The role of callback in survey data for nonresponse adjustment

(In-class presentation)

Outline

1. Paradata, callback data, and missing data

2. Real-world example

3. Identification and semiparametric inference

4. Extensions

5. Simulations and application

6. Discussion

Outline

1. Paradata, callback data, and missing data

2. Real-world example

3. Identification and semiparametric inference

4. Extensions

5. Simulations and application

6. Discussion

Paradata

Paradata are the records tracking the collection process of survey data (Couper, 1998).

Table 1: Data structure of a sampling survey

Questionnaire data				Paradata
ID	X_1	$\cdots$	X_p	Process of data collection
1	x_{11}	$\cdots$	x_{1p}	date, time, length,
2	x_{21}	$\cdots$	x_{2p}	mode of communication,
$\vdots$	$\vdots$	$\vdots$	$\vdots$	attitude of the respondent,
$\vdots$	$\vdots$	$\vdots$	$\vdots$	record of contacts,
n	x_{n1}	$\cdots$	x_{np}	gender of interviewer
Survey data				

Paradata are widely available in modern surveys, e.g., U.S. National Health Interview Survey (NHIS), British Survey of Social Attitudes (BSSA), and the European Social Survey (ESS).

Missing data and callback data

Callback data: a traditional form of paradata.

The frame data are often prone to missingness.

In many surveys interviewers continue to contact nonrespondents and the contact attempts are recorded. Sometimes called level-of-effort data (Biemer et al., 2013).

Table 2: Data structure of a sampling survey with callbacks

Frame/Questionnaire data				Contact attempts			
ID	X	Y	R	R_1	R_2	$\dots$	R_K
1	x_1	y_1	1	1	1	$\cdots$	1
2	x_2	y_2	1	0	1	$\cdots$	1
:	:	NA	0	0	0	$\cdots$	0
:	:	NA	0	0	0	$\cdots$	0
n	x_n	y_n	1	0	0	$\cdots$	1

Use of callback data in social sciences

Appropriate use of callback data can improve survey quality and compensate for deficiencies in surveys.

- Callbacks have routinely been used to monitor response rates and to study how design features affect contact and cooperation in the course of data collection (Bates, 2003; Groves & Couper, 1998); e.g., calls made during weekday evenings and on weekends are more likely to be responded.
- Kreuter (2013) provides a comprehensive literature review on the use of paradata in analyses of survey data.
- Olson (2013) reviewed categories of paradata and challenges and opportunities in using paradata for nonresponse adjustment.

Missing data analysis

Missingness mechanisms (Rubin, 1976; Little & Rubin, 2002)

- Missing at random (MAR) $R \perp\!\!\!\perp Y \mid X$;
- Missing not at random (MNAR) $R \not\!\perp\!\!\!\perp Y \mid X$;

A large body of the missing data analysis literature rests on missingness at random (MAR) or ignorability. Parametric approaches: Likelihood-based inference; Imputation. Semiparametric approaches: Regression-based estimation (REG); Inverse probability weighting (IPW); Doubly robust estimation (DR).

The biggest challenge for MNAR is identification: the joint distribution is not uniquely determined from the observed data distribution.

Missing data analysis

Strategies to achieve identification:

- restrictive parametric models, e.g., Heckman (1979)'s selection model; counterexamples, the normal-logistic model (Wang et al., 2014; Miao et al., 2016);
- instrumental variable approach (Manski, 1985; Newey, 2009; Das et al., 2003; Sun et al., 2018; Liu et al., 2020; Tchetgen Tchetgen & Wirth, 2017),
- shadow variable approach (D'Haultfœuille, 2010; Miao & Tchetgen Tchetgen, 2016; Wang et al., 2014; Zhao & Shao, 2015; Kott, 2014).

Missing data analysis

Researchers have traditionally sought auxiliary variables from the sampling frame, but it is surprisingly difficult for practitioners and the difficulty is amplified in multipurpose studies where multiple survey variables are concerned and multiple auxiliary variables are necessitated.

Moreover, instrumental and shadow variable approaches invoke additional no interaction or completeness conditions that further limit their use.

Lastly, they break down if the auxiliary variables also have missing values, e.g., due to failure of contact in surveys.

In contrast, callback data offer an important source of auxiliary information for nonresponse adjustment. Follow ups are commonly made in many surveys to increase the response rate, the contact process are recorded by interviewers and are often kept for all units.

Using callback data for missing data analysis

Callback data had not been used widely in statistical analysis until recently.

- The early idea of Politz & Simmons (1949): use the number of nights that a respondent had been at home during the past week to account for the "not at homes"by weighting;
- The "continuum of resistance" model: nonrespondents are more similar to delayed respondents than they are to early respondents, so that the most reluctant respondents can be used to approximate the nonrespondents. (Lin & Schaeffer, 1995; Groves & Couper, 1998; Kreuter et al., 2010; Little, 1982).
- Model the joint likelihood of the callbacks and the frame variables; require untenable assumptions to achieve identification, e.g., assume the response probabilities are equal across different attempts or levels of the frame variables (Biemer et al., 2013; Drew & Fuller, 1980).
- Chen et al. (2018); Zhang et al. (2018) generalize the Heckman (1979) Selection model to incorporate callbacks to improve efficiency.
- Daniels et al. (2015) advocate the use of pattern mixture models for sensitivity analysis.

Using callback data for missing data analysis

Most notably, Alho (1990); Kim & Im (2014); Qin & Follmann (2014); Guan et al. (2018) employ propensity score models to make nonresponse adjustment with callbacks and propose inverse probability weighted and empirical likelihood-based estimators.

Their model:

$$\begin{aligned}&\text{logit } f(R_k = 1 \mid R_{k-1} = 0, X, Y) = \alpha_{k0} + \alpha_{k1} X + \gamma_k Y \\ &\text{with } \alpha_{k1} = \alpha_1, \quad \gamma_k = \gamma \text{ for } k = 1, 2,\end{aligned}$$

So far, identification of semiparametric and nonparametric propensity score models with callbacks is not available.

