs}}, RY_{\text{mis}}; \theta, \pi) = \{\pi(Y_{\text{obs}}) p_Y(Y; \theta)\}^R \{\overline{\pi}(Y_{\text{obs}}) p_{\text{obs}}(Y_{\text{obs}})\}^{1-R}.$$

- First note that π is identifiable (since it pertains to a conditional distribution of the observed data).
- Given θ_1 and θ_2, set
  $f(R = 1, Y_{\text{obs}}, Y_{\text{mis}}; \theta_1, \pi) = f(R = 1, Y_{\text{obs}}, Y_{\text{mis}}; \theta_2, \pi)$. We have

$$\pi(Y_{\text{obs}}) p_Y(Y; \theta_1) = \pi(Y_{\text{obs}}) p_Y(Y; \theta_2).$$

  Use the positivity condition (1.2) to cancel out $\pi(Y_{\text{obs}})$. The result follows from the identifiability of θ in $p_Y(Y; \theta)$.

Identifiability with Missing Data

Non-identifiability under NMAR

- On the other hand, NMAR does not preserve identifiability. Because without the assumption that the complete cases are a representative sample of the (sub)population, information contained in the missing values cannot be inferred from the observed ones and is thus irredeemably lost.
- So under NMAR, some aspects of the distribution of full data will become unidentifiable.
- More importantly, whether the missing mechanism is MAR or NMAR cannot be identified from the data. That is, for a MAR situation, there exists an NMAR situation that could have given rise to the same observed data as the MAR one.

Identifiability with Missing Data

Non-identifiability under NMAR

- Use the Subpopulation 1 in Figure 1.1 as an illustration (See Figure 1.2).
- It could be that all the half-filled circles are all black, like the observed ones, and the missing pattern is completely at random with probability $5/7$ (a scenario of MAR).
- But it also could be that the half-filled circles had all sorts of different colors, and were set to missing if they were non-black (a scenario of NMAR).
- Both situations could have generated the observed data. Neither is more or less plausible than the other per the observed data.

Identifiability with Missing Data

Non-identifiability under NMAR

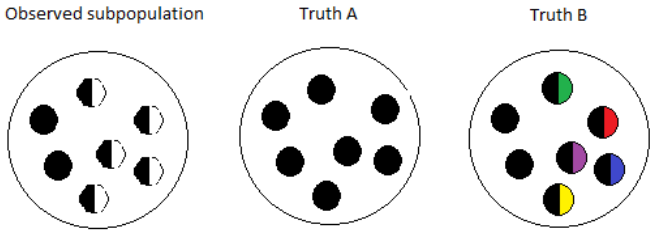

Figure 1.2: An illustration of non-identifiability under NMAR. The observed subpopulation could have been generated by Truth A combined with a MAR mechanism with missing probability 5/7. It could also have been generated by Truth B combined with a (deterministic) missing mechanism that the right half of the circle is set missing if it is non-black. It cannot be told from the observed data which situation is more plausible.

Identifiability with Missing Data

Non-identifiability under NMAR

- Here is a simple example for the non-identifiability under NMAR.
- Let $Y \sim$ Binomial$(1, \theta)$ and suppose we have an iid sample of Y except that some observations are missing (indicated by $R = 0$). The interest is in making inference on θ.
- Denote $p_{yr} = \Pr(Y = y, R = r)$ $y = 1, 0$, $r = 1, 0$.

Table 1.4: Binomial distribution with missing values.

		R	
		1	0
Y	1	p_{11}	p_{10}
	0	p_{01}	p_{00}

Identifiability with Missing Data

Non-identifiability under NMAR

- We can only observe the Y_i with $R_i = 1$, that is, the counts in the two cells of the left column in Table 1.4. So without any assumptions on the missing mechanism, we can only identify (gain information about) p_{11} and p_{01} from the observed data.
- Since $\theta = p_{11} + p_{10}$ and p_{10} is not identifiable, θ is no identifiable.

