# Likelihood Adjusted for Nonignorable Missing Covariate Values with Unspecified Propensity in Generalized Linear Models

Jun Shao

University of Wisconsin-Madison

Joint work with Fang Fang (East China Normal University) and Jiwei Zhao (SUNY-Buffalo)

November 13, 2015

# Outline

# Background

- Missing data is a common phenomenon in many applications in areas such as clinical trials, economics, sample surveys, and social sciences.
- **Missing Completely at Random (MCAR)**: The propensity of missing data is unrelated to any value, whether missing or observed.
- **Missing at Random (MAR)**: The propensity of missing data is unrelated to the missing values, but may be related to the observed values.
- Both MCAR and MAR are **ignorable** missing
  Solutions: Well-developed.

# Nonignorable Missing Data

- **Non-Ignorable Missing (NI)**: The propensity of missing data is related to the missing values, even after conditioning on all observed data.
- *Example: It commonly occurs when people do not want to reveal something very personal (such as income, age, weight, sexual preference, etc.).*
- Solutions: Difficult to handle and the solution is limited.

# The problem we consider

- Consider a GLM with **nonignorable** missing covariate values.
- $Y$: the response variable, $X = (U, Z)$: the covariate vector

  $p(Y|X,\beta) = \exp\{Y\eta - b(\eta) + c(Y)\}, \eta = \eta(\beta_c + \beta_u^\tau U + \beta_z^\tau Z)$
- $Y$ and $Z$ are fully observed, $U$ may have missing components
- $R$: the indicator of whether $U$ is fully observed.

$$P(R = 1|Y, U, Z) = P(R = 1|Y, U) \quad (1)$$

We call $Z$ an instrument variable.

# Existing methods

- For nonignorable missing, Robins and Ritov (1997) shows that either $P(R=1|Y,U)$ or $P(U|Z)$ has to be parametric.

- **Full Parametric Methods**: assume both $P(R=1|Y,U)$ and $P(U|Z)$ are parametric
  - Lipsitz et al. (1999, SIM)
  - Ibrahim et al. (1999, JRSSB)
  - Herring and Ibrahim (2002, Biostatistics)
  - Stubbendick and Ibrahim (2003, Biometrics; 2006, Sinica)
  - Huang et al. (2005, Biometrics)
  - Ibrahim and Molenberghs (2009, Test)

- **Semiparametric Pseudo Likelihood**: assume $P(U|Z)$ is parametric but $P(R=1|Y,U)$ is unspecified
  - Zhao and Shao (2015, JASA)

# The pseudo likelihood method

- By (1) and Bayes formula,

$$\begin{aligned} p(Z|Y,U,R=1) &= p(Z|Y,U) \\ &= \frac{p(Y|U,Z,\beta)p(U|Z,\gamma)p(Z)}{\int p(Y|U,z,\beta)p(U|z,\gamma)p(z)dz} \end{aligned} \quad (2)$$

- Pseudo likelihood estimator $(\tilde{\beta}, \tilde{\gamma})$ is the maximizer of

$$L(\beta', \gamma') = \prod_{i:r_i=1} \frac{p(y_i|u_i, z_i, \beta')p(u_i|z_i, \gamma')}{\sum_{j=1}^{N} p(y_i|u_i, z_j, \beta')p(u_i|z_j, \gamma')} \quad (3)$$

- Problems

  – When $\beta_z$ is close to 0, $(\tilde{\beta}_c, \tilde{\beta}_u)$ is very inefficient since its asymptotic variance diverges to infinity. However, $\tilde{\beta}_z$ is fine.

  – The pseudo likelihood (3) does not use any data from $(y_i, u_i, r_i = 0)$.

# The Proposed Method

We propose a two-stage method:

- Stage 1: Estimate $\gamma$ and $\beta_z$ based on the pseudo likelihood.
- Stage 2: Estimate $\beta$ by maximizing a likelihood adjusted for missing covariate values using the estimated $\gamma$ and $\beta_z$ from stage 1. To solve the likelihood equation, we propose an iterative algorithm.

# Adjusted likelihood for $\beta$

- If there is no missing data, the likelihood equation is

$$S(\beta') = \sum_{i=1}^{N} g(u_i, z_i, \beta') \left\{ y_i - h(\beta'_c + \beta'^{\tau}_u u_i + \beta'^{\tau}_z z_i) \right\} = 0,$$

where $h(\beta'_c + \beta'^{\tau}_u u_i + \beta'^{\tau}_z z_i) = \nabla_\eta b(\eta_i) = E(y_i | u_i, z_i)$ and $g(u_i, z_i, \beta') = \nabla_{\beta'} h / \nabla^2_{\eta\eta} b(\eta_i)$.