Using callback data for missing data analysis

In contrast, we consider a fundamentally nonparametric identification strategy:

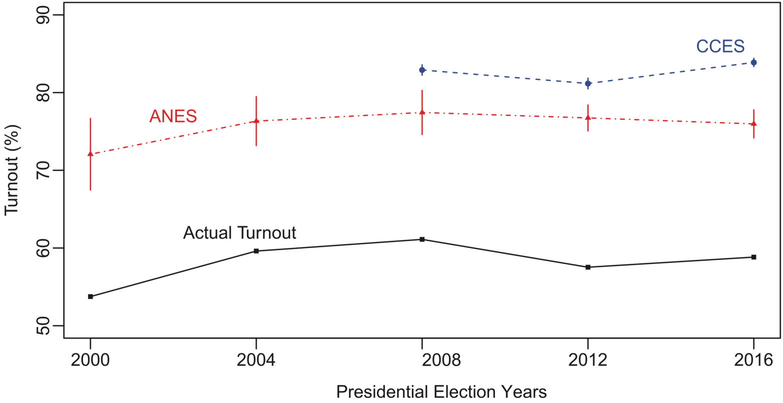
- we propose an identifying assumption that allows for nonparametric and nonlinear propensity score models,
- establish the semiparametric theory
- and propose a suite of semiparametric estimators including doubly robust ones.

Outline

1. Paradata, callback data, and missing data
2. Real-world example
3. Identification and semiparametric inference
4. Extensions
5. Simulations and application
6. Discussion

Overestimation of turnout

For decades, overestimation of turnout has been a classic problem in election surveys, and researchers have struggled with how to adjust for turnout bias.



Source of overestimation bias

- Nonresponse bias (voter overrepresentation):
  - those who do not vote are less likely to answer a political questionnaire → MNAR
- Measurement bias (voter overreporting):
  - the interviewer/respondent interaction in surveys makes respondents susceptible to social desirability response bias

ANES NRFU Survey on 2020 U.S. presidential election

Mailed study including 8,000 addresses:

- began on January 28, 2021 with an advance postcard
- the first class invitation was mailed on February 1
- followed by a first class replacement questionnaire on March 2
- and a second replacement questionnaire on March 30
- a reminder postcard was sent on February 16, and another one on April 5.

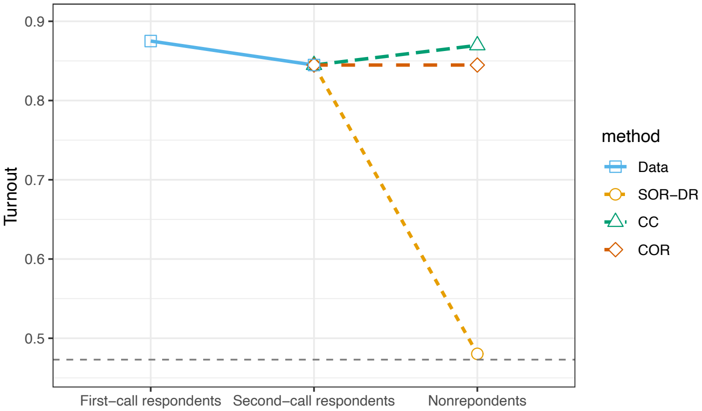
ANES NRFU Survey on 2020 U.S. presidential election

- The number of completed questionnaires is 3,779 which gives the unweighted response rate of 48.3%, and the weighted NRFU response rate of 56.6%.

- Whether vote in the 2020 presidential election is asked in three versions of the questionnaire. The voting-age population (VAP) and voting-eligible population (VEP) turnout for the 2020 presidential election is 61.5% and 66.2%, respectively. Because the ANES sampling framework excludes noncitizens, we use the VEP rate as an approximation to the true turnout of ANES population.

- The weighted voter turnout based on the respondents is over 85%, which is much higher than the VEP or VAP rate, indicating a severe overestimation bias.

Can we obtain a good estimate?



Outline

1. Paradata, callback data, and missing data

2. Real-world example

3. Identification and semiparametric inference

4. Extensions

5. Simulations and application

6. Discussion

Identification

Notation: Frame variables (X, Y), X fully observed covariates, Y the outcome prone to missing values.

Callback data $R_k, k = 1, 2$ the response state of Y with $R_k = 1$ if Y is available in the kth call and $R_k = 0$ otherwise. $\Rightarrow R_2 \geq R_1$.

Define the odds ratio functions for the response propensity in the the first and second calls as follows,

$$\begin{aligned}
\Gamma_1(X, Y) &= \log \frac{f(R_1 = 1 \mid X, Y) f(R_1 = 0 \mid X, Y = 0)}{f(R_1 = 0 \mid X, Y) f(R_1 = 1 \mid X, Y = 0)}, \\
\Gamma_2(X, Y) &= \log \frac{f(R_2 = 1 \mid R_1 = 0, X, Y) f(R_2 = 0 \mid R_1 = 0, X, Y = 0)}{f(R_2 = 0 \mid R_1 = 0, X, Y) f(R_2 = 1 \mid R_1 = 0, X, Y = 0)}.
\end{aligned}$$

Γ_k measures the impact of the missing outcome on the response propensity, the degree of nonignorable missingness, the resistance to respond caused by the outcome.

Identification

Assumption 1.

(i) Callback: $R_2 \geq R_1$;

(ii) Positivity: $0 < f(R_1 = 1 \mid X, Y) < 1$ and $0 < f(R_2 = 1 \mid R_1 = 0, X, Y) < 1$ for all (X, Y);

(iii) Stableness of resistance: $\Gamma_1(X, Y) = \Gamma_2(X, Y) = \Gamma(X, Y)$.

Theorem 1. Under Assumption 1, $f(X, Y, R_1, R_2)$ is identified.

No parametric models for the propensity scores or restrictions on the effects of covariates are imposed.

Identification

An immediate application to the linear logistic model.

Proposition 1. Assuming that $\text{logit } \pi_k(X, Y) = \text{logit } f(R_k = 1 \mid R_{k-1} = 0, X, Y) = \alpha_{k0} + \alpha_{k1}X + \gamma Y$, then α_{k0}, α_{k1}, and γ are identified.

Alho (1990); Kim & Im (2014); Guan et al. (2018) have to assume that $\alpha_{k1} = \alpha_1$.

