Chapter 1

Introduction

Incomplete Data

- Due to no direct measurement
- Due to refusal / Don't know / not available
- Due to uncertainty in the measurement
- Due to design
- Due to self-selection

Example 1: No direct measurement

- A study of managers of Iowa farmer cooperatives ($n = 98$)
- Five variables
  - x_1: Knowledge (knowledge of the economic phase of management directed toward profit-making in a business and product knowledge)
  - x_2: Value Orientation (tendency to rationally evaluate means to an economic end)
  - x_3: Role Satisfaction (gratification obtained by the manager from performing the managerial role)
  - x_4: Past Training (amount of formal education)
  - y: Role performance
- We are interested in estimating parameters in the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Example 1 (Cont'd)

Measure	No. of Items	Mean	Reliability
x_1 Knowledge	26	1.38	0.6096
x_2 Value orientation	30	2.88	0.6386
x_3 Role satisfaction	11	2.46	0.8002
x_4 Past training	1	2.12	1.0000
y Role performance	24	0.0589	0.8230

Example 1 (Cont'd)

- Ordinary least squares method

$$\begin{aligned}\hat{Y} &= -0.9740 + \underset{(0.0535)}{0.2300X_1} + \underset{(0.0356)}{0.1199X_2} + \underset{(0.0375)}{0.0560X_3} + \underset{(0.0392)}{0.1099X_4}\end{aligned}$$

- Errors-in-variable estimates

$$\begin{aligned}\hat{Y} &= -1.1828 + \underset{(0.1288)}{0.3579X_1} + \underset{(0.0794)}{0.1549X_2} + \underset{(0.0510)}{0.0613X_3} + \underset{(0.0447)}{0.0715X_4}\end{aligned}$$

Reference:

Warren, White, and Fuller (1974). "An Errors-In-Variables Analysis of Managerial Role Performance", JASA, 69, p 886-893.

Example 2. Asthma Study Data (Pigott, 2001)

Variable descriptions

Variable	Definition	Possible values	Mean	N
Asthma belief	Level of confidence	1= little confidence 5= lots of confidence	4.057	154
Group	Treatment or control	0 = treatment 1 = control	0.558	154
Symsev	Severity of asthma symptoms in 2 weeks	0 = no symptoms 3 = severe symptoms	0.235	141
Reading	Standardized state reading test scores	Grade equivalent scores, from 1.10 to 8.10	3.443	79
Age		Ranging from 8 to 14	10.586	152
Gender		0 =Male 1 = Female	0.442	154
Allergy	No. of allergies	Range from 0 to 7	2.783	83

Missing Data Patterns

Symsev	Reading	Age	Allergy	# of cases	% of cases
O	O	O	O	19	12.3
M	O	O	O	1	0.6
O	M	O	O	54	35.1
O	O	O	M	56	36.4
M	M	O	O	9	5.8
M	O	O	M	1	0.6
O	M	O	M	10	6.5
O	O	M	M	2	1.3
M	M	O	M	2	1.3
				154	100.0

Example 2 (Cont'd)

Results (CC: Complete Case, ML: Maximum Likelihood)

Variable	CC analysis		ML analysis	
	B	SE	B	SE
Intercept	4.617	0.838	4.083	0.362
Trt group	-0.550	0.276	-0.132	0.112
Symsev	-0.315	0.161	-0.480	0.144
Reading	0.409	0.096	0.218	0.039
Age	-0.211	0.115	-0.089	0.043
Gender	0.198	0.189	0.084	0.104
Allergy	-0.005	0.057	0.063	0.029

Reference:

Pigott (2001). "A Review of Methods for Missing Data", *Educational Research and Evaluation*, 7, 353-383.

Example 3: 2009 Local Area Labor Force survey in Korea.

- Large scale survey with about $n = 157K$ sample households.
- Obtain the employment status: Employed, Unemployed, Not in labor force.
- To obtain response, interviewers visit the sample households up to four times. That is, the current rule allows for three follow-ups.

