Recent Advances in the analysis of missing data with non-ignorable missingness

Jae-Kwang Kim

Department of Statistics, Iowa State University

July 4th, 2014
1 Introduction

2 Full likelihood-based ML estimation

3 GMM method

4 Partial Likelihood approach

5 Exponential tilting method

6 Concluding Remarks
Observed likelihood

- \((X, Y)\): random variable, \(y\) is subject to missingness
- \(f(y \mid x; \theta)\): model of \(y\) on \(x\)
- \(g(\delta \mid x, y; \phi)\): model of \(\delta\) on \((x, y)\)
- Observed likelihood

\[
L_{obs}(\theta, \phi) = \prod_{\delta_i=1} f(y_i \mid x_i; \theta) g(\delta_i \mid x_i, y_i; \phi) \\
\times \prod_{\delta_i=0} \int f(y_i \mid x_i; \theta) g(\delta_i \mid x_i, y_i; \phi) \, dy_i
\]

- Under what conditions the parameters are identifiable (or estimable)?
Suppose that we can decompose the covariate vector $x$ into two parts, $u$ and $z$, such that

$$g(\delta | y, x) = g(\delta | y, u)$$  \hspace{1cm} (1)

and, for any given $u$, there exist $z_{u,1}$ and $z_{u,2}$ such that

$$f(y | u, z = z_{u,1}) \neq f(y | u, z = z_{u,2}).$$  \hspace{1cm} (2)

Under some other minor conditions, all the parameters in $f$ and $g$ are identifiable.
Remark

• Condition (1) means

\[ \delta \perp z \mid y, u. \]

• That is, given \((y, u)\), \(z\) does not help in explaining \(\delta\).

• Thus, \(z\) plays the role of instrumental variable in econometrics:

\[ f(y^* \mid x^*, z^*) = f(y^* \mid x^*), \quad Cov(z^*, x^*) \neq 0. \]

Here, \(y^* = \delta, x^* = (y, u)\), and \(z^* = z\).

• We may call \(z\) the **nonresponse instrument** variable.

• Rigorous theory developed by Wang et al. (2014).
Remark

- MCAR (Missing Completely at random): $P(\delta \mid y)$ does not depend on $y$.
- MAR (Missing at random): $P(\delta \mid y) = P(\delta \mid y_{obs})$
- NMAR (Not Missing at random): $P(\delta \mid y) \neq P(\delta \mid y_{obs})$
- Thus, MCAR is a special case of MAR.
Parameter estimation under the existence of nonresponse instrument variable

- Full likelihood-based ML estimation
- Generalized method of moment (GMM) approach (Section 6.3 of KS)
- Conditional likelihood approach (Section 6.2 of KS)
- Pseudo likelihood approach (Section 6.4 of KS)
- Exponential tilting method (Section 6.5 of KS)
- Latent variable approach (Section 6.6 of KS)
1 Introduction

2 Full likelihood-based ML estimation

3 GMM method

4 Partial Likelihood approach

5 Exponential tilting method

6 Concluding Remarks
Full likelihood-based ML estimation

- Wish to find $\hat{\eta} = (\hat{\theta}, \hat{\phi})$, that maximizes the observed likelihood

$$L_{obs}(\eta) = \prod_{\delta_i=1} f(y_i | x_i; \theta) g(\delta_i | x_i, y_i; \phi) \times \prod_{\delta_i=0} \int f(y_i | x_i; \theta) g(\delta_i | x_i, y_i; \phi) dy_i$$

- Mean score theorem: Under some regularity conditions, finding the MLE of maximizing the observed likelihood is equivalent to finding the solution to

$$\bar{S}(\eta) \equiv E\{S(\eta) | y_{obs}, \delta; \eta\} = 0,$$

where $y_{obs}$ is the observed data. The conditional expectation of the score function is called mean score function.
Example 1 [Example 2.5 of KS]

1. Suppose that the study variable $y$ is randomly distributed with Bernoulli distribution with probability of success $p_i$, where

$$p_i = p_i(\beta) = \frac{\exp(x'_i\beta)}{1 + \exp(x'_i\beta)}$$

for some unknown parameter $\beta$ and $x_i$ is a vector of the covariates in the logistic regression model for $y_i$. We assume that 1 is in the column space of $x_i$.

2. Under complete response, the score function for $\beta$ is

$$S_1(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta)) x_i.$$
Example 1 (Cont’d)

3. Let $\delta_i$ be the response indicator function for $y_i$ with distribution $Bernoulli(\pi_i)$ where

$$\pi_i = \frac{\exp(x_i'\phi_0 + y_i\phi_1)}{1 + \exp(x_i'\phi_0 + y_i\phi_1)}.$$ 

We assume that $x_i$ is always observed, but $y_i$ is missing if $\delta_i = 0$.