Parameterization

We focus on estimation of the outcome mean $\mu = E(Y)$.

$$\begin{aligned} f(Y, R_1, R_2 \mid X) \quad = \quad & c_1(X) \cdot f(R_1 \mid X, Y = 0) \cdot \exp\{(R_1 - 1)\Gamma(X, Y)\} \\ & \cdot f(Y \mid R_2 = 1, R_1 = 0, X) \\ & \cdot f(R_2 \mid R_1 = 0, X, Y)^{1-R_1} f(R_2 = 0 \mid R_1 = 0, X, Y)^{-1} \end{aligned}$$

The baseline propensity scores
$A_1(X) = f(R_1 = 1 \mid X, Y = 0)$,
$A_2(X) = f(R_2 = 1 \mid R_1 = 0, X, Y = 0)$,
the odds ratio function $\Gamma(X, Y)$,
and the second-call outcome distribution
$f(Y \mid R_2 = 1, R_1 = 0, X)$.

Can be parameterized separately without hindering congeniality.

Hereafter, let $\pi_1(X, Y) = f(R_1 = 1 \mid X, Y)$, $\pi_2(X, Y) = f(R_2 = 1 \mid R_1 = 0, X, Y)$, $f_2(Y \mid X) = f(Y \mid R_2 = 1, R_1 = 0, X)$.

IPW estimation

An inverse probability weighted estimator of the outcome mean is

$$\hat{\mu}_{\mathrm{ipw}} = \hat{E}\left\{\frac{R_2}{\hat{p}_2} Y\right\}.$$

$\hat{p}_2 = \hat{\pi}_1 + \hat{\pi}_2(1 - \hat{\pi}_1)$ is an estimator of $f(R_2 = 1 \mid X, Y)$.

IPW estimation

An inverse probability weighted estimator of the outcome mean is

$$\hat{\mu}_{\text{ipw}} = \hat{E}\left\{\frac{R_2}{\hat{p}_2}Y\right\}.$$

$\hat{p}_2 = \hat{\pi}_1 + \hat{\pi}_2(1 - \hat{\pi}_1)$ is an estimator of $f(R_2 = 1 \mid X, Y)$.

Characterization of propensity scores

$$\begin{aligned} 0 &= E\left\{\frac{R_1}{\pi_1} - 1 \mid X, Y\right\}, \\ 0 &= E\left\{\frac{R_2 - R_1}{\pi_2} - (1 - R_1) \mid X, Y\right\}, \Leftrightarrow \\ 0 &= E\left\{\frac{R_2}{\pi_2} - 1 \mid R_1 = 0, X, Y\right\}. \end{aligned}$$

IPW estimation

An inverse probability weighted estimator of the outcome mean is

$$\hat{\mu}_{\mathrm{ipw}} = \hat{E}\left\{\frac{R_2}{\hat{p}_2}Y\right\}.$$

$\hat{p}_2 = \hat{\pi}_1 + \hat{\pi}_2(1-\hat{\pi}_1)$ is an estimator of $f(R_2 = 1 \mid X, Y)$.

Specify $\pi_1(\alpha_1,\gamma), \pi_2(\alpha_2,\gamma) \Leftrightarrow \{A_1(X;\alpha_1), A_2(X;\alpha_2), \Gamma(X,Y;\gamma)\}$, solve

$$\begin{aligned}
0 &= \hat{E}\left[\left\{\frac{R_1}{\pi_1(\alpha_1,\gamma)} - 1\right\}\cdot V_1(X)\right], && (1)\\
0 &= \hat{E}\left[\left\{\frac{R_2 - R_1}{\pi_2(\alpha_2,\gamma)} - (1-R_1)\right\}\cdot V_2(X)\right], && (2)\\
0 &= \hat{E}\left[\left\{\frac{R_2 - R_1}{\pi_2(\alpha_2,\gamma)} - \frac{1-\pi_1(\alpha_1,\gamma)}{\pi_1(\alpha_1,\gamma)}R_1\right\}\cdot U(X,Y)\right], && (3)
\end{aligned}$$

A natural choice $V_1(X) = \partial A_1(X;\alpha_1)/\partial\alpha_1, V_2(X) = \partial A_2(X;\alpha_2)/\partial\alpha_2$, $U(X,Y) = \partial\Gamma(X,Y;\gamma)/\partial\gamma$.

IPW estimation

Assuming the common-slope logistic model

$$\text{logit } \pi_k(X,Y) = \alpha_{k0} + \alpha_1 X + \gamma Y$$

for $k = 1, 2$, Kim & Im (2014) estimate the parameters by solving

$$\begin{aligned} 0 &= \hat{E}\left[\left\{\frac{R_1}{\pi_1(\alpha_{10},\alpha_1,\gamma)} - 1\right\}\cdot 1\right], \\ 0 &= \hat{E}\left[\left\{\frac{R_2 - R_1}{\pi_2(\alpha_{20},\alpha_1,\gamma)} - \frac{1-\pi_1(\alpha_{10},\alpha_1,\gamma)}{\pi_1(\alpha_{10},\alpha_1,\gamma)} R_1\right\}\cdot (1, X^{\mathrm{T}}, Y)^{\mathrm{T}}\right], \end{aligned}$$

which can be obtained by letting $V_1(X) = 1$ and $U(X,Y) = (1, X^{\mathrm{T}}, Y)^{\mathrm{T}}$, respectively.

However, if $\text{logit } \pi_k(X,Y) = \alpha_{k0} + \alpha_{k1} X + \gamma Y$, then the above two equations fail.

In contrast, we can still consistently estimate the parameters with $V_1(X) = V_2(X) = (1, X^{\mathrm{T}})^{\mathrm{T}}$ and $U(X,Y) = Y$, for example.

The proposed IPW estimation is a generalization of the calibration estimator of Kim & Im (2014).