Example 3 (Cont'd)

Realized Responses from the Korean LF survey data

status	t=1	t=2	t=3	t=4	No response
Employment	81,685	46,926	28,124	15,992	
Unemployment	1,509	948	597	352	32,350
Not in LF	57,882	32,308	19,086	10,790	

Example 3 (Cont'd)

	First Response at t-th visit				No Response
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	
Response Rate (%)	42.94	24.40	14.55	8.26	9.85
Ave. Unemp. Rate (%)	1.81	1.98	2.08	2.15	?

Response propensity seems to be correlated with the unemployment rate.

Reference:

Kim, J.K. and Im, J. (2014). "Propensity score weighting adjustment with several follow-ups", *Biometrika* **101**, 439-448.

Example 4: BMI data example

- Korean Longitudinal Study of Aging (KLoSA) data (http://www.kli.re.kr/klosa/en/about/introduce.jsp)
- Original sample measures height and weight from survey questions (N=9,842)
- A validation sample (n=505) is randomly selected from the original sample to obtain physical measurement for the height and weight.

Reference:

Y. Xu, J.K. Kim, and Y. Li. (2017). "Semiparametric estimation for measurement error models with validation data", Canadian Journal of Statistics 45, 185–201.

Outline

Table: **Outline of a 15-week lecture**

Weeks	Chapter	Topic
1-2	1-2	Introduction. Likelihood-based approach
3-5	3	Computation
6-7	4	Imputation
8-9	5-6	Multiple Imputation & Fractional Imputation
10-11	7	Propensity scoring approach
12	8	Nonignorable missing data
13-14		Causal inference
15		Final Presentation

Overview lecture: Measurement errors in the outcome variable

Bayes theorem

- Bayes theorem

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{\sum_c P(D \mid C)P(C)}$$

where

  - C: true status (e.g. disease status) (Unobservable)
  - D: measurement (e.g. test result) (Observable)

- Conditional Bayes theorem

$$P(C \mid D, X) = \frac{P(D \mid C, X)P(C \mid X)}{\sum_c P(D \mid C, X)P(C \mid X)}, \tag{1}$$

where X is covariates (observable) .

- May assume that

$$P(D \mid C) = P(D \mid C, X).$$

  This is often called the non-differentiable measurement error assumption.

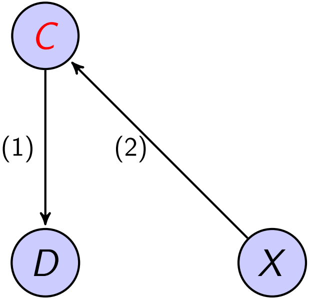
- Under the non-differentiable measurement error assumption, we can express (1) as

$$P(C \mid D, X) = \frac{P(D \mid C)P(C \mid X)}{\sum_c P(D \mid C)P(C \mid X)}, \tag{2}$$

Assumptions

- Two models in (2):
  1. $P(D \mid C)$: data model
  2. $P(C \mid X)$: process model
- Data model is known (or directly estimated from a validation sample)
- The process model has a probability structure. That is, we may use $P(C \mid X) = P(C \mid X; \theta)$ for some $\theta \in \Omega \subset \mathbb{R}^p$.
- The true status C is not observed, but we observe D and X.
- To make the presentation simple, we will assume C is binary with support $\{0, 1\}$.

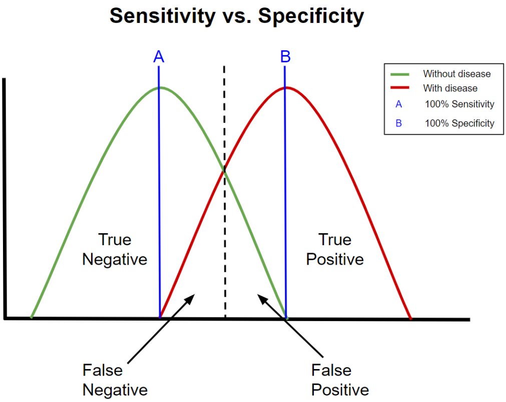
Measurement error model framework



(1): Data model (known),
(2): Process model (known up to θ).