Exercise 1.1

1. *Given the distribution of the observed data, i.e., fixing up p_{11} and p_{01}, specify the range of possible θ.*
2. *For each possible θ in that range, express the missing data mechanism $\pi_y := \Pr(R = 1|Y = y), y = 1, 0$, in terms of p_{11}, p_{01}, and θ.*
3. *Under MAR, show that θ is identifiable by expressing it as an explicit function of p_{11} and p_{01}.*

Identifiability with Missing Data

Non-identifiability under NMAR

- All (non-)identifiability talked about so far is nonparametric identifiability.
- Under NMAR, when parametric models are specified for the missing data mechanism (selection models) and for the sampling distribution, the parameters may be identifiable. So may the MAR assumption.
- For example, let $Y \sim N(\mu, \sigma^2)$ and assume a logistic selection model

$$\Pr(R = 1|Y = y) = \frac{e^{\psi_0+\psi_1 y}}{1 + e^{\psi_0+\psi_1 y}}. \tag{1.3}$$

- One can show that the parameters $(\psi_0, \psi_1, \mu, \sigma^2)$ are identifiable under (1.3). Hence, the MAR assumption is identifiable (why?).

Identifiability with Missing Data

Non-identifiability under NMAR

- In fact, though the MAR assumption is not testable nonparametrically, one can posit parametric selection models in which MAR is identifiable and testable.
- Thus to check the MAR assumption, one can compare the analysis results under MAR and under NMAR with the parametric selection model and see how things differ. This is called *sensitivity analysis*.
- Marked difference suggests that the MAR assumption might be untenable.
- A lack of difference, however, does not verify the MAR assumption. It just means that the assumption is not sensitive to that particular selection model.

Identifiability with Missing Data

Non-identifiability under NMAR

Exercise 1.2

With the selection model (1.3), show that the parameters are identifiable with the observed data likelihood

$$\Pr(Y=y,R=r)=\left\{\frac{e^{\psi_0+\psi_1 y}}{1+e^{\psi_0+\psi_1 y}}\sigma^{-1}\phi\left(\frac{y-\mu}{\sigma}\right)\right\}^{r} \times\left\{\int\frac{1}{1+e^{\psi_0+\psi_1 y}}\sigma^{-1}\phi\left(\frac{y-\mu}{\sigma}\right)dy\right\}^{1-r},$$

where ϕ is the density of the standard normal distribution.

Identifiability with Missing Data

Non-identifiability under NMAR

Exercise 1.3

Without the selection model (1.3), show that the MAR assumption is not identifiable even with the normal assumption on Y by completing the following.

Given $\pi_0 := \Pr(R = 1|Y = y)$ under MAR and (μ, σ^2), find a non-constant function $\pi(y) \in [0, 1]$ and (μ_1, σ_1^2) such that

$$\pi(y)\sigma_1^{-1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) = \pi_0\sigma^{-1}\phi\left(\frac{y-\mu}{\sigma}\right), \forall y \in \mathbb{R}.$$

Contents

Overview of Statistical Methods

- We will discuss four common approaches to statistical inference with missing data. These are
  1. **Maximum Likelihood Estimation** (MLE) by the Expecation-Maximization (EM) algorithm
  2. **Multiple Imputation** (MI)
  3. **Fully Bayesian methods** (FB)
  4. **Weighted Estimating Equations** (WEE)
- The first three methods are based on likelihoods. WEE is based on estimating equations and is closely associated with semiparametric inference.
- The focus of this course is on MLE (via the EM) and WEE, and we will be mostly concerned with data that are MAR.

Overview of Statistical Methods

Maximum likelihood via the EM algorithm

- The Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is a general iterative algorithm that may be used to find MLEs in incomplete data problems.
- EM is most useful when maximization from the full data likelihood is straightforward while maximization based on the observed data likelihood is difficult.
- The basic idea of EM is to augment the data (likelihood) so that the observed data likelihood resembles a full data likelihood, so that it can be maximized using standard techniques.