Outcome regression-based estimation

An outcome regression-based estimator of the outcome mean is

$$\hat{\mu}_{\mathrm{reg}} = \hat{E}\{R_2 Y + (1 - R_2) E(Y \mid X, R_2 = 0)\}, \tag{4}$$

impute the missing outcome values with $E(Y \mid X, R_2 = 0)$.

$$E(Y \mid X, R_2 = 0) = \frac{E\{e^{-\Gamma} Y \mid R_2 = 1, R_1 = 0, X\}}{E\{e^{-\Gamma} \mid R_2 = 1, R_1 = 0, X\}},$$

Specify $\{\pi_1(\alpha_1, \gamma), f_2(Y \mid X; \beta)\}$,

$$\begin{aligned} 0 &= E\left\{(R_2 - R_1) \cdot \frac{\partial \log f_2(Y \mid X; \beta)}{\partial \beta}\right\}, \\ 0 &= E\left[(1 - R_2)\{U(X, Y) - E\{U(X, Y) \mid X, R_2 = 0; \beta, \gamma\}\}\right]. \end{aligned}$$

Outcome regression-based estimation

An outcome regression-based estimator of the outcome mean is

$$\hat{\mu}_{\text{reg}} = \hat{E}\{R_2 Y + (1 - R_2) E(Y \mid X, R_2 = 0)\}, \tag{5}$$

impute the missing outcome values with $E(Y \mid X, R_2 = 0)$.

$$E(Y \mid X, R_2 = 0) = \frac{E\{e^{-\Gamma} Y \mid R_2 = 1, R_1 = 0, X\}}{E\{e^{-\Gamma} \mid R_2 = 1, R_1 = 0, X\}},$$

Specify $\{\pi_1(\alpha_1, \gamma), f_2(Y \mid X; \beta)\}$ and solve

$$\begin{aligned} 0 &= \hat{E}\left\{R_2(1 - R_1) \cdot \frac{\partial \log f_2(Y \mid X; \beta)}{\partial \beta}\right\}, \\ 0 &= \hat{E}\left[\left\{\frac{R_1}{\pi_1(\alpha_1, \gamma)} - R_2\right\} U(X, Y) - (1 - R_2) E\{U(X, Y) \mid X, R_2 = 0; \beta, \gamma \end{aligned}$$

where $U(X, Y)$ is a user-specified function with dimension equal to that of (α_1, γ).

Semiparametric efficiency theory

The stableness of resistance assumption defines the model

$$\mathcal{M}_{\mathrm{npr}} = \left\{ f(X, Y, R_1, R_2) : \begin{array}{l} \text{Assumption 1 holds;} \\ A_1, A_2, \Gamma, f_2 \text{ are unrestricted.} \end{array} \right\}$$

Proposition 2. The observed-data tangent space for $\mathcal{M}_{\mathrm{npr}}$ is

$$\mathcal{T} = \{h(O) = R_1 h_1(x, y) + R_2 h_2(x, y) + h_3(x); E\{h(O)\} = 0, E\{h^2(O)\} < \infty$$

where $h_1(x, y)$, $h_2(x, y)$ and $h_3(x)$ are arbitrary measurable and square-integrable functions.

The observed-data tangent space is the entire Hilbert space of observed-data functions with mean zero and finite variance, equipped with the usual inner product.

Proof of Proposition 2

The observed-data likelihood for a single observation can be written as

$$\begin{aligned} f(O) &= f(X)f(Y, R_1 = 1 \mid X)^{R_1} f(R_2 = 1, Y \mid R_1 = 0, X)^{(R_2 - R_1)} \\ &\quad \cdot f(R_2 = 0 \mid R_1 = 0, X)^{1-R_2} f(R_1 = 0 \mid X)^{1-R_1} \\ &= f(X)\{f(Y \mid R_1 = 1, X)f(R_1 = 1 \mid X)\}^{R_1} \\ &\quad \cdot \{f(Y \mid R_1 = 0, X)f(R_2 = 1 \mid R_1 = 0, X, Y)\}^{(R_2 - R_1)} \\ &\quad \cdot f(R_2 = 0 \mid R_1 = 0, X)^{1-R_2} f(R_1 = 0 \mid X)^{1-R_1}. \end{aligned}$$

We write $A_1(X), A_2(X), \Gamma(X, Y), f_1(Y \mid X) = f(Y \mid R_1 = 1, X)$ as A_1, A_2, Γ, f_1 for short where it does not cause confusion.

Proof of Proposition 2

Then the observed-data likelihood can be written as:

$$\begin{aligned} f(O) &= f(X)\left[f_1 \frac{e^{A_1}}{e^{A_1}+\int e^{-\Gamma} f_1 dY}\right]^{R_1}\left[\frac{e^{A_2+\Gamma}}{1+e^{A_2+\Gamma}} \frac{e^{-\Gamma} f_1}{\int e^{-\Gamma} f_1 dY}\right]^{R_2-R_1} \\ & \cdot\left[\frac{\int \frac{e^{-\Gamma}}{1+e^{A_2+\Gamma}} f_1 dY}{\int e^{-\Gamma} f_1 dY}\right]^{1-R_2}\left[\frac{\int e^{-\Gamma} f_1 dY}{e^{A_1}+\int e^{-\Gamma} f_1 dY}\right]^{1-R_1} . \end{aligned}$$

and the log-likelihood is

$$\begin{aligned} \log f(O) &= \log f(X)+R_2 \log f_1-\log \left(e^{A_1}+\int e^{-\Gamma} f_1 dY\right) \\ & +R_1 A_1-(R_2-R_1) \log(1+e^{-A_2-\Gamma})-(R_2-R_1)\Gamma \\ & +(1-R_2) \log \left(\int \frac{e^{-\Gamma} f_1}{1+e^{A_2+\Gamma}} dY\right) . \end{aligned}$$

Proof of Proposition 2

Consider a regular parametric submodel $f_t(X, Y, R_1, R_2)$ indexed by t with $f_{t=0}(X, Y, R_1, R_2) = f(X, Y, R_1, R_2)$, and let $f_t(X), f_{1,t}, A_{1,t}, A_{2,t}, \Gamma_t$ be the corresponding parameterization. Letting $S_1 = \partial \log f_{1,t} / \partial t$, $S_X = \partial \log f_t(X) / \partial t$ are the scores of f_1 and $f(X)$, $\{A_1, A_2, \Gamma\}$ are pathwise derivatives of $A_{1,t}$, $A_{2,t}$, Γ_t evaluated at $t = 0$, the observed-data score in this submodel, i.e., the score of $f_t(O)$ is

$$
\begin{aligned}
S(O) \quad = \quad & R_2 \left[S_1 - \pi_2(A_2 + \Gamma) + A_2 - E\{S_1 - \Gamma - \pi_2(A_2 + \Gamma) \mid R_2 = 0, X\} \right] \\
& + R_1\{A_1 - A_2 + \pi_2(A_2 + \Gamma)\} \\
& + S_X - f(R_1 = 1 \mid X) A_1 - f(R_1 = 0 \mid X) \cdot E\{S_1 - \Gamma \mid R_1 = 0, X\} \\
& + E\{S_1 - \Gamma - \pi_2(A_2 + \Gamma) \mid R_2 = 0, X\}.
\end{aligned}
$$

The tangent space has a complicated form; nonetheless, one can show that it is in fact equal to the entire Hilbert space of observed-data functions with mean zero and finite variance, equipped with the usual inner product.