4. Under missing data, the mean score function for $\beta$, the conditional expectation of the original score function given the observed data, is

$$\bar{S}_1(\beta, \phi) = \sum_{\delta_i=1} \{y_i - p_i(\beta)\} x_i + \sum_{\delta_i=0}^{\frac{1}{y_i}=0} w_i(y; \beta, \phi) \{y - p_i(\beta)\} x_i,$$

where $w_i(y; \beta, \phi)$ is the conditional probability of $y_i = y$ given $x_i$ and $\delta_i = 0$:

$$w_i(y; \beta, \phi) = \frac{P_\beta(y_i = y \mid x_i) P_\phi(\delta_i = 0 \mid y_i = y, x_i)}{\sum_{z=0}^{1} P_\beta(y_i = z \mid x_i) P_\phi(\delta_i = 0 \mid y_i = z, x_i)}$$

Thus, $\bar{S}_1(\beta, \phi)$ is also a function of $\phi$. 
Example 1 (Cont’d)

5 If the response mechanism is MAR so that $\phi_1 = 0$, then

$$w_i (y; \beta, \phi) = \frac{P_\beta (y_i = y \mid x_i)}{\sum_{z=0}^1 P_\beta (y_i = z \mid x_i)} = P_\beta (y_i = y \mid x_i)$$

and so

$$\tilde{S}_1 (\beta, \phi) = \sum_{\delta_i=1} \{y_i - p_i (\beta)\} x_i = \tilde{S}_1 (\beta).$$

6 If MAR does not hold, then $(\hat{\beta}, \hat{\phi})$ can be obtained by solving $\tilde{S}_1 (\beta, \phi) = 0$ and $\tilde{S}_2 (\beta, \phi) = 0$ jointly, where

$$\tilde{S}_2 (\beta, \phi) = \sum_{\delta_i=1} \{\delta_i - \pi (\phi; x_i, y_i)\} (x_i, y_i)$$

$$+ \sum_{\delta_i=0}^{1} \sum_{y=0}^{1} w_i (y; \beta, \phi) \{\delta_i - \pi_i (\phi; x_i, y)\} (x_i, y).$$
• Interested in finding $\hat{\eta}$ that maximizes $L_{obs}(\eta)$. The MLE can be obtained by solving $S_{obs}(\eta) = 0$, which is equivalent to solving $\bar{S}(\eta) = 0$ by the mean score theorem.

• EM algorithm provides an alternative method of solving $\bar{S}(\eta) = 0$ by writing

$$\bar{S}(\eta) = E \left\{ S_{com}(\eta) \mid y_{obs}, \delta; \eta \right\}$$

and using the following iterative method:

$$\hat{\eta}^{(t+1)} \leftarrow \text{solve } E \left\{ S_{com}(\eta) \mid y_{obs}, \delta; \hat{\eta}^{(t)} \right\} = 0.$$
5. EM algorithm

Definition

Let $\eta^{(t)}$ be the current value of the parameter estimate of $\eta$. The EM algorithm can be defined as iteratively carrying out the following E-step and M-steps:

- **E-step**: Compute
  \[ Q \left( \eta \mid \eta^{(t)} \right) = E \left\{ \ln f (y, \delta; \eta) \mid y_{\text{obs}}, \delta, \eta^{(t)} \right\} \]

- **M-step**: Find $\eta^{(t+1)}$ that maximizes $Q(\eta \mid \eta^{(t)})$ w.r.t. $\eta$. 
Example 1 (Cont’d)

- E-step:

$$\bar{S}_1 \left( \beta \mid \beta^{(t)}, \phi^{(t)} \right) = \sum_{\delta_i=1} \{y_i - p_i (\beta)\} x_i + \sum_{\delta_i=0}^{1} \sum_{j=0} w_{ij(t)} \{j - p_i (\beta)\} x_i,$$

where

$$w_{ij(t)} = Pr(Y_i = j \mid x_i, \delta_i = 0; \beta^{(t)}, \phi^{(t)})$$

$$= \frac{Pr(Y_i = j \mid x_i; \beta^{(t)}) Pr(\delta_i = 0 \mid x_i, j; \phi^{(t)})}{\sum_{y=0}^{1} Pr(Y_i = y \mid x_i; \beta^{(t)}) Pr(\delta_i = 0 \mid x_i, y; \phi^{(t)})}$$

and

$$\bar{S}_2 \left( \phi \mid \beta^{(t)}, \phi^{(t)} \right) = \sum_{\delta_i=1} \{\delta_i - \pi_i (x_i, y_i; \phi)\} (x'_i, y'_i)$$

$$+ \sum_{\delta_i=0}^{1} \sum_{j=0} w_{ij(t)} \{\delta_i - \pi_i (x_i, j; \phi)\} (x'_i, j)'.$$
Example 1 (Cont’d)

- **M-step:**
  The parameter estimates are updated by solving
  \[
  \begin{bmatrix}
  \bar{S}_1 \left( \beta \mid \beta^{(t)}, \phi^{(t)} \right) \\
  \bar{S}_2 \left( \phi \mid \beta^{(t)}, \phi^{(t)} \right)
  \end{bmatrix} = (0, 0)
  \]
  for \( \beta \) and \( \phi \).

- For categorical missing data, the conditional expectation in the E-step can be computed using the weighted mean with weights \( w_{ij(t)} \). Ibrahim (1990) called this method **EM by weighting**.
Monte Carlo EM

Motivation: Monte Carlo samples in the EM algorithm can be used as imputed values.

Monte Carlo EM

1. In the EM algorithm defined by

   - **[E-step]** Compute
     \[
     Q \left( \eta \mid \eta^{(t)} \right) = E \left\{ \ln f (y, \delta; \eta) \mid y_{\text{obs}}, \delta; \eta^{(t)} \right\}
     \]
   
   - **[M-step]** Find \( \eta^{(t+1)} \) that maximizes \( Q \left( \eta \mid \eta^{(t)} \right) \),

     E-step is computationally cumbersome because it involves integral.