Overview of Statistical Methods

Maximum likelihood via the EM algorithm

- Specifically, denote the full data of the whole sample as D, the observed part of the sample as D_{obs}, and the missing part of the sample as D_{mis}. In the previously used notation, $D = \{Y_i, i = 1, \cdots, n\}$ and $D_{\text{obs}} = \{(R_i, Y_{\text{obs},i}, R_i Y_{\text{mis},i})\}$.
- Let $l_n(\theta|D)$ denote the full data log-likelihood. The EM algorithm consists of an "E step" and an "M step". The M step is especially simple to describe since it uses whatever computational methods that are appropriate in the full data case.
- That is, the M step performs maximum likelihood estimation of θ using the "augmented" log-likelihood obtained from the E step. It treats this augmented log-likelihood as if it were a full data log-likelihood.

Overview of Statistical Methods

Maximum likelihood via the EM algorithm

- The E step computes the expected value of the full data log-likelihood given both the observed data and a current estimate of the parameters.
- E-Step: Let $\theta^{(t)}$ be the current estimate of the parameter θ. The E step computes

$$Q(\theta|\theta^{(t)}) := E\{l_n(\theta)|D_{\text{obs}}, \theta^{(t)}\}.$$

- Note that the conditional expectation is taken assuming $\theta^{(t)}$ is the "true" parameter.
- In the iid case with two-levels of missingness under MAR,

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \Big[R_i l(\theta|Y_i) + (1 - R_i) E\{l(\theta|Y_i)|Y_{\text{obs},i}, R_i = 0, \theta^{(t)}\} \Big],$$

where $l(\theta|y)$ is the log-likelihood for a single observation of full data Y.

Maximum likelihood via the EM algorithm

- Note that under MAR

$$E\{l(\theta|Y_i)|Y_{\text{obs},i}, R_i = 0, \theta^{(t)}\} = E\{l(\theta|Y_i)|Y_{\text{obs},i}, \theta^{(t)}\}.$$

  So that the E-step pertains only to the sampling distribution of Y and has nothing to do with the selection model.
- M-Step: The M step computes $\theta^{(t+1)}$ by maximizing the expected log-likelihood found in the E step:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}).$$

- These two steps are iterated until convergence.

Maximum likelihood via the EM algorithm

- Here is a toy example. Suppose the full data are iid $Y_i \sim N(\theta, 1)$, $i = 1, \cdots, n$. Further assume that we observe the first m of them, and the remaining $n - m$ observations are MCAR.
- The full data log-likelihood is, up to a constant,

$$l_n(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \theta)^2.$$

- We first do the M step. Set $\frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)}) = 0$. We have

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta}E[l_n(\theta)|D_{\text{obs}}, \theta^{(t)}] &= E\left[\frac{\partial}{\partial\theta}l_n(\theta)\Big|D_{\text{obs}}, \theta^{(t)}\right] \\ &= \sum_{i=1}^{n}(E[Y_i|D_{\text{obs}}, \theta^{(t)}] - \theta) \end{aligned}$$

Maximum likelihood via the EM algorithm

- So
$$\theta^{(t+1)} = n^{-1} \sum_{i=1}^{n} E[Y_i | D_{\text{obs}}, \theta^{(t)}].$$
- Now, the E step amounts to computing
$$E[Y_i | D_{\text{obs}}, \theta^{(t)}] = \begin{cases} Y_i, & i = 1, \cdots, m \\ \theta^{(t)}, & i = m+1, \cdots, n \end{cases}$$
- So the $(t+1)$th iteration is
$$\theta^{(t+1)} = \frac{\sum_{i=1}^{m} Y_i + (n-m)\theta^{(t)}}{n}.$$

Multiple imputation

- The technique of multiple imputation involves creating multiple "full" datasets by filling in values for the missing data. Then, each filled-in dataset is analyzed as if it were a full dataset.
- The inferences for the filled-in datasets are then combined into one result, by averaging over the filled-in datasets.
- The most popular way of doing MI is to sample from a posterior predictive distribution under the Bayesian framework.
- MI is a proper imputation technique in the sense that the uncertainty contained in the missing values are acknowledged by creating multiple full datasets.