Semiparametric efficiency theory

Theorem 2. The efficient influence function for μ in the nonparametric model $\mathcal{M}_{\mathrm{npr}}$ is

$$\begin{aligned}\mathrm{IF}(O;\mu) \quad = \quad & \left\{\frac{R_1}{\pi_1} - \frac{R_1}{\pi_2}\frac{1-\pi_1}{\pi_1} + \frac{R_2 - R_1}{\pi_2^2}\right\} Y \\ & - \left\{\frac{R_1}{\pi_1} - \frac{R_1}{\pi_2}\frac{1-\pi_1}{\pi_1} + \frac{R_2 - R_1}{\pi_2^2} - 1\right\} \frac{E(Y/\pi_2 \mid X, R_2 = 0)}{E(1/\pi_2 \mid X, R_2 = 0)} \\ & -\mu.\end{aligned}$$

Semiparametric efficiency theory

Under MAR, without callback data,

$$\mathrm{IF}_{\mathrm{aipw}}(O;\mu) = \frac{R_2}{f(R_2=1\mid X)}Y - \left\{\frac{R_2}{f(R_2=1\mid X)} - 1\right\}E(Y\mid R_2=1,X) - \mu.$$

Under MAR, with callback data, $\mathcal{M}_{\mathrm{npr}}^{\mathrm{mar}} = \{\mathcal{M}_{\mathrm{npr}} : \Gamma = 0\}$,

Proposition 3. In $\mathcal{M}_{\mathrm{npr}}^{\mathrm{mar}}$, the efficient influence function for μ is still $\mathrm{IF}_{\mathrm{aipw}}(O;\mu)$, which is more efficient than $\mathrm{IF}(O;\mu)$.

Under MAR, $\mathrm{IF}(O;\mu)$ does not reduce to $\mathrm{IF}_{\mathrm{aipw}}(O;\mu)$.

The efficiency bound does not change in the presence of callback data under MAR.

Doubly robust estimation

An estimator of μ motivated by the efficient influence function is

$$\begin{aligned}\hat{\mu}_{\mathrm{dr}} &= \hat{E}\left[\left\{\frac{R_1}{\hat{\pi}_1} - \frac{R_1}{\hat{\pi}_2}\frac{1-\hat{\pi}_1}{\hat{\pi}_1} + \frac{R_2 - R_1}{\hat{\pi}_2^2}\right\} Y\right] \\ &\quad -\hat{E}\left[\left\{\frac{R_1}{\hat{\pi}_1} - \frac{R_1}{\hat{\pi}_2}\frac{1-\hat{\pi}_1}{\hat{\pi}_1} + \frac{R_2 - R_1}{\hat{\pi}_2^2} - 1\right\} \frac{E(Y/\hat{\pi}_2 \mid X, R_2 = 0; \hat{\beta}_{\mathrm{dr}}, \gamma_{\mathrm{dr}})}{E(1/\hat{\pi}_2 \mid X, R_2 = 0; \hat{\beta}_{\mathrm{dr}}, \gamma_{\mathrm{dr}})}\right].\end{aligned}$$

where the nuisance parameters $(\hat{\alpha}_{1,\mathrm{dr}}, \hat{\alpha}_{2,\mathrm{dr}}, \hat{\beta}, \hat{\gamma}_{\mathrm{dr}})$ are obtained by solving

$$\begin{aligned} 0 &= \hat{E}\left\{R_2(1-R_1)\cdot\frac{\partial \log f_2(Y \mid X; \beta)}{\partial \beta}\right\}, \\ 0 &= \hat{E}\left[\left\{\frac{R_1}{\pi_1(\alpha_1, \gamma)} - 1\right\}\cdot V_1(X)\right], \\ 0 &= \hat{E}\left[\left\{\frac{R_2 - R_1}{\pi_2(\alpha_2, \gamma)} - (1 - R_1)\right\}\cdot V_2(X)\right], \\ 0 &= \hat{E}\left[\left\{\frac{R_2 - R_1}{\pi_2(\alpha_2, \gamma)} - \frac{1 - \pi_1(\alpha_1, \gamma)}{\pi_1(\alpha_1, \gamma)} R_1\right\}\cdot \{U(X, Y) - E(U(X, Y) \mid X, R_2 = \end{aligned}$$

Doubly robust estimation

Theorem 3. Under Assumption 1 and certain regularity conditions, $(\hat{\alpha}_{1,\mathrm{dr}}, \hat{\gamma}_{\mathrm{dr}}, \hat{\mu}_{\mathrm{dr}})$ are consistent and asymptotically normal provided one of the following conditions holds:

- $A_1(X;\alpha_1), \Gamma(X,Y;\gamma)$ and $A_2(X;\alpha_2)$ are correctly specified; or
- $A_1(X;\alpha_1), \Gamma(X,Y;\gamma)$ and $f_2(Y \mid X;\beta)$ are correctly specified.

Furthermore, $\hat{\mu}_{\mathrm{dr}}$ attains the semiparametric efficiency bound for the nonparametric model $\mathcal{M}_{\mathrm{npr}}$ when all models $\{A_1(X;\alpha_1), A_2(X;\alpha_2), \Gamma(X,Y;\gamma), f_2(Y \mid X;\beta)\}$ are correct.

$(\hat{\alpha}_{1,\mathrm{dr}}, \hat{\gamma}_{\mathrm{dr}}, \hat{\mu}_{\mathrm{dr}})$ are doubly robust against misspecification of $A_2(X;\alpha_2)$ and $f_2(Y \mid X;\beta)$, provided that the first-call propensity score $\pi_1(X,Y;\alpha_1,\gamma)$ (i.e., $A_1(X;\alpha_1), \Gamma(X,Y;\gamma)$) is correctly specified.