2. Wei and Tanner (1990): In the E-step, first draw

   \[
   y_{mis}^{*(1)}, \cdots, y_{mis}^{*(m)} \sim f \left( y_{mis} \mid y_{\text{obs}}, \delta; \eta^{(t)} \right)
   \]

   and approximate

   \[
   Q \left( \eta \mid \eta^{(t)} \right) \approx \frac{1}{m} \sum_{j=1}^{m} \ln f \left( y_{\text{obs}}, y_{mis}^{*(j)}, \delta; \eta \right).
   \]
Example 2 [Example 3.15 of KS]

- Suppose that
  \[ y_i \sim f(y_i \mid x_i; \theta) \]
  Assume that \( x_i \) is always observed but we observe \( y_i \) only when \( \delta_i = 1 \) where \( \delta_i \sim Bernoulli [\pi_i(\phi)] \) and
  \[ \pi_i(\phi) = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 y_i)}. \]

- To implement the MCEM method, in the E-step, we need to generate samples from
  \[ f(y_i \mid x_i, \delta_i = 0; \hat{\theta}, \hat{\phi}) = \frac{f(y_i \mid x_i; \hat{\theta})\{1 - \pi_i(\hat{\phi})\}}{\int f(y_i \mid x_i; \hat{\theta})\{1 - \pi_i(\hat{\phi})\} dy_i}. \]
Example 2 (Cont’d)

- We can use the following rejection method to generate samples from $f(y_i | x_i, \delta_i = 0; \hat{\theta}, \hat{\phi})$:

  1. Generate $y_i^*$ from $f(y_i | x_i; \hat{\theta})$.
  2. Using $y_i^*$, compute

$$\pi_i^*(\hat{\phi}) = \frac{\exp(\hat{\phi}_0 + \hat{\phi}_1 x_i + \hat{\phi}_2 y_i^*)}{1 + \exp(\hat{\phi}_0 + \hat{\phi}_1 x_i + \hat{\phi}_2 y_i^*)}.$$

Accept $y_i^*$ with probability $1 - \pi_i^*(\hat{\phi})$.

  3. If $y_i^*$ is not accepted, then goto Step 1.
Example 2 (Cont’d)

- Using the \( m \) imputed values of \( y_i \), denoted by \( y_i^{*(1)}, \ldots, y_i^{*(m)} \), and the M-step can be implemented by solving

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} S \left( \theta; x_i, y_i^{*(j)} \right) = 0
\]

and

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \delta_i - \pi(\phi; x_i, y_i^{*(j)}) \right\} \left( 1, x_i, y_i^{*(j)} \right) = 0,
\]

where \( S (\theta; x_i, y_i) = \partial \log f(y_i \mid x_i; \theta)/\partial \theta. \)
1 Introduction

2 Full likelihood-based ML estimation

3 GMM method

4 Partial Likelihood approach

5 Exponential tilting method

6 Concluding Remarks
Basic setup

- \((X, Y)\): random variable
- \(\theta\): Defined by solving
  \[
  E\{U(\theta; X, Y)\} = 0.
  \]
- \(y_i\) is subject to missingness
  \[
  \delta_i = \begin{cases} 
  1 & \text{if } y_i \text{ responds} \\
  0 & \text{if } y_i \text{ is missing.}
  \end{cases}
  \]
- Want to find \(w_i\) such that the solution \(\hat{\theta}_w\) to
  \[
  \sum_{i=1}^{n} \delta_i w_i U(\theta; x_i, y_i) = 0
  \]
  is consistent for \(\theta\).
Basic Setup

• **Result 1:** The choice of
  
  \[ w_i = \frac{1}{E(\delta_i \mid x_i, y_i)} \]  

  makes the resulting estimator \( \hat{\theta}_w \) consistent.

• **Result 2:** If \( \delta_i \sim \text{Bernoulli}(\pi_i) \), then using \( w_i = 1/\pi_i \) also makes the resulting estimator consistent, but it is less efficient than \( \hat{\theta}_w \) using \( w_i \) in (3).
Parameter estimation: GMM method

- Because $z$ is a nonresponse instrumental variable, we may assume
  \[ P(\delta = 1 \mid x, y) = \pi(\phi_0 + \phi_1 u + \phi_2 y) \]
  for some $(\phi_0, \phi_1, \phi_2)$.

- Kott and Chang (2010): Construct a set of estimating equations such as
  \[
  \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi(\phi_0 + \phi_1 u_i + \phi_2 y_i)} - 1 \right\} (1, u_i, z_i) = 0
  \]
  that are unbiased to zero.

- May have overidentified situation: Use the generalized method of moments (GMM).
Suppose that we are interested in estimating the parameters in the regression model
\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \] (4)
where \( E(e_i | x_i) = 0. \)

Assume that \( y_i \) is subject to missingness and assume that
\[ P(\delta_i = 1 | x_{1i}, x_{i2}, y_i) = \frac{\exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}. \]

Thus, \( x_{2i} \) is the nonresponse instrument variable in this setup.
Example 3 (Cont’d)

- A consistent estimator of $\phi$ can be obtained by solving

$$
\hat{U}_2(\phi) \equiv \sum_{i=1}^{n} \left\{ \frac{\delta}{\pi(\phi; x_{1i}, y_i)} - 1 \right\} (1, x_{1i}, x_{2i}) = (0, 0, 0).
$$

(5)

Roughly speaking, the solution to (5) exists almost surely if $E\{\partial \hat{U}_2(\phi)/\partial \phi\}$ is of full rank in the neighborhood of the true value of $\phi$. If $x_2$ is vector, then (5) is overidentified and the solution to (5) does not exist. In the case, the GMM algorithm can be used.