Data Model

- $C = 1$ means disease status
  - sensitivity (true positive rate): $P(D = 1 \mid C = 1) = 1 - \alpha$
  - specificity (true negative rate): $P(D = 0 \mid C = 0) = 1 - \beta$



Process Model

- We may use a statistical model for the probability of $C = 1$, denoted by $P(C = 1 \mid X; \theta)$ with unknown θ.
- For example, we may use a logistic regression model

$$P(C = 1 \mid X; \theta) = \frac{\exp(\mathbf{x}'\theta)}{1 + \exp(\mathbf{x}'\theta)} := \pi(\mathbf{x}; \theta)$$

- If θ is known, then we can compute

$$P(C = 1 \mid X, D) = \frac{P(D \mid C = 1)P(C = 1 \mid X; \theta)}{\sum_c P(D \mid C = c)P(C = c \mid X; \theta)}$$

as a denoised version of classification.

Remark

- The parameters in the data model are assumed to be known.
- If the true labels were observed, then we could use the following maximum likelihood method for parameter estimation.

$$\hat{\theta} = \arg\max_{\theta} \ell_C(\theta), \tag{3}$$

where

$$\ell_C(\theta) = \sum_{i=1}^{n} \left[C_i \log \pi(\mathbf{x}_i; \theta) + (1 - C_i) \log\{1 - \pi(\mathbf{x}_i; \theta)\} \right]$$

is the log-likelihood function of θ using the C-values.

- More generally, we may use a general form of loss function $\ell(\pi, C)$ associated with a classifier $\pi(\mathbf{x})$.

- Idea:
  1. Compute the conditional probability for

$$
\begin{aligned}
P(D=1 \mid X;\theta) &= \sum_{c=0}^{1} P(D=1 \mid C=c)P(C=c \mid X;\theta) \\
&= \beta\{1-\pi(X;\theta)\} + (1-\alpha)\pi(X;\theta) \\
&:= \tilde{\pi}(X;\theta)
\end{aligned}
$$

  2. Construct the loss function for θ using the noisy label:

$$
\ell_{obs}(\theta) = \sum_{i=1}^{n} \left[D_i \log \tilde{\pi}_d(\mathbf{x}_i;\theta) + (1-D_i)\log\{1-\tilde{\pi}_d(\mathbf{x}_i;\theta)\} \right]. \quad (4)
$$

  3. Compute the maximizer of $L_{obs}(\theta)$:

$$
\hat{\theta} = \arg\max_{\theta} \ell_{obs}(\theta).
$$

Learning with noisy labels: 2. EM algorithm

- First define the log-likelihood function using true label C:

$$\ell_C(\theta) = \sum_{i=1}^{n} [C_i \log \pi(X_i; \theta) + (1 - C_i) \log\{1 - \pi(X_i; \theta)\}]$$

- Iterative computation:
  - **E-step**: Given the current parameter $\theta^{(t)}$, compute

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E\{\ell_C(\theta) \mid X, D; \theta^{(t)}\} \\ &= \sum_{i=1}^{n} \left[\hat{C}_i^{(t)} \log \pi(X_i; \theta) + (1 - \hat{C}_i^{(t)}) \log\{1 - \pi(X_i; \theta)\} \right], \end{aligned}$$

    where $\hat{C}_i^{(t)} = E\left(C_i \mid X_i, D_i; \theta^{(t)}\right)$.
  - **M-step**: Update θ by