Takeaway points

- We propose the stableness of resistance assumption for identification, which is so far the most parsimonious condition characterizing the most flexible model for nonresponse adjustment with callbacks;
- we establish identification and develop IPW, outcome-regression based, and doubly robust estimation methods;
- we establish the semiparametric efficiency theory for using callbacks.

Outline

1. Paradata, callback data, and missing data

2. Real-world example

3. Identification and semiparametric inference

4. Extensions

5. Simulations and application

6. Discussion

Estimation of a general smooth functional

Consider estimation of θ defined by the solution to a given estimating equation $E\{m(X,Y;\theta)\}=0$. Assuming $\partial E\{m(\theta)\}/\partial\theta$ is non-singular, IPW, outcome regression-based, and doubly robust estimation of θ can be obtained simply by replacing $Y-\mu$ with m in the corresponding estimating equations of μ.

The efficient influence function for δ in the nonprametric model $\mathcal{M}$ is $\mathrm{IF}(O;\theta)=-[\partial E\{m(\theta)\}/\partial\theta]^{-1}\phi(O)$, where

$$\begin{aligned}\phi(O) \quad = \quad & \left\{\frac{R_1}{\pi_1}-\frac{R_1}{\pi_2}\frac{1-\pi_1}{\pi_1}+\frac{R_2-R_1}{\pi_2^2}-1\right\}\left\{\frac{E(m/\pi_2 \mid X, R_2=0)}{E(1/\pi_2 \mid X, R_2=0)}-m\right\} \\ & +m.\end{aligned}$$

Estimation under an alternative parameterization

An alternative parameterization is to model $\{\pi_1, \pi_2, f_1\}$.

There is a one-to-one mapping between these two parameterizations

$$\{\pi_1, \pi_2, f_1\} \leftrightarrow \{\pi_1, \pi_2, f_2\}.$$

An IPW, an outcome regression-based, and a doubly robust and locally efficient estimator can also be developed under this parameterization, which are omitted here.

If response rate is relatively high in the first call, modeling f_1 is recommended, but otherwise, modeling f_2 is recommended.

Estimation with multiple callbacks

Identification equally holds under the stableness of resistance (for the first two calls) assumption for multiple callbacks, with no restrictions on the third and later callbacks;

IPW, outcome regression-based, and doubly robust estimators developed with two callbacks still work, but no longer efficient.

For multiple callbacks, derivation of the tangent space and the efficient influence function is in general complicated which involves projection.

Estimation of the causal effect on the untreated

Suppose we are interested in the effect of a binary treatment R (1 for treatment and 0 for control) on an outcome Y. Following from the potential outcomes framework for causal inference, we let Y_1, Y_0 denote the potential outcomes would be observed if the treatment were set to $R = 1, 0$, respectively. A causal effect is defined as a comparison of the potential outcomes–for instance, the average causal effect is $E\{Y_1 - Y_0\}$. The observed outcome is a realization of the potential outcome under the treatment a unit actually received, i.e., $Y = RY_1 + (1 - R)Y_0$. Therefore, causal inference is inherently a missing data problem because Y_1 is only observed for the treated units and Y_0 only for control units.

Estimation of the causal effect on the untreated

If the treatment assignment mechanism is ignorable, i.e., R is independent of $\{Y_1, Y_0\}$ conditional on fully observed covariates X, then the potential outcome distribution is identified and estimation of the potential outcome mean is analogous to the MAR setting in missing data analysis.

However, a key challenge for causal inference in observational studies is unmeasured confounding, i.e., some variables correlated with both the treatment and the outcome are not measured. In the presence of unmeasured confounding, ignorability does not hold and the missingness of potential outcomes is MNAR, and hence the potential outcome distributions are not identified without extra model assumptions or auxiliary data.

The instrumental variable is an influential tool for confounding bias adjustment (Wright, 1928; Hernán & Robins, 2006; Angrist et al., 1996), and recently, Miao et al. (2018), Shi et al. (2020), Tchetgen Tchetgen et al. (2020), Lipsitch et al. (2010), Kuroki & Pearl (2014), and Ogburn & VanderWeele (2012) establish a proximal inference framework for confounding bias adjustment. The instrumental variable approach entails an instrumental variable that is independent of the unmeasured confounders, correlated with the treatment, and does not directly affect the outcome; and the proximal inference rests on two proxies of the unmeasured confounder; besides, both approaches require additional conditions to achieve identified.

Estimation of the causal effect on the untreated

Here we show how to use callbacks for treatment to identify causal effects in the presence of unmeasured confounding. Suppose we make two calls to promote the treatment—for instance, the injection of Covid-19 vaccine; and units received the treatment (e.g., two shots of Covid-19 vaccine) in the first call will not be called again.

Let R_1, R_2 denote treatment states in these two calls, and Y the level of antibody six months after taking the shot. We have that $R_2 \geq R_1$. Assume that $Y_{R_1=0,R_2=1} = Y_{R_1=1,R_2=1} \equiv Y_{R_2=1}$, i.e., the effect of vaccine is the same whenever it is taken in these two calls. Particularly, we are interested in the average treatment effect on the untreated $ATUT = E\{Y_1 - Y_0 \mid R_2 = 0\}$.

Estimation of the causal effect on the untreated

Not all people are willing to take the treatment due to concerns about safety, side effect, etc. We assume that a unit's resistance to taking the treatment remains the same in these two calls, i.e., the odds ratio functions of $f\{R_1 = 1 \mid X, Y_{R_2=1}\}$ and $f\{R_2 = 1 \mid X, Y_{R_2=1}, R_1 = 0\}$ are the same.

Applying the above identification results, we can identify the distribution of $Y_{R_2=1}$ and $E(Y_{R_2=1} \mid R_2 = 0)$ and thus $ATUT = E(Y_{R_2=1} - Y_{R_2=0} \mid R_2 = 0)$, the effect of vaccine on those who ultimately do not take the vaccine in the two calls. In contrast to the instrumental variable approach and the proximal inference, this strategy does not involve instrumental or proxy variables, but makes use of the history or callbacks for treatment and invokes the stableness of resistance assumption to achieve identification.