- Finding the solution to $\hat{U}_2(\phi) = 0$ can be obtained by finding the minimizer of $Q(\phi) = \hat{U}_2(\phi)' \hat{U}_2(\phi)$ or $Q_W(\phi) = \hat{U}_2(\phi)' W \hat{U}_2(\phi)$ where $W = \{V(\hat{U}_2)\}^{-1}$. 
Example 3 (Cont’d)

• Once the solution $\hat{\phi}$ to (5) is obtained, then a consistent estimator of $\beta = (\beta_0, \beta_1, \beta_2)$ can be obtained by solving

\[
\hat{U}_1(\beta, \hat{\phi}) \equiv \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} \{ y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \} (1, x_{1i}, x_{2i}) = (0, 0, 0)
\] (6)

for $\beta$. 
The asymptotic variance of the GMM estimator \( \hat{\phi} \) that minimizes \( \hat{U}_2(\phi)'\hat{\Sigma}^{-1}\hat{U}_2(\phi) \) is

\[
V(\hat{\phi}) \cong \left( \Gamma'\Sigma^{-1}\Gamma \right)^{-1}
\]

where

\[
\Gamma = E\{\partial \hat{U}_2(\phi)/\partial \phi \}
\]

\[
\Sigma = V(\hat{U}_2).
\]

The variance is estimated by

\[
(\hat{\Gamma}'\hat{\Sigma}^{-1}\hat{\Gamma})^{-1},
\]

where \( \hat{\Gamma} = \partial \hat{U}/\partial \phi \) evaluated at \( \hat{\phi} \) and \( \hat{\Sigma} \) is an estimated variance-covariance matrix of \( \hat{U}_2(\phi) \) evaluated at \( \hat{\phi} \).
Asymptotic Properties

- The asymptotic variance of $\hat{\beta}$ obtained from (6) with $\hat{\phi}$ computed from the GMM can be obtained by

$$V(\hat{\theta}) \approx \left( \Gamma'_a \Sigma_a^{-1} \Gamma_a \right)^{-1}$$

where

$$\Gamma_a = E\{\partial \hat{U}(\theta)/\partial \theta\}$$
$$\Sigma_a = V(\hat{U})$$
$$\hat{U} = (\hat{U}'_1, \hat{U}'_2)'$$

and $\theta = (\beta, \phi)$. 
1 Introduction

2 Full likelihood-based ML estimation

3 GMM method

4 Partial Likelihood approach

5 Exponential tilting method

6 Concluding Remarks
Partial Likelihood approach

- A classical way of likelihood-based approach for parameter estimation under non-ignorable nonresponse is to maximize $L_{\text{obs}}(\theta, \phi)$ with respect to $(\theta, \phi)$, where

$$L_{\text{obs}}(\theta, \phi) = \prod_{\delta_i=1} f(y_i \mid x_i; \theta) g(\delta_i \mid x_i, y_i; \phi)$$

$$\times \prod_{\delta_i=0} \int f(y_i \mid x_i; \theta) g(\delta_i \mid x_i, y_i; \phi) \, dy_i$$

- Such approach can be called full likelihood-based approach because it uses full information available in the observed data.
- However, it is well known that such full likelihood-based approach is quite sensitive to the failure of the assumed model.
- On the other hand, partial likelihood-based approach (or conditional likelihood approach) uses a subset of the sample.
Conditional Likelihood approach

Idea

- Since
  \[ f(y \mid x)g(\delta \mid x, y) = f_1(y \mid x, \delta)g_1(\delta \mid x), \]
  for some \( f_1 \) and \( g_1 \), we can write

  \[
  L_{\text{obs}}(\theta) = \prod_{\delta_i = 1} f_1(y_i \mid x_i, \delta_i = 1) g_1(\delta_i \mid x_i) \\
  \times \prod_{\delta_i = 0} \int f_1(y_i \mid x_i, \delta_i = 0) g_1(\delta_i \mid x_i) dy_i
  \]

  \[
  = \prod_{\delta_i = 1} f_1(y_i \mid x_i, \delta_i = 1) \times \prod_{i=1}^n g_1(\delta_i \mid x_i).
  \]

- The conditional likelihood is defined to be the first component:

  \[
  L_c(\theta) = \prod_{\delta_i = 1} f_1(y_i \mid x_i, \delta_i = 1) = \prod_{\delta_i = 1} \frac{f(y_i \mid x_i; \theta)\pi(x_i, y_i)}{\int f(y \mid x_i; \theta)\pi(x_i, y)dy},
  \]

  where \( \pi(x, y_i) = Pr(\delta_i = 1 \mid x_i, y_i) \).
Example

- Assume that the original sample is a random sample from an exponential distribution with mean $\mu = 1/\theta$. That is, the probability density function of $y$ is $f(y; \theta) = \theta \exp(-\theta y)I(y > 0)$.
- Suppose that we observe $y_i$ only when $y_i > K$ for a known $K > 0$.
- Thus, the response indicator function is defined by $\delta_i = 1$ if $y_i > K$ and $\delta_i = 0$ otherwise.
Example