- Some components of $U$ may be always observed but not used as instruments. Let $U = (U_1, U_2)$, where $U_1$ may have missing values and $U_2$ is always observed.

- We consider an adjusted likelihood equation:

$$\sum_{i=1}^{N} g(u^*_{i1}, u_{i2}, z_i, \beta') \{ y_i - h(\beta'_c + \beta'^{\tau}_{u1} u^*_{i1} + \beta'^{\tau}_{u2} u_{i2} + \beta'^{\tau}_z z_i) \} = 0,$$

where $u^*_{i1}$ is a function of observed data.

# What $u_{i1}^*$ should we use?

- $u_{i1}^* = E(u_{i1}|u_{i2}, z_i)$ does not work since usually

$$h(\beta_c+\beta_{u1}^\tau u_{i1}^*+\beta_{u2}^\tau u_{i2}+\beta_z^\tau z_i) \neq E\{h(\beta_c+\beta_{u1}^\tau u_{i1}+\beta_{u2}^\tau u_{i2}+\beta_z^\tau z_i)|u_{i2}, z_i\}$$

- The above equation holds if

$$u_{i1}^*(\beta) = u_{i1}^{(0)}+\frac{\beta_{u1}}{\|\beta_{u1}\|^2}\left\{h^{-1}(\mu_i(\beta)) - \beta_c - \beta_{u1}^\tau u_{i1}^{(0)} - \beta_{u2}^\tau u_{i2} - \beta_z^\tau z_i\right\} \quad (4)$$

where $\mu_i(\beta)$ denotes the quantity on the right hand side of above equation, and $u_{i1}^{(0)} = E(u_{i1}|u_{i2}, z_i)$.

- This leads to the following valid likelihood equation:

$$\sum_{i=1}^N g(u_{i1}^*(\beta'), u_{i2}, z_i, \beta')\left\{y_i - h(\beta_c' + \beta_{u1}'^\tau u_{i1}^*(\beta') + \beta_{u2}'^\tau u_{i2} + \beta_z'^\tau z_i)\right\} = 0.$$

# An iterated algorithm

Denote

$$S(\beta'|\beta'') = \sum_{i=1}^{N} g(u_{i1}^*(\beta''), u_{i2}, z_i, \beta') \left\{ y_i - h(\beta_c' + \beta_{u1}'^{\tau} u_{i1}^*(\beta'') + \beta_{u2}'^{\tau} u_{i2} + \beta_z'^{\tau} z_i) \right\} = \tag{5}$$

**Algorithm:**

0. For each $i$, generate $\{u_{i1}^m, m = 1, \cdots, M\}$ from $p(u_{i1}|u_{i2}, z_i, \hat{\gamma})$.
1. At the $t$th iteration, compute $u_{i1}^*(\hat{\beta}^{(t)})$ according to (4) with $\beta = \hat{\beta}^{(t)}$, $u_{i1}^{(0)}$ replaced by $E(u_{i1}|u_{i2}, z_i, \hat{\gamma})$, and $\mu_i(\hat{\beta}^{(t)})$ approximated by

$$\mu_i^{(t)}(\hat{\beta}^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} h\left( \hat{\beta}_c^{(t)} + \hat{\beta}_{u1}^{(t)^{\tau}} u_{i1}^m + \hat{\beta}_{u2}^{(t)^{\tau}} u_{i2} + \hat{\beta}_z^{(t)^{\tau}} z_i \right).$$

2. Replace $u_{i1}^*(\beta'')$ in (5) by $u_{i1}^*(\hat{\beta}^{(t)})$ and compute $\hat{\beta}^{(t+1)}$ by solving $S(\beta'|\hat{\beta}^{(t)}) = 0$.
3. Execute 1-2 until $\{\hat{\beta}^{(t)}, t = 1, 2, \cdots\}$ converges to $\hat{\beta}$.

# Theoretical Results

**Theorem 1:** Under some regularity conditions, we have

(A) For any fixed $t$, if $\hat{\beta}^{(t)}$ converges to $\beta$ in probability, then its one-step update $\hat{\beta}^{(t+1)}$ also converges to $\beta$ in probability, as $N \to \infty$.

(B) If $\hat{\beta}^{(1)}$ converges to $\beta$ in probability as $N \to \infty$, then

$$P\left(\|\hat{\beta}^{(t)} - \hat{\beta}\| \geq \|\hat{\beta}^{(t+1)} - \hat{\beta}\| \text{ for all } t \text{ and } S(\hat{\beta}|\hat{\beta}) = 0\right) \to 1,$$

where $\hat{\beta} = \lim_{t\to\infty} \hat{\beta}^{(t)}$.