Overview of Statistical Methods

Multiple imputation

- Improper imputation techniques involve ad-hoc ways of filling in the missing values, such as substituting the sample mean, fitted values, or other values.
- Some improper imputation techniques include: hot deck imputation, where recently recorded units in the sample are substituted for the unobserved values, mean imputation, where means from sets of recorded values are substituted; and regression imputation, where missing values for a subject are filled in by predicted values from the regression on the known variables for that subject.
- Proper imputation such as MI has a solid theory and leads to valid large sample inferences for the parameters, whereas improper imputation does not.

Overview of Statistical Methods

Multiple imputation

The basic idea behind MI is as follows:

1. Construct K "full" datasets by inserting in the missing values drawn from a Bayesian posterior predictive distribution;
2. Obtain $\widehat{\theta}^{(k)}$ for the kth imputed dataset, $k = 1, ..., K$.
3. The parameter estimate is $\widehat{\theta} = K^{-1} \sum_{k=1}^{K} \theta^{(k)}$.
4. To compute the variance estimate, let $\widehat{V}^{(k)}$ denote the variance estimate from the kth imputed full dataset, obtained by, e.g., the inverse information matrix.
5. Compute
   - Within imputation variation: $\overline{V} = K^{-1} \sum_{k=1}^{K} \widehat{V}^{(k)}$
   - Between imputation variation: $\widehat{B} = K^{-1} \sum_{k=1}^{K} (\widehat{\theta}^{(k)} - \widehat{\theta})^{\otimes 2}$

   where $a^{\otimes 2} = aa^{\mathrm{T}}$ for any vector a.
6. The variance of $\widehat{\theta}$ is given by

$$\widehat{V}^{MI} = \overline{V} + (1 + K^{-1})\widehat{B}.$$

Overview of Statistical Methods

Fully Bayesian methods

- Fully Bayesian methods for missing data involve specifying priors on all of the parameters. The missing values as well as the parameters are then sampled from their respective posterier distributions via the Gibbs sampler.
- FB with missing values only involves the incorporation of an extra layer in the Gibbs steps compared to the full data case.
- The fundamental reason for this conceptual simplicity is that the Bayesian framework sees no difference between data and parameter by treating both as random variables.
- Thus, Bayesian methods can easily accommodate missing data without requiring extra modeling assumptions or new techniques for inference.

Overview of Statistical Methods

Fully Bayesian methods

- To describe the basic framework, let $L(D|\theta)$ denote the full data likelihood, and let $q(\theta)$ denote prior for θ.
- Our goal is to make inferences with the posterior distribution of θ based on the observed data.
- To that end, we conduct the following steps iteratively by Gibbs sampling (we use $p(A|B)$ as a generic notation for the conditional distribution of A given B):
  1. Draw D_{mis} from $p(D_{\text{mis}}|\theta, D_{\text{obs}})$
  2. Draw θ from $p(\theta|D_{\text{obs}}, D_{\text{mis}})$, where D_{mis} is from the previous draw.

Fully Bayesian methods

- Both $p(D_{\text{mis}}|\theta, D_{\text{obs}})$ and $p(\theta|D)$ are proportional to

$$L(D|\theta)q(\theta).$$

  So, techniques such as Metropolis-Hastings algorithm can be used in each sampling step.
- By Gibbs sampling theory, after a period of "burn-in" iterations, the θs thus drawn eventually follow $p(\theta|D_{obs})$.

Weighted estimating equations

- The weighted estimating equation approach starts with some existing estimating function based on the full data, say, $m(Y;\theta)$. This estimating function is valid in the sense that $Em(Y;\theta_0) = 0$, where θ_0 is the true value of θ.
- An example of estimation functions is the score function in a parametric model.
- With full data, the estimating equation is

$$\sum_{i=1}^{n} m(Y_i;\theta) = 0,$$

  where the root $\widehat{\theta}_n$ can be calculated by the Newton-Raphson algorithm.