- To compute the maximum likelihood estimator from the observed likelihood, note that

\[ S_{\text{obs}}(\theta) = \sum_{\delta_i=1} \left( \frac{1}{\theta} - y_i \right) + \sum_{\delta_i=0} \left\{ \frac{1}{\theta} - E(y_i \mid \delta_i = 0) \right\}. \]

- Since

\[ E(Y \mid y > K) = \frac{1}{\theta} - \frac{K \exp(-\theta K)}{1 - \exp(-\theta K)}, \]

the maximum likelihood estimator of \( \theta \) can be obtained by the following iteration equation:

\[ \left\{ \hat{\theta}^{(t+1)} \right\}^{-1} = \bar{y}_r - \frac{n-r}{r} \left\{ \frac{K \exp(-K\hat{\theta}^{(t)})}{1 - \exp(-K\hat{\theta}^{(t)})} \right\}, \tag{7} \]

where \( r = \sum_{i=1}^n \delta_i \) and \( \bar{y}_r = r^{-1} \sum_{i=1}^n \delta_i y_i \).
Conditional Likelihood approach

Example

- Since \( \pi_i = Pr(\delta_i = 1 \mid y_i) = I(y_i > K) \) and \( E(\pi_i) = E\{I(y_i > K)\} = \exp(-K\theta) \), the conditional likelihood reduces to

\[
\prod_{\delta_i=1} \theta \exp\{-\theta(y_i - K)\}.
\]

The maximum conditional likelihood estimator of \( \theta \) is

\[
\hat{\theta}_c = \frac{1}{\bar{y}_r - K}.
\]

Since \( E(y \mid y > K) = \mu + K \), the maximum conditional likelihood estimator of \( \mu \), which is \( \hat{\mu}_c = 1/\hat{\theta}_c \), is unbiased for \( \mu \).
Remark

- Under some regularity conditions, the solution $\hat{\theta}_c$ that maximizes $L_c(\theta)$ satisfies

$$I_c^{1/2}(\hat{\theta}_c - \theta) \xrightarrow{\mathcal{L}} N(0, I)$$

where

$$I_c(\theta) = -E \left\{ \frac{\partial}{\partial \theta'} S_c(\theta) \mid x_i; \theta \right\} = \sum_{i=1}^{n} \left[ E \left\{ S_i S_i' \pi_i \mid x_i; \theta \right\} - \frac{E (S_i \pi_i \mid x_i; \theta)}{E (\pi_i \mid x_i; \theta)} \otimes^2 \right],$$

$$S_c(\theta) = \frac{\partial \ln L_c(\theta)}{\partial \theta}, \text{ and } S_i(\theta) = \frac{\partial \ln f(y_i \mid x_i; \theta)}{\partial \theta}.$$  

- Works only when $\pi(x, y)$ is a known function.
- Does not require nonresponse instrumental variable assumption.
- Popular for biased sampling problem.
Pseudo Likelihood approach

Idea

- Consider bivariate \((x_i, y_i)\) with density \(f(y | x; \theta) h(x)\) where \(y_i\) are subject to missingness.
- We are interested in estimating \(\theta\).
- Suppose that \(Pr(\delta = 1 | x, y)\) depends only on \(y\). (i.e. \(x\) is nonresponse instrument)
- Note that \(f(x | y, \delta) = f(x | y)\).
- Thus, we can consider the following conditional likelihood

\[
L_c(\theta) = \prod_{\delta_i = 1} f(x_i | y_i, \delta_i = 1) = \prod_{\delta_i = 1} f(x_i | y_i).
\]

- We can consider maximizing the pseudo likelihood

\[
L_p(\theta) = \prod_{\delta_i = 1} \frac{f(y_i | x_i; \theta) \hat{h}(x_i)}{\int f(y_i | x; \theta) \hat{h}(x) dx},
\]

where \(\hat{h}(x)\) is a consistent estimator of the marginal density of \(x\).
Idea

- We may use the empirical density in $\hat{h}(x)$. That is, $\hat{h}(x) = 1/n$ if $x = x_i$. In this case,

$$L_c(\theta) = \prod_{\delta_i=1} \frac{f(y_i \mid x_i; \theta)}{\sum_{k=1}^n f(y_i \mid x_k; \theta)}.$$  

- We can extend the idea to the case of $x = (u, z)$ where $z$ is a nonresponse instrument. In this case, the conditional likelihood becomes

$$\prod_{i: \delta_i=1} p(z_i \mid y_i, u_i) = \prod_{i: \delta_i=1} \frac{f(y_i \mid u_i, z_i; \theta) p(z_i \mid u_i)}{\int f(y_i \mid u_i, z; \theta) p(z \mid u_i) dz}.$$ (8)
Pseudo Likelihood approach

- Let \( \hat{p}(z|u) \) be an estimated conditional probability density of \( z \) given \( u \). Substituting this estimate into the likelihood in (8), we obtain the following pseudo likelihood:

\[
\prod_{i: \delta_i = 1} \frac{f(y_i | u_i, z_i; \theta) \hat{p}(z_i | u_i)}{\int f(y_i | u_i, z; \theta) \hat{p}(z | u_i) dz}.
\] (9)