# Theoretical Results

**Theorem 2:** $\hat{\gamma}$ is consistent and asymptotically normal with an explicit influence function.

**Theorem 3:** $\hat{\beta}$ is consistent and asymptotically normal with an explicit asymptotical variance.

**Theorem 4:** The asymptotical variance estimator by substitution technique is consistent.

# Simulation studies

We first considered the following case (A):

(A) $Y$ is binary with $P(Y = 1|U, Z) = \text{expit}\{-1 - U + \beta_z Z\}$, $U$ and $Z$ are univariate, $U|Z \sim N(-1 + 2Z^2, 1)$, $Z \sim N(1, 1)$, $U$ has missing values, and $P(R = 1|Y, U) = \Phi(1 + U - Y)$, where $\Phi$ is the standard normal distribution function.

The percentages of complete data were around 76%. We considered the combination of $N = 300, 500, 1000$ and $\beta_z = 0, 2, 4$.

# Simulation results for case (A) with $N = 300$

| | method | parameter | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_c$=-1 | $\beta_u$=-1 | $\beta_z$=0 | $\beta_c$=-1 | $\beta_u$=-1 | $\beta_z$=2 | $\beta_c$=-1 | $\beta_u$=-1 | $\beta_z$=4 |
| BIAS | FULL | -0.055 | -0.051 | 0.045 | -0.049 | -0.025 | 0.082 | -0.034 | -0.029 | 0.116 |
| | CC | -0.741 | 0.242 | 0.095 | -0.744 | 0.072 | 0.348 | -0.697 | -0.058 | 0.580 |
| | PL | | not computed | | -0.163 | -0.326 | 0.193 | -0.136 | -0.122 | 0.249 |
| | AL | -0.138 | -0.163 | 0.115 | -0.080 | -0.067 | 0.176 | -0.029 | -0.044 | 0.164 |
| SD | FULL | 0.258 | 0.187 | 0.374 | 0.283 | 0.152 | 0.455 | 0.334 | 0.134 | 0.577 |
| | CC | 0.415 | 0.245 | 0.551 | 0.489 | 0.184 | 0.664 | 0.510 | 0.162 | 0.756 |
| | PL | | not computed | | 0.869 | 1.036 | 0.664 | 0.693 | 0.474 | 0.838 |
| | AL | 0.452 | 0.493 | 0.523 | 0.434 | 0.300 | 0.784 | 0.434 | 0.210 | 0.867 |
| SE | FULL | 0.247 | 0.179 | 0.359 | 0.279 | 0.145 | 0.430 | 0.317 | 0.128 | 0.549 |
| | CC | 0.377 | 0.259 | 0.518 | 0.447 | 0.182 | 0.612 | 0.492 | 0.155 | 0.725 |
| | PL | | not computed | | 0.838 | 0.772 | 0.620 | 0.657 | 0.417 | 0.856 |
| | AL | 0.412 | 0.436 | 0.502 | 0.405 | 0.270 | 0.711 | 0.413 | 0.193 | 0.805 |
| CP | FULL | 0.953 | 0.946 | 0.946 | 0.952 | 0.945 | 0.941 | 0.951 | 0.955 | 0.951 |
| | CC | 0.506 | 0.780 | 0.957 | 0.693 | 0.898 | 0.941 | 0.774 | 0.944 | 0.922 |
| | PL | | not computed | | 0.956 | 0.913 | 0.944 | 0.937 | 0.928 | 0.943 |
| | AL | 0.971 | 0.959 | 0.963 | 0.955 | 0.944 | 0.943 | 0.945 | 0.943 | 0.948 |

BIAS: bias of the estimator; SD: standard deviation; SE: estimated standard deviation; CP: coverage probability of 95% confidence interval

FULL: full data; CC: complete case; PL: pseudo likelihood; AL: proposed adjusted likelihood

# Comparison to MLE

We then considered the following case (B):

(B) $Y|U,Z \sim N(1+U+\beta_{z1}Z_1+\beta_{z2}Z_2,1)$ with univariate $U$ and two-dimensional $Z=(Z_1,Z_2)$, $U|Z \sim N(2+Z_1-2Z_2^2,1)$, $Z_1$ and $Z_2$ are independently $\sim N(1,1)$, and $P(R=1|Y,U)=\Phi(-2-U+|Y|)$.

We considered $N=500$ and $\beta_z=(\beta_{z1},\beta_{z2})=(0,0)$ or $(1,-2)$. The percentages of complete data were 50% and 60%, respectively.