Outline

1. Paradata, callback data, and missing data
2. Real-world example
3. Identification and semiparametric inference
4. Extensions
5. Simulations and application
6. Discussion

Data generating mechanism

Data generating model

$$\pi_1 = \text{expit }(\alpha_1^{\mathrm{T}} X + \gamma Y), \pi_2 = \text{expit }(\alpha_2^{\mathrm{T}} W_1 + \gamma Y), f_2(Y \mid X) \sim N(\beta^{\mathrm{T}} W_2, \sigma^2)$$

Four scenarios with different choices of (W_1, W_2)

	TT	FT	TF	FF
	$W_1 = X, W_2 = X$	$W_1 = \widetilde{X}, W_2 = X$	$W_1 = X, W_2 = \widetilde{X}$	$W_1 = \widetilde{X}, W_2 = \widetilde{X}$
α_1^{T}	(-1, 0.5, 0.2)	(-0.3, -0.7, 0.7)	(-1, 1, -0.1)	(-0.3, 0.5, 1)
α_2^{T}	(1, 0.5, 0.2)	(-0.3, 1.9, 0.9)	(0.5, 1, -0.1)	(-0.4, 0.8, 0)
β^{T}	(2.5, 2.3, 1.6)	(-1, 5.4, 4)	(-0.5, 5, -1)	(-1.5, 4, 3)
γ	0.16	0.1	0.5	0.25
σ	1.2	2	0.4	0.25

Working model for estimation

$$\pi_1 = \text{expit }(\alpha_1^{\mathrm{T}} X + \gamma Y), \pi_2 = \text{expit }(\alpha_2^{\mathrm{T}} X + \gamma Y), f_2(Y \mid X) \sim N(\beta^{\mathrm{T}} X, \sigma^2)$$

Bias plots

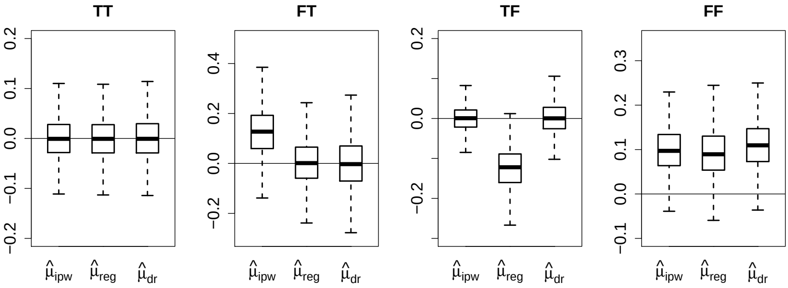


Figure 1: Bias for estimators of μ.

Bias plots

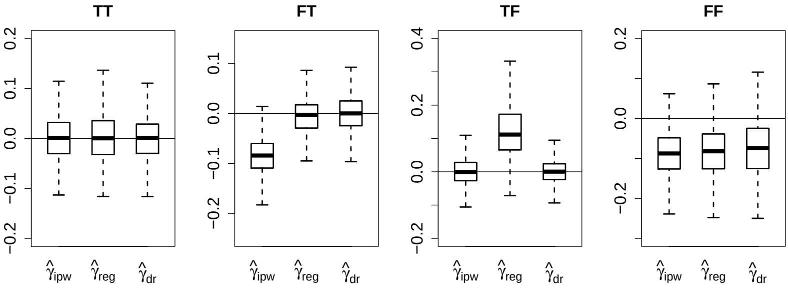


Figure 2: Bias for estimators of γ.

Application to the ANES NRFU Survey

Table 3: Estimates of voter turnout

		turnout rate	**odds ratio parameter γ**	
method	estimate	95% confidence interval	estimate	p-value
CC	0.870	(0.853,0.886)	—	—
MAR	0.809	(0.783, 0.835)	—	—
COR	0.857	(0.837,0.876)	—	—
IPW	0.662	(0.544,0.780)	1.561	0.0086
REG	0.623	(0.428,0.817)	1.866	0.0497
DR	0.666	(0.523,0.808)	1.486	0.0124

Application to the ANES NRFU Survey

Table 4: The coefficients of covariates in propensity score models

	The first contact stage		**The second contact stage**	
variable	estimate	p-value	estimate	p-value
intercept	-2.356	$<$0.0001**	-2.563	$<$0.0001**
m1sent: advance postcard	-0.200	0.1771	-0.281	0.3137
version: on page 1	0.197	0.1779	0.227	0.3434
title: long	-0.130	0.3817	-0.392	0.1709
incvis: visible	0.105	0.4700	0.178	0.4788
race: black	-0.296	0.1734	-0.130	0.6891
gender: male	-0.398	0.0055**	-0.606	0.0306**
age2: 30-59	0.671	0. 0012**	0.574	0.0521*
age3: 60$+$	1.487	$<$0.0001**	1.797	$<$0.0001**
education: some college	0.318	0.2850	0.387	0.3654

Application to the Consumer Expenditure Surveys

The Consumer Expenditure Surveys (CE) program provides data on the buying habits of U.S. consumers, collected by U.S. Bureau of Labor Statistics. The survey releases the callback data.

We use the public-use microdata collected in the fourth quarter of 2018 for illustration.

Outcomes: Y_1 and Y_2 are the log of last quarter's expenditures on housing and on utilities, fuels, and public services, respectively.

9709 households we analyze, 1992 responded in the first stage $(R_1 = 1, R_2 = 1)$, 3287 responded later $(R_1 = 0, R_2 = 1)$, and 4430 never responded $(R_1 = 0, R_2 = 0)$.

We use logistic propensity score models and a bivariate normal outcome model. The point estimates of outcome means obtained by different methods are close, however, the doubly robust estimator has a smaller variance.

Application to the Consumer Expenditure Surveys

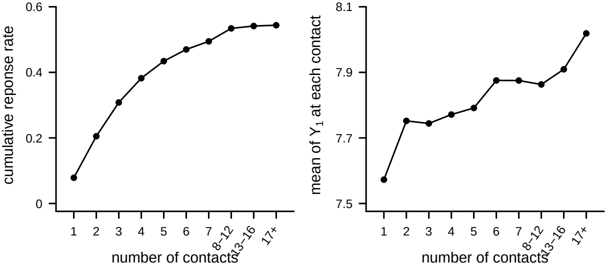


Figure 3: Response rates and outcomes mean for respondents for the CES application.