- The pseudo maximum likelihood estimator (PMLE) of \( \theta \), denoted by \( \hat{\theta}_p \), can be obtained by solving

\[
S_p(\theta; \hat{\alpha}) \equiv \sum_{\delta_i = 1} [S(\theta; x_i, y_i) - E\{S(\theta; u_i, z, y_i) | y_i, u_i; \theta, \hat{\alpha}\}] = 0
\]

for \( \theta \), where \( S(\theta; x, y) = S(\theta; u, z, y) = \partial \log f(y | x; \theta) / \partial \theta \) and

\[
E\{S(\theta; u_i, z, y_i) | y_i, u_i; \theta, \hat{\alpha}\} = \frac{\int S(\theta; u_i, z, y_i)f(y_i | u_i, z; \theta)p(z | u_i; \hat{\alpha}) dz}{\int f(y_i | u_i, z; \theta)p(z | u_i; \hat{\alpha}) dz}.
\]
The Fisher-scoring method for obtaining the PMLE is given by

\[
\hat{\theta}_p^{(t+1)} = \hat{\theta}_p^{(t)} + \left\{ \mathcal{I}_p \left( \hat{\theta}^{(t)}, \hat{\alpha} \right) \right\}^{-1} S_p(\hat{\theta}^{(t)}, \hat{\alpha})
\]

where

\[
\mathcal{I}_p \left( \theta, \hat{\alpha} \right) = \sum_{\delta_i = 1} \left[ E \left\{ S(\theta; u_i, z, y_i)^{\otimes 2} \mid y_i, u_i; \theta, \hat{\alpha} \right\} - E \left\{ S(\theta; u_i, z, y_i) \mid y_i, u_i; \theta, \hat{\alpha} \right\}^{\otimes 2} \right] .
\]

Variance estimation is very complicated. Jackknife or bootstrap can be used.
1 Introduction

2 Full likelihood-based ML estimation

3 GMM method

4 Partial Likelihood approach

5 Exponential tilting method

6 Concluding Remarks
Exponential tilting method

Motivation

- Observed likelihood function can be written
  \[ L_{obs}(\phi) = \prod_{i=1}^{n} \{\pi_i(\phi)\}^{\delta_i} \left[ \int \{1 - \pi_i(\phi)\} f(y|x_i) dy \right]^{1-\delta_i}, \]

  where \( f(y|x) \) is the true conditional distribution of \( y \) given \( x \).

- To find the MLE of \( \phi \), we solve the mean score equation \( \bar{S}(\phi) = 0 \), where
  \[ \bar{S}(\phi) = \sum_{i=1}^{n} [\delta_i S_i(\phi) + (1 - \delta_i) E\{S_i(\phi)|x_i, \delta_i = 0\}] \tag{10} \]

  where \( S_i(\phi) = \{\delta_i - \pi_i(\phi)\} \frac{\partial \logit \pi_i(\phi)}{\partial \phi} \) is the score function of \( \phi \) for the density \( g(\delta | x, y; \phi) = \pi^\delta (1 - \pi)^{1-\delta} \) with \( \pi = \pi(x, y; \phi) \).
Motivation

• The conditional expectation in (10) can be evaluated by using

\[ f(y|x, \delta = 0) = f(y|x) \frac{P(\delta = 0|x, y)}{E\{P(\delta = 0|x, y)|x\}} \]  \hspace{1cm} (11)

Two problems occur:

1. Requires correct specification of \( f(y \mid x; \theta) \). Known to be sensitive to the choice of \( f(y \mid x; \theta) \).
2. Computationally heavy: Often uses Monte Carlo computation.
Exponential tilting method

Remedy (for Problem One)

Idea

Instead of specifying a parametric model for $f(y | x)$, consider specifying a parametric model for $f(y | x, \delta = 1)$, denoted by $f_1(y | x)$. In this case,

$$E\{S_i(\phi) | x_i, \delta_i = 0\} = \frac{\int S_i(\phi)f_1(y | x_i)O(x_i, y; \phi)dy}{\int f_1(y | x_i)O(x_i, y; \phi)dy}$$

where

$$O(x, y; \phi) = \frac{1 - \pi(\phi; x, y)}{\pi(\phi; x, y)}.$$
Remark

- Based on the following identity

\[
\begin{align*}
    f_0(y_i | x_i) &= f_1(y_i | x_i) \times \frac{O(x_i, y_i)}{E\{O(x_i, Y_i) | x_i, \delta_i = 1\}},
\end{align*}
\]  \hspace{1cm} (12)

where \( f_\delta (y_i | x_i) = f (y_i | x_i, \delta_i = \delta) \) and

\[
O(x_i, y_i) = \frac{Pr(\delta_i = 0 | x_i, y_i)}{Pr(\delta_i = 1 | x_i, y_i)}
\]  \hspace{1cm} (13)

is the conditional odds of nonresponse.

- \textit{Kim and Yu (2011)} considered a Kernel-based nonparametric regression method of estimating \( f(y | x, \delta = 1) \) to obtain \( E(Y | x, \delta = 0) \).
• If the response probability follows from a logistic regression model

\[
\pi(u_i, y_i) \equiv Pr(\delta_i = 1 | u_i, y_i) = \frac{\exp(\phi_0 + \phi_1 u_i + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 u_i + \phi_2 y_i)},
\]

(14)

the expression (12) can be simplified to

\[
f_0(y_i | x_i) = f_1(y_i | x_i) \times \frac{\exp(\gamma y_i)}{E \{\exp(\gamma Y) | x_i, \delta_i = 1\}},
\]

(15)

where \( \gamma = -\phi_2 \) and \( f_1(y | x) \) is the conditional density of \( y \) given \( x \) and \( \delta = 1 \).