$$\theta^{(t+1)} = \arg\max Q(\theta \mid \theta^{(t)}). \tag{5}$$

Remark

- In the E-step, we use Bayes theorem

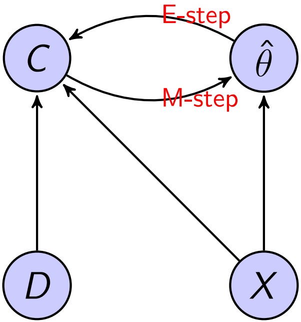
$$
\begin{aligned}
E\left(C_i \mid \mathbf{x}_i, D_i; \theta^{(t)}\right) &= P\left(C_i = 1 \mid \mathbf{x}_i, D_i; \theta^{(t)}\right) \\
&= \frac{P\left(C_i = 1 \mid \mathbf{x}_i; \theta^{(t)}\right) P(D_i \mid C_i = 1)}{\sum_{c=0}^{1} P\left(C_i = c \mid \mathbf{x}_i; \theta^{(t)}\right) P(D_i \mid C_i = c)}.
\end{aligned}
$$

- To implement the M-step in (5), note that the MLE procedure in (3) is a mapping from $S = \{(\mathbf{x}_i, C_i), i = 1, \cdots, n\}$ to $\hat{\theta}$. That is, $\hat{\theta} = \hat{\theta}(S)$.
- Because

$$
\ell_{com}(\theta) = \sum_{i=1}^{n} \left[C_i \log \pi(\mathbf{x}_i; \theta) + (1 - C_i) \log\{1 - \pi(\mathbf{x}_i; \theta)\} \right],
$$

the M-step in (5) can be expressed as $\theta^{(t+1)} = \hat{\theta}(S^{(t)})$, where $S^{(t)} = \{(\mathbf{x}_i, \hat{C}_i^{(t)}), i = 1, \cdots, n\}$.

EM algorithm



Prediction

- Best prediction: Expectation from the prediction model at $\theta = \hat{\theta}$

$$\hat{C}_i^* = E\left(C_i \mid D_i, X_i; \hat{\theta} \right) \tag{6}$$

This is a denoised version of D_i.

- Prediction model is obtained by combining data model with process model using Bayes theorem:

$$P\left(C_i = 1 \mid X_i, D_i; \hat{\theta} \right) = \frac{P(C_i = 1 \mid X_i; \hat{\theta}) P(D_i \mid C_i = 1)}{\sum_{c=0}^{1} P(C_i = c \mid X_i; \hat{\theta}) P(D_i \mid C_i = c)}$$

Prediction error

- Let
$$C_i^* = E\left(C_i \mid D_i, X_i; \theta\right) := C_i^*(\theta).$$
- Prediction error of $\hat{C}_i^* = C_i^*(\hat{\theta})$ in (6):
$$\hat{C}_i^* - C_i = \{C_i^*(\theta) - C_i\} + \left\{C_i^*(\hat{\theta}) - C_i^*(\theta)\right\}. \tag{7}$$
- In (7), the first part is the genuine prediction error and the second part is the error due to the uncertainty in $\hat{\theta}$.
- Mean Squared Prediction Error:
$$\begin{aligned} MSPE(\hat{C}_i^*) &\doteq E\{(C_i^* - C_i)^2\} + B_i V(\hat{\theta}) B_i' \\ &= E\{V\left(C_i \mid D_i, X_i\right)\} + B_i V(\hat{\theta}) B_i', \end{aligned}$$

where $B_i = \partial C_i^*(\theta)/\partial\theta$.

Statistical Methods (Summary)

Basic Steps

1. Model Specification
   - Data model
   - Process model
2. Parameter estimation
   - Direct maximization of marginal likelihood
   - EM alghorithm
3. Best prediction
   - Derive the predictive model using Bayes formula
   - Best prediction is obtained by computing the expectation of the prediction model evaluated at MLE.
4. Uncertainty quantification
   - Linearization or Bootstrap
   - Bayesian approach