Application to the Consumer Expenditure Surveys

The estimates of the odds ratio are negative, suggesting that high-consuming people are less likely to respond. The expenditures on housing has a larger impact on the response propensity; this may related to the survey process, which takes home visits as one of the main ways of contact.

Table 5: Point estimate, 95% confidence interval, and p-values

	μ_1				γ_1		
	IPW	REG	DR	CC	IPW	REG	DR
Estimate	7.850	7.859	7.842	7.756	-0.265	-0.252	-0.238
CI or p.v.	(7.769, 7.931)	(7.808, 7.910)	(7.800, 7.884)	(7.734, 7.778)	0.008	0.002	0
	μ_2				γ_2		
	IPW	REG	DR	CC	IPW	REG	DR
Estimate	6.339	6.352	6.345	6.285	-0.030	-0.048	-0.056
CI or p.v.	(6.257, 6.422)	(6.299, 6.405)	(6.296, 6.393)	(6.264, 6.306)	0.806	0.615	0.521

Outline

1. Paradata, callback data, and missing data

2. Real-world example

3. Identification and semiparametric inference

4. Extensions

5. Simulations and application

6. Discussion

Some other extensions

With the assist of callback data one can test MAR because it is a special case of our stableness of resistance assumption.

Identification remains if this assumption holds for any two adjacent calls by applying our identifying strategy to the subsurvey starting with these two calls.

We assumed a single action response process—a call attempt either succeeds or fails; however, samples can be classified into one of several dispositions, e.g., interview, refusal, other non-response or final non-contact (Biemer et al., 2013).

Our approach admits a vector of variables missing or observed simultaneously, which is often the case when the missingness is due to failure of contact, but in practice it is possible that different frame variables are observed in different call attempts and one needs to account for complex patterns of missingness.

Explore the idea of the stableness resistance in causal inference, case-control studies, and longitudinal studies.

Thanks!

References

Alho, J. M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika* **77**, 617–624.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91**, 444–455.

Bates, N. (2003). Contact histories in personal visit surveys: The survey of income and program participation (sipp) methods panel. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Biemer, P. P., Chen, P. & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **176**, 147–168.

Chen, B., Li, P. & Qin, J. (2018). Generalization of heckman selection model to nonignorable nonresponse using call-back information. *Statistica Sinica* **28**, 1761–1785.

Couper, M. (1998). Measuring survey quality in a casic environment. In *Proceedings of the Survey Research Methods Section of the ASA at JSM1998*. pp. 41–49.

References

Daniels, M. J., Jackson, D., Feng, W. & White, I. R. (2015). Pattern mixture models for the analysis of repeated attempt designs. *Biometrics* **71**, 1160–1167.

Das, M., Newey, W. K. & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies* **70**, 33–58.

D'Haultfœuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics* **154**, 1–15.

Drew, J. & Fuller, W. A. (1980). Modeling nonresponse in surveys with callbacks. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. John Wiley & Sons.

Guan, Z., Leung, D. H. & Qin, J. (2018). Semiparametric maximum likelihood inference for nonignorable nonresponse with callbacks. *Scandinavian Journal of Statistics* **45**, 962–984.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

References

Hernán, M. A. & Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology* **17**, 360–372.

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

Kott, P. S. (2014). Calibration weighting when model and calibration variables can differ. In *Contributions to Sampling Statistics*, F. Mecatti, L. P. Conti & G. M. Ranalli, eds. Cham: Springer, pp. 1–18.

Kreuter, F. (2013). Improving surveys with paradata: Introduction. In *Kreuter, F. Improving Surveys with Paradata: Analytic Uses of Process Information. Wiley*. Wiley Online Library, pp. 1–9.

Kreuter, F., Müller, G. & Trappmann, M. (2010). Nonresponse and measurement error in employment research: making use of administrative data. *Public Opinion Quarterly* **74**, 880–906.

Kuroki, M. & Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101**, 423–437.

Lin, I.-F. & Schaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly* **59**, 236–258.

References

Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388.

Little, R. J. (1982). Models for nonresponse in sample surveys. *Journal of the American statistical Association* **77**, 237–250.

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley: New York.

Liu, L., Miao, W., Sun, B., Robins, J. & Tchetgen, E. T. (2020). Identification and inference for marginal average treatment effect on the treated with an instrumental variable. *Statistica Sinica* **30**, 1517–1541.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**, 313–333.

Miao, W., Ding, P. & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Miao, W., Geng, Z. & Tchetgen Tchetgen, E. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105**, 987–993.

References

Miao, W. & Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **103**, 475–482.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal* **12**, S217–S229.

Ogburn, E. L. & VanderWeele, T. J. (2012). On the nondifferential misclassification of a binary confounder. *Epidemiology* **23**, 433–439.

Olson, K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science* **645**, 142–170.

Politz, A. & Simmons, W. (1949). An attempt to get the "not at homes"into the sample without callbacks. *Journal of the American Statistical Association* **44**, 9–16.

Qin, J. & Follmann, D. A. (2014). Semiparametric maximum likelihood inference by using failed contact attempts to adjust for nonignorable nonresponse. *Biometrika* **101**, 985–991.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

References

Shi, X., Miao, W., Nelson, J. C. & Tchetgen Tchetgen, E. (2020). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B* **82**, 521–540.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. & Tchetgen, E. J. T. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica* **28**, 1965.

Tchetgen Tchetgen, E. J. & Wirth, K. E. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* , in press.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X. & Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982* .

Vansteelandt, S., Rotnitzky, A. & Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94**, 841–860.

Wang, S., Shao, J. & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.

References

WRIGHT, P. G. (1928). *Tariff on Animal and Vegetable Oils*. New York: Macmillan.

ZHANG, Y., CHEN, H. & ZHANG, N. (2018). Bayesian inference for nonresponse two-phase sampling. *Statistica Sinica* **28**, 2167–2187.

ZHAO, J. & SHAO, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* **110**, 1577–1590.