• Model (15) states that the density for the nonrespondents is an exponential tilting of the density for the respondents. The parameter \( \gamma \) is the tilting parameter that determines the amount of departure from the ignorability of the response mechanism. If \( \gamma = 0 \), the the response mechanism is ignorable and \( f_0(y|x) = f_1(y|x) \).
Problem Two

How to compute

\[ E\{S_i(\phi) \mid x_i, \delta_i = 0\} = \frac{\int S_i(\phi)O(x_i, y; \phi)f_1(y \mid x_i)dy}{\int O(x_i, y; \phi)f_1(y \mid x_i)dy} \]

without relying on Monte Carlo computation?
Computation for

\[ E_1\{Q(x_i, Y) \mid x_i\} = \int Q(x_i, y)f_1(y \mid x_i)dy. \]

If \( x_i \) were null, then we would approximate the integration by the empirical distribution among \( \delta = 1 \).

Use

\[
\int Q(x_i, y)f_1(y \mid x_i)dy = \int Q(x_i, y)\frac{f_1(y \mid x_i)}{f_1(y)}f_1(y)dy
\]

\[ \propto \sum_{\delta_j=1} Q(x_i, y_j)\frac{f_1(y_j \mid x_i)}{f_1(y_j)} \]

where

\[ f_1(y) = \int f_1(y \mid x)f(x \mid \delta = 1)dx \]

\[ \propto \sum_{\delta_i=1} f_1(y \mid x_i). \]
In practice, \( f_1(y \mid x) \) is unknown and is estimated by \( \hat{f}_1(y \mid x) = f_1(y \mid x; \hat{\gamma}) \).

Thus, given \( \hat{\gamma} \), a fully efficient estimator of \( \phi \) can be obtained by solving

\[
S_2(\phi, \hat{\gamma}) \equiv \sum_{i=1}^{n} \left\{ \delta_i S(\phi; x_i, y_i) + (1 - \delta_i) \bar{S}_0(\phi \mid x_i; \hat{\gamma}, \phi) \right\} = 0, \tag{16}
\]

where

\[
\bar{S}_0(\phi \mid x_i; \hat{\gamma}, \phi) = \frac{\sum_{j=1}^{\delta_i} S(\phi; x_i, y_j) f_1(y_j \mid x_i; \hat{\gamma}) O(\phi; x_i, y_j)/\hat{f}_1(y_j)}{\sum_{j=1}^{\delta_i} f_1(y_j \mid x_i; \hat{\gamma}) O(\phi; x_i, y_j)/\hat{f}_1(y_i)}
\]

and

\[
\hat{f}_1(y) = n_R^{-1} \sum_{i=1}^{n} \delta_i f_1(y \mid x_i; \hat{\gamma}).
\]

May use EM algorithm to solve (16) for \( \phi \).
Exponential tilting method

- **Step 1:** Use the responding part of \((x_i, y_i)\), obtain \(\hat{\gamma}\) in the model \(f_1(y \mid x; \gamma)\).

\[
S_1(\gamma) \equiv \sum_{\delta_i=1} S_1(\gamma; x_i, y_i) = 0. \tag{17}
\]

- **Step 2:** Given \(\hat{\gamma}\) from Step 1, obtain \(\hat{\phi}\) by solving (16):

\[
S_2(\phi, \hat{\gamma}) = 0.
\]

- **Step 3:** Using \(\hat{\phi}\) computed from Step 2, the PSA estimator of \(\theta\) can be obtained by solving

\[
\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} U(\theta; x_i, y_i) = 0, \tag{18}
\]

where \(\hat{\pi}_i = \pi_i(\hat{\phi})\).
Remark

- In many cases, $x$ is categorical and $f_1(y \mid x)$ can be fully nonparametric.
- If $x$ has a continuous part, nonparametric Kernel smoothing can be used.
- The proposed method seems to be robust against the failure of the assumed model on $f_1(y \mid x; \gamma)$.
- Asymptotic normality of PSA estimator can be obtained & Linearization method can be used for variance estimation (Details skipped)
- By augmenting the estimating function, we can also impose a calibration constraint such as

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} x_i = \sum_{i=1}^{n} x_i.$$
Exponential tilting method