Weighted estimating equations

- By Z estimation theory, $\widehat{\theta}_n$ is consistent and asymptotically normal, whose variance can be estimated by

$$\left\{\sum_{i=1}^{n} \dot{m}(Y_i;\widehat{\theta}_n)\right\}^{-1} \sum_{i=1}^{n} m(Y_i;\widehat{\theta}_n)^{\otimes 2} \left\{\sum_{i=1}^{n} \dot{m}(Y_i;\widehat{\theta}_n)^{\mathrm{T}}\right\}^{-1},$$

  where $\dot{m}(y;\theta) = \frac{\partial}{\partial\theta} m(y;\theta)$ and $a^{\otimes 2} = aa^{\mathrm{T}}$ for any vector a.
- With incomplete data, applying the full-data estimation function to the complete cases (CC analysis) may lead to bias because the complete cases need not be a random sample of the population. That is to say, the estimating equation

$$\sum_{i=1}^{n} R_i m(Y_i;\theta) = 0$$

  is generally invalid unless the data are MCAR.

Weighted estimating equations

- To correct for the non-representativeness of the complete cases, each case is be inversely weighted by its selection probability.
- Assume MAR and let $\pi(Y_{\mathsf{obs}}) = \Pr(R = 1|Y)$. The inverse probability weighted (IPW) estimating equation is

$$\sum_{i=1}^{n} \frac{R_i}{\pi(Y_{\mathsf{obs},i})} m(Y_i;\theta) = 0.$$

- The IPW estimating function is valid because

$$\begin{aligned} E\left\{\frac{R}{\pi(Y_{\mathsf{obs}})} m(Y;\theta_0)\right\} &= E\left[\frac{E(R|Y)}{\pi(Y_{\mathsf{obs}})} m(Y;\theta_0)\right] \\ &= Em(Y;\theta_0) \\ &= 0. \end{aligned}$$

Overview of Statistical Methods

Weighted estimating equations

- The selection probability (propensity score) $\pi(Y_{\text{obs}})$ is typically unknown. In that case, a parametric model $\pi(Y_{\text{obs}};\psi)$ can be built.
- An estimator $\widehat{\psi}_n$ can be found by MLE using the data $(R_i, Y_{\text{obs},i}), i = 1, \cdots, n$.
- Then, the estimated selection probabilities from the parametric model are inserted into the IPW estimating equations:

$$\sum_{i=1}^{n} \frac{R_i}{\pi(Y_{\text{obs},i};\widehat{\psi}_n)} m(Y_i;\theta) = 0.$$

Overview of Statistical Methods

Weighted estimating equations

- The WEE approach is very useful in causal inference under the counter-factual framework.
- Suppose each subject could be subject to either treatment or control, indicated by $W = 1$ and 0, respectively. The outcome is denoted as $Y(w)$ had the subject been assigned to group w, $w = 0, 1$. So, each subject has two *potential* outcomes
- The average causal treatment effect is defined as

$$EY(1) - EY(0).$$

- However, only the outcome associated with the treatment group to which the subject is actually assigned is observed, i.e., $Y = WY(1) + (1 - W)Y(0)$. So, this is a missing data problem.

Overview of Statistical Methods

Weighted estimating equations

- As in any missing data problem, the missingness mechanism, or the treatment assignment mechanism, is very important to the inference.
- In completely randomized experiments, the difference of unweighted averages is a valid estimator fo the average causal treatment effect:

$$\frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} (1 - W_i) Y_i}{n - \sum_{i=1}^{n} W_i}.$$

- In observational studies, it is not realistic to assume that the assignment mechanism is completely random. Let Z denote a set of pre-treatment variables on which the treatment assignment may depend.

Overview of Statistical Methods

Weighted estimating equations

- We further make the standard assumption that the potential outcomes are independent of treatment assignment given the pre-treatment variables:

$$\{Y(1), Y(0)\} \perp\!\!\!\perp W | Z.$$

- This assumption basically says there is no unmeasured confounders for the relationship between potential outcomes and treatment assignment. It corresponds to MAR in missing data terminology.