We further consider some methods based on maximum likelihood estimation: MLE-C: MLE with a correct propensity model; MLE-W: MLE with a wrong propensity model; MLE-MAR: MLE assuming missing at random.

# Simulation results for case (B)

| | method | parameter | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_c$=1 | $\beta_u$=1 | $\beta_{z1}$=0 | $\beta_{z2}$=0 | $\beta_c$=1 | $\beta_u$=1 | $\beta_{z1}$=1 | $\beta_{z2}$=-2 |
| BIAS | FUILL | 0.002 | 0.001 | 0.000 | 0.000 | 0.006 | 0.000 | -0.002 | -0.003 |
| | CC | 0.543 | 0.018 | 0.059 | -0.186 | 0.207 | 0.014 | 0.027 | -0.056 |
| | PL | not computed | | | | 0.004 | -0.002 | 0.003 | -0.005 |
| | AL | 0.008 | 0.001 | 0.000 | 0.000 | 0.008 | 0.000 | -0.002 | -0.005 |
| | MLE-C | -0.001 | 0.001 | 0.000 | 0.000 | 0.006 | 0.000 | -0.002 | -0.004 |
| | MLE-W | -0.066 | -0.006 | 0.026 | -0.012 | -0.123 | -0.023 | 0.038 | -0.061 |
| | MLE-MAR | 0.311 | 0.018 | 0.014 | -0.054 | 0.165 | 0.017 | -0.012 | 0.006 |
| SD | FULL | 0.086 | 0.015 | 0.050 | 0.079 | 0.085 | 0.015 | 0.048 | 0.078 |
| | CC | 0.151 | 0.024 | 0.069 | 0.132 | 0.113 | 0.019 | 0.059 | 0.090 |
| | PL | not computed | | | | 0.125 | 0.025 | 0.071 | 0.112 |
| | AL | 0.131 | 0.022 | 0.069 | 0.115 | 0.119 | 0.023 | 0.065 | 0.110 |
| | MLE-C | 0.104 | 0.018 | 0.058 | 0.089 | 0.097 | 0.017 | 0.053 | 0.082 |
| | MLE-W | 0.111 | 0.018 | 0.059 | 0.091 | 0.099 | 0.018 | 0.053 | 0.086 |
| | MLE-MAR | 0.106 | 0.018 | 0.059 | 0.092 | 0.096 | 0.017 | 0.054 | 0.082 |
| SE | FULL | 0.084 | 0.015 | 0.047 | 0.076 | 0.084 | 0.015 | 0.048 | 0.075 |
| | CC | 0.138 | 0.020 | 0.068 | 0.112 | 0.109 | 0.018 | 0.059 | 0.086 |
| | PL | not computed | | | | 0.118 | 0.023 | 0.070 | 0.101 |
| | AL | 0.129 | 0.022 | 0.066 | 0.109 | 0.119 | 0.023 | 0.063 | 0.109 |
| | MLE-C | 0.104 | 0.017 | 0.055 | 0.085 | 0.094 | 0.017 | 0.052 | 0.079 |
| | MLE-W | 0.111 | 0.018 | 0.056 | 0.087 | 0.098 | 0.018 | 0.053 | 0.083 |
| | MLE-MAR | 0.105 | 0.018 | 0.057 | 0.088 | 0.094 | 0.017 | 0.053 | 0.079 |
| CP | FULL | 0.946 | 0.950 | 0.942 | 0.934 | 0.946 | 0.955 | 0.948 | 0.935 |
| | CC | 0.014 | 0.786 | 0.862 | 0.614 | 0.523 | 0.883 | 0.930 | 0.890 |
| | PL | not computed | | | | 0.929 | 0.916 | 0.938 | 0.922 |
| | AL | 0.942 | 0.944 | 0.936 | 0.942 | 0.948 | 0.941 | 0.930 | 0.944 |
| | MLE-C | 0.954 | 0.942 | 0.938 | 0.912 | 0.943 | 0.939 | 0.935 | 0.929 |
| | MLE-W | 0.884 | 0.934 | 0.910 | 0.918 | 0.744 | 0.735 | 0.887 | 0.863 |
| | MLE-MAR | 0.150 | 0.794 | 0.932 | 0.892 | 0.574 | 0.826 | 0.931 | 0.934 |

# NHANES data analysis

- We analyzed a data set from the National Health and Nutrition Examination Survey (NHANES 2005), which is designed to assess the health and nutritional status of adults and children in the United States.
- $Y$: indicator of hypertension
- $X$: $age$, $gender$, $dxa$: body fat measured by Dual-energy x-ray absorptiometry, $bmi$: body mass index

$$P(Y=1|dxa, age, gender, bmi) = \text{logit}\{\beta_1+\beta_2 dxa+\beta_3 age+\beta_4 gender\}.$$

- $U_1$: $dxa$
- Consider five options of $Z = age$, $Z = bmi$, $Z = (age, gender)$, $Z = (age, bmi)$, and $Z = (gender, bmi)$.
- $U_2$: components of $(age, gender, bmi)$ that are not in $Z$.