**Example 4 (Example 6.5 of KS)**

- Assume that both $x_i = (z_i, u_i)$ and $y_i$ are categorical with category $\{(i, j); i \in S_z \times S_u\}$ and $S_y$, respectively.
- We are interested in estimating $\theta_k = Pr(Y = k)$, for $k \in S_y$.
- Now, we have nonresponse in $y$ and let $\delta_i$ be the response indicator function for $y_i$. We assume that the response probability satisfies
  \[ Pr(\delta = 1 | x, y) = \pi(u, y; \phi). \]
- To estimate $\phi$, we first compute the observed conditional probability of $y$ among the respondents:
  \[ \hat{p}_1(y | x_i) = \frac{\sum_{\delta_j=1} l(x_j = x_i, y_j = y)}{\sum_{\delta_j=1} l(x_j = x_i)}. \]
The EM algorithm can be implemented by (16) with
\[
\bar{S}_0(\phi \mid x_i; \phi) = \frac{\sum_{\delta_j=1} S(\phi; \delta_i, u_i, y_j) \hat{p}_1(y_j \mid x_i) O(\phi; u_i, y_j) / \hat{p}_1(y_j)}{\sum_{\delta_j=1} \hat{p}_1(y_j \mid x_i) O(\phi; u_i, y_j) / \hat{p}_1(y_j)},
\]
where \( O(\phi; u, y) = \{1 - \pi(u, y; \phi)\} / \pi(u, y; \phi) \) and
\[
\hat{p}_1(y) = n_R^{-1} \sum_{i=1}^n \delta_i \hat{p}_1(y \mid x_i).
\]

Alternatively, we can use
\[
\bar{S}_0(\phi \mid x_i; \phi) = \frac{\sum_{y \in S_y} S(\phi; \delta_i, u_i, y) \hat{p}_1(y \mid x_i) O(\phi; u_i, y)}{\sum_{y \in S_y} \hat{p}_1(y \mid x_i) O(\phi; u_i, y)}.
\]
Example 4 (Cont’d)

• Once \( \hat{\pi}(u, y) = \pi(u, y; \hat{\phi}) \) is computed, we can use

\[
\hat{\theta}_{k, ET} = n^{-1} \left\{ \sum_{\delta_i=1} l(y_i = k) + \sum_{\delta_i=0} \sum_{y \in S_y} w_{iy}^* l(y = k) \right\},
\]

where \( w_{iy}^* \) is the fractional weights computed by

\[
w_{iy}^* = \frac{\{\hat{\pi}(u_i, y)^{-1} - 1\} \hat{p}_1(y|x_i)}{\sum_{y \in S_y} \{\hat{\pi}(u_i, y)^{-1} - 1\} \hat{p}_1(y|x_i)}.
\]
Exit Poll: The Assembly election (2012 Gang-dong district in Seoul)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Party A</th>
<th>Party B</th>
<th>Other</th>
<th>Refusal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>20-29</td>
<td>93</td>
<td>115</td>
<td>4</td>
<td>28</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>30-39</td>
<td>104</td>
<td>233</td>
<td>8</td>
<td>82</td>
<td>427</td>
</tr>
<tr>
<td></td>
<td>40-49</td>
<td>146</td>
<td>295</td>
<td>5</td>
<td>49</td>
<td>495</td>
</tr>
<tr>
<td></td>
<td>50-</td>
<td>560</td>
<td>350</td>
<td>3</td>
<td>174</td>
<td>1,087</td>
</tr>
<tr>
<td>Female</td>
<td>20-29</td>
<td>106</td>
<td>159</td>
<td>8</td>
<td>62</td>
<td>335</td>
</tr>
<tr>
<td></td>
<td>30-39</td>
<td>129</td>
<td>242</td>
<td>5</td>
<td>70</td>
<td>446</td>
</tr>
<tr>
<td></td>
<td>40-49</td>
<td>170</td>
<td>262</td>
<td>5</td>
<td>69</td>
<td>506</td>
</tr>
<tr>
<td></td>
<td>50-</td>
<td>501</td>
<td>218</td>
<td>7</td>
<td>211</td>
<td>937</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1,809</td>
<td>1,874</td>
<td>45</td>
<td>745</td>
<td>4,473</td>
</tr>
<tr>
<td>Truth</td>
<td></td>
<td>62,489</td>
<td>57,909</td>
<td>1,624</td>
<td></td>
<td>122,022</td>
</tr>
</tbody>
</table>
### Table: Analysis result: Gang-dong district in Seoul

<table>
<thead>
<tr>
<th>Method</th>
<th>Party A</th>
<th>Party B</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adjustment</td>
<td>48.5</td>
<td>50.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Adjustment (Age * Sex)</td>
<td>49.0</td>
<td>49.8</td>
<td>1.2</td>
</tr>
<tr>
<td>New Method</td>
<td>51.0</td>
<td>47.7</td>
<td>1.2</td>
</tr>
<tr>
<td>Truth</td>
<td>51.2</td>
<td>47.5</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Comparison of the methods (%)

Notes:
- Jae-Kwang Kim (ISU)
- Nonignorable missing
- July 4th, 2014
## Analysis result in Seoul (48 Seats)

<table>
<thead>
<tr>
<th>Method</th>
<th>Party A</th>
<th>Party B</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adjustment</td>
<td>10</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>Adjustment (Age* Sex)</td>
<td>10</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>New Method</td>
<td>15</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>Truth</td>
<td>16</td>
<td>30</td>
<td>2</td>
</tr>
</tbody>
</table>

*Jae-Kwang Kim (ISU)*

Nonignorable missing

July 4th, 2014
1 Introduction

2 Full likelihood-based ML estimation

3 GMM method

4 Partial Likelihood approach

5 Exponential tilting method

6 Concluding Remarks
Concluding remarks

• Uses a model for the response probability.
• Parameter estimation for response model can be implemented using the idea of maximum likelihood method.
• Instrumental variable needed for identifiability of the response model.
• Likelihood-based approach vs GMM approach
• Less tools for model diagnostics or model validation
• Promising areas of research