Weighted estimating equations

- Similar to the general missing data case, we can use the following IPW estimator for the average causal treatment effect:

$$n^{-1}\sum_{i=1}^{n}\frac{W_iY_i}{\pi(Z_i;\psi)} - n^{-1}\sum_{i=1}^{n}\frac{(1-W_i)Y_i}{1-\pi(Z_i;\psi)}, \tag{1.4}$$

  where $\pi(Z;\psi) = \Pr(W=1|Z)$ is a model for the propensity score, and ψ can be estimated based on the data $(W_i, Z_i), i = 1, \cdots, n$.

Exercise 1.4

Show that (1.4) is unbiased for the average causal treatment effect (assuming ψ is at its true value).

Example: Bivariate outcome with a missing component

- Here is hypothetical example: UW-Madison Division of Recreational Sports offered a one-semester fitness program designed to help participants lose weight.
- To assess how this program is doing, they randomly selected 100 participants and measured their BMI values at enrollment and after completion of the program. The aim is to see how the average BMIs change before and after treatment.
- However, some of recruits did not go through the training program, so their post-treatment BMI value is missing.
- We assume that the decision for non-compliance depends solely on the pre-treatment BMI. So the post-treatment BMI is MAR.

Example: Bivariate outcome with a missing component

- To put the question into statistical framework, the full data consist of a bivariate outcome (Y_1, Y_2) with Y_2 possibly missing.
- So the observed data consist of

$$(R_i, Y_{1i}, R_i Y_{2i}), \quad i = 1, \cdots, n.$$

- The aim is to estimate $EY_1 - EY_2$. Since we can certainly estimate EY_1 by the sample average of fully observed Y_1, we focus on the estimation of EY_2.

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- We first consider MLE using the EM algorithm.
- In that case we need to have a model for the full data (Y_1, Y_2). Assume that

$$(Y_1, Y_2) \sim N\left\{\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right\}.$$

- Denote $\theta = (\mu, \Sigma)$.
- From here on, for simplicity in describing the algorithms, we use small-case letters to denote the iid sample

$$(r_i, y_{1i}, r_i y_{2i}), \quad i = 1, \cdots, n.$$

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- We first look at the M step.
- Under mild regularity conditions (which hold in this case),

$$\frac{\partial}{\partial\theta}Q(\theta|\theta^{(j)}) := \frac{\partial}{\partial\theta}E\left[l(\theta|D)\Big|D_{\text{obs}},\theta^{(j)}\right] = E\left[\frac{\partial}{\partial\theta}l(\theta|D)\Big|D_{\text{obs}},\theta^{(j)}\right].$$

  So that the M step amounts to solving the conditional expectation of the score function.
- By A1.2, the M step can be explicitly expressed as

$$\mu^{(j+1)} = E[\overline{y}|D_{\text{obs}},\theta^{(j)}], \quad \Sigma^{(j+1)} = n^{-1}E\left[\sum_{i=1}^{n}(y_i-\mu^{(j+1)})^{\otimes 2}\Big|D_{\text{obs}},\theta^{(j)}\right],$$

  where $a^{\otimes 2} = aa^{\mathrm{T}}$ for any vector a.

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- Now, the E step.
- Since all y_1 belong to D_{obs}, we have that

$$\mu_1^{(j)} = \widehat{\mu}_1 := \overline{y}_1, \qquad \forall j.$$

- Using the conditional expectation formula given in A1.2.3, we have

$$\mu_2^{(j+1)} = n^{-1}\sum_{i=1}^{n}E[y_{2i}|y_{1i},r_iy_{2i},\theta^{(j)}] = n^{-1}\sum_{i=1}^{n}\widehat{y}_{2i}^{(j)},$$

  where $\widehat{y}_{2i}^{(j)} = y_{2i}$ if $r_i = 1$ and $\widehat{y}_{2i}^{(j)} = \mu_2^{(j)} + \sigma_{12}^{(j)}{\sigma_{11}^{(j)}}^{-1}(y_{1i}-\widehat{\mu}_1)$ if $r_i = 0$.