# NHANES data analysis results

| effect | method | instrument $Z$ | estimate | standard error | p-value |
|---|---|---|---|---|---|
| intercept | AL | age | -5.5091 | 0.6834 | 0.000 |
| | | age,gender | -5.5269 | 0.6644 | 0.000 |
| | | age,bmi | -5.4321 | 0.6844 | 0.000 |
| | | bmi | -5.4511 | 0.6251 | 0.000 |
| | | gender,bmi | -5.5045 | 0.6127 | 0.000 |
| | PL | age | -5.9139 | 5.2883 | 0.263 |
| | | age,gender | -2.8827 | 7.8250 | 0.713 |
| | | age,bmi | -0.8004 | 8.2188 | 0.922 |
| | | gender,bmi | -12.236 | 272.92 | 0.964 |
| | CC | | -5.3175 | 0.7156 | 0.000 |
| | MLE-MAR | | -5.0807 | 0.6044 | 0.000 |
| dxa | AL | age | 0.0394 | 0.0150 | 0.009 |
| | | age,gender | 0.0347 | 0.0125 | 0.006 |
| | | age,bmi | 0.0314 | 0.0102 | 0.002 |
| | | bmi | 0.0313 | 0.0117 | 0.007 |
| | | gender,bmi | 0.0333 | 0.0117 | 0.004 |
| | PL | age | 0.0003 | 0.0261 | 0.992 |
| | | age,gender | -0.0571 | 0.1802 | 0.751 |
| | | age,bmi | -0.0999 | 0.2147 | 0.642 |
| | | gender,bmi | 0.0030 | 5.9060 | 0.999 |
| | CC | | 0.0076 | 0.0126 | 0.549 |
| | MLE-MAR | | 0.0223 | 0.0102 | 0.028 |

# NHANES data analysis results

| effect | method | instrument $Z$ | estimate | standard error | p-value |
|---|---|---|---|---|---|
| age | AL | age | 0.0478 | 0.0094 | 0.000 |
| | | age,gender | 0.0521 | 0.0099 | 0.000 |
| | | age,bmi | 0.0530 | 0.0088 | 0.000 |
| | | bmi | 0.0534 | 0.0076 | 0.000 |
| | | gender,bmi | 0.0528 | 0.0081 | 0.000 |
| | PL | age | 0.0702 | 0.0098 | * |
| | | age,gender | 0.0707 | 0.0125 | 0.000 |
| | | age,bmi | 0.0676 | 0.0139 | * |
| | | gender,bmi | 0.0146 | 6.6283 | 0.998 |
| | CC | | 0.0674 | 0.0100 | 0.000 |
| | MLE-MAR | | 0.0538 | 0.0080 | 0.000 |
| gender | AL | age | 0.4064 | 0.2359 | 0.085 |
| | | age,gender | 0.3785 | 0.1890 | 0.045 |
| | | age,bmi | 0.2963 | 0.1755 | 0.091 |
| | | bmi | 0.2957 | 0.1872 | 0.114 |
| | | gender,bmi | 0.3352 | 0.1922 | 0.081 |
| | PL | age | 0.3122 | 0.7247 | 0.667 |
| | | age,gender | -0.0370 | 0.3754 | 0.992 |
| | | age,bmi | -1.4205 | 3.6459 | 0.697 |
| | | gender,bmi | -0.1819 | 0.2298 | * |
| | CC | | 0.0593 | 0.2094 | 0.777 |
| | MLE-MAR | | 0.1821 | 0.1780 | 0.306 |

*not available

# Concluding Remarks

- We propose a novel approach to handle generalized linear models with nonignorable missing covariate data without any parametric model on the propensity.
- The pseudo likelihood method only works when an appropriate instrument is available. Our proposed method also needs to specify an instrument but is more flexible in choosing it.
- The proposed method is more stable and usually more efficient than the pseudo likelihood method.
- The proposed method needs a correct parametric model on $p(U|Z,\gamma)$. It is almost unavoidable since we put no parametric assumption on the propensity.

# The End

**Thank you.**

**Questions or Comments?**