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- For $\Sigma^{(j+1)}$, note that

$$(y_i - \mu^{(j+1)})^{\otimes 2} = \begin{pmatrix} \left(y_{1i} - \mu_1^{(j+1)}\right)^2 & \left(y_{1i} - \mu_1^{(j+1)}\right)\left(y_{2i} - \mu_2^{(j+1)}\right) \\ & \left(y_{2i} - \mu_2^{(j+1)}\right)^2 \end{pmatrix}.$$

- The $(1,1)$th term is constant under the conditional expectation. The $(1,2)$th term is linear in y_{2i}, so its conditional expectation is to replace y_{2i} with $\widehat{y}_{2i}^{(j)}$.

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- The conditional expectation of the $(2,2)$th term is $\left(y_{2i} - \mu_2^{(j+1)}\right)^2$ if $r_i = 1$ and is $\sigma_{2|1}^{(j)} + \left(\widehat{y}_{2i}^{(j)} - \mu_2^{(j+1)}\right)^2$, where $\sigma_{2|1}^{(j)} = \sigma_{22}^{(j)} - {\sigma_{12}^{(j)}}^2 {\sigma_{11}^{(j)}}^{-1}$.
- Denote this term as $\widehat{V}_{2i}^{(j)}$.
- So,

$$\Sigma^{(j+1)} = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} (y_{1i} - \widehat{\mu}_1)^2 & (y_{1i} - \widehat{\mu}_1)\left(\widehat{y}_{2i}^{(j)} - \mu_2^{(j+1)}\right) \\ & \widehat{V}_{2i}^{(j)} \end{pmatrix},$$

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- In sum, to compute the MLEs, we first compute the non-iterative part:

$$\widehat{\mu}_1 = \overline{y}_1, \quad \widehat{\sigma}_{11} = n^{-1} \sum_{i=1}^{n} (y_{1i} - \overline{y}_1)^2 .$$

- At the $(j+1)$th iteration with parameter $\theta^{(j)}$, compute

$$\widehat{y}_{2i}^{(j)} = r_i y_{2i} + (1 - r_i)\{\mu_2^{(j)} + \sigma_{12}^{(j)} \widehat{\sigma}_{11}^{-1} (y_{1i} - \widehat{\mu}_1)\}.$$

  Update $\mu_2^{(j+1)} = n^{-1} \sum_{i=1}^{n} \widehat{y}_{2i}^{(j)}$ and

$$\sigma_{12}^{(j+1)} = n^{-1} \sum_{i=1}^{n} (y_{1i} - \widehat{\mu}_1) \left(\widehat{y}_{2i}^{(j)} - \mu_2^{(j+1)} \right).$$

Example: Bivariate outcome with a missing component
MLE with EM algorithm

- Then, compute

$$\widehat{V}_{2i}^{(j)} = \begin{cases} \left(y_{2i} - \mu_2^{(j+1)} \right)^2, & r_i = 1 \\ \sigma_{22}^{(j)} - {\sigma_{12}^{(j)}}^2 \widehat{\sigma}_{11}^{-1} + \left(\widehat{y}_{2i}^{(j)} - \mu_2^{(j+1)} \right)^2, & r_i = 0 \end{cases}$$

  Update $\sigma_{22}^{(j+1)} = n^{-1} \sum_{i=1}^{n} \widehat{V}_{2i}^{(j)}$.

- Note that this step is optional if we are only interested in estimating μ_2, because the iterative steps of μ_2 do not involve σ_{22}.

Example: Bivariate outcome with a missing component
WEE

- A nonparametric estimator for $\mu_2 := EY_2$ with full data is $n^{-1}\sum_{i=1}^n Y_{2i}$.
- Under MAR, the CC estimator

$$\frac{\sum_{i=1}^n R_i Y_{2i}}{\sum_{i=1}^n R_i}$$

  is biased as R may be (marginally) correlated with Y_2.
- We build a model for the selection probability $\pi(Y_1;\psi) = \Pr(R=1|Y)$, say, logistic regression model, i.e.,

$$\pi(Y_1;\psi) = \frac{e^{\psi_0+\psi_1 Y_1}}{1+e^{\psi_0+\psi_1 Y_1}},$$

  and estimate ψ by its MLE $\widehat{\psi}$ based on $(R_i, Y_{1i}), i = 1, \cdots, n$.

Example: Bivariate outcome with a missing component
WEE

- Then, a valid estimator for μ_2 is the inverse probability weighted (IPW) estimator

$$\widehat{\mu}_2^{IPW} = n^{-1}\sum_{i=1}^n \frac{R_i Y_{2i}}{\pi(Y_{1i};\widehat{\psi})}.$$

- Compared with the MLE method, which needs a parametric model for the distribution of Y, the IPW does not require such a model.
- However, the IPW requires a parametric model for the missingness mechanism.
- In this sense, the IPW is *semiparametric*.

Example: Bivariate outcome with a missing component
WEE

- Interestingly, there is a way to combine the strengths of the two approaches.
- First, let's build a regression model for Y_2 on Y_1: $\mu(Y_1;\beta) = E[Y_2|Y_1]$, e.g., a linear regression model

$$\mu(Y_1;\beta) = \beta_0 + \beta_1 Y_1.$$

- The parameter estimate $\widehat{\beta}$ can be computed using least squares by the CC analysis on $\{(Y_{2i}, Y_{1i}) : R_i = 1, i = 1, \cdots, n\}$ (why is CC analysis valid here?).

Example: Bivariate outcome with a missing component
WEE

- Consider the following estimator

$$\widehat{\mu}_2^{DR} = n^{-1}\sum_{i=1}^{n}\frac{R_i Y_{2i}}{\pi(Y_{1i};\widehat{\psi})} + n^{-1}\sum_{i=1}^{n}\left(1 - \frac{R_i}{\pi(Y_{1i};\widehat{\psi})}\right)\mu(Y_{1i};\widehat{\beta}).$$

- This estimator is *doubly robust* (DR) in the sense that it is valid when *either* the π model *or* the μ model is correct.
- To see this, fixing ψ and β at their true values, one can show that the expectation of

$$\frac{RY_2}{\pi(Y_1;\psi)} + \left(1 - \frac{R}{\pi(Y_1;\psi)}\right)\mu(Y_1;\beta)$$

is μ_2 when either model is true. See A1.3.

Overview of Statistical Methods

Example: Bivariate outcome with a missing component
WEE

- Still more interesting is the fact that when both models are correct, $\widehat{\mu}_2^{DR}$ has smaller (asymptotic) variance than $\widehat{\mu}_2^{IPW}$.
- A variety of DR semiparametric approaches have been developed to account for missing observations without making strict parametric assumptions.
- A general DR approach using weighted estimating equations has been proposed by Robins, Rotnitzky, and Zhao (1994). The general weighted estimating equations (Robins and Ritov, 1997) are doubly robust in the sense that, in order to obtain a valid estimate of the parameters, either the missing data mechanism or the conditional distribution of the missing data given the observed data, has to be correctly specified, but not both.

References

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1-38.
- Falkson, G., Cnaan, A., Simson, I. W., Dayal, Y., Falkson, H., Smith, T. J., & Haller, D. G. (1990). A randomized phase II study of acivicin and 4'deoxydoxorubicin in patients with hepatocellular carcinoma in an Eastern Cooperative Oncology Group study. American journal of clinical oncology, 13, 510-515.
- Falkson, G., Lipsitz, S., Borden, E. Simson, I.W. & Haller, D. (1994) A ECOG ran domized Phase II study of beta Interferon and Menogoril. American Journal Of Clinical Oncology, 18, 287-292.
- Robins, James M., Andrea Rotnitzky, & Lue Ping Zhao (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. Journal of the American Statistical Association, 89, 846-866.
- Robins, J. M. & Ritov, Y. A. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models. Statistics in Medicine, 16, 285-319.
- Ware, J. H., Dockery, D. W., Spiro III, A., Speizer, F. E., & Ferris Jr, B. G. (1984). Passive Smoking, Gas Cooking, and Respiratory Health of Children Living in Six Cities 1-3. American Review of Respiratory Disease, 129, 366-374.