Semiparametric models with data missing by design and inverse probability weighted empirical processes: partial results and open problems

Jon A. Wellner

Joint work with Norman Breslow

University of Washington

- joint work with Norman E. Breslow,
  University of Washington
- Talk at Joint Statistical Meetings, August 9, 2006
- *Email: jaw@stat.washington.edu*
  http: //www.stat.washington.edu/jaw/jaw.research.html

Outline

1. Semiparametric models with missing data by design
2. Horovitz - Thompson estimators
3. Finite sampling empirical processes for stratified sampling
4. Applying the Praestgaard - Wellner theorem
5. Summary; problems and open questions

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space
- Assume:

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space
- Assume:
    - There exist $\sqrt{n}-$consistent, asymptotically Gaussian ML estimators $(\widehat{\theta}_n, \widehat{\eta}_n)$ of θ and η under i.i.d. random sampling (i.e. complete data).

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space
- Assume:
    - There exist $\sqrt{n}-$consistent, asymptotically Gaussian ML estimators $(\widehat{\theta}_n, \widehat{\eta}_n)$ of θ and η under i.i.d. random sampling (i.e. complete data).
    - Scores $\dot{l}_\theta$ and $\dot{l}_\eta n = B_{\theta,\eta} h$, $h \in \mathcal{H} \subset \mathcal{B}$ in a Donsker class $\mathcal{F}$.

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space
- Assume:
  - There exist $\sqrt{n}-$consistent, asymptotically Gaussian ML estimators $(\widehat{\theta}_n, \widehat{\eta}_n)$ of θ and η under i.i.d. random sampling (i.e. complete data).
  - Scores $\dot{l}_\theta$ and $\dot{l}_\eta n = B_{\theta,\eta}h$, $h \in \mathcal{H} \subset \mathcal{B}$ in a Donsker class $\mathcal{F}$.
  - Information operator $\dot{l}_\eta^T \dot{l}_\eta = B_0^* B_0$ continuously invertible on its range

1. Semiparametric models with missing data by design

- **Setting:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
- parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space
- Assume:
  - There exist $\sqrt{n}-$consistent, asymptotically Gaussian ML estimators $(\widehat{\theta}_n, \widehat{\eta}_n)$ of θ and η under i.i.d. random sampling (i.e. complete data).
  - Scores $\dot{l}_\theta$ and $\dot{l}_\eta n = B_{\theta,\eta} h$, $h \in \mathcal{H} \subset \mathcal{B}$ in a Donsker class $\mathcal{F}$.
  - Information operator $\dot{l}_\eta^T \dot{l}_\eta = B_0^* B_0$ continuously invertible on its range
  - $(\widehat{\theta}_n, \widehat{\eta}_n)$ are consistent for (θ_0, η_0).

- Missing data – by design! X not observed for all items / individuals

- **Missing data** – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
    - $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
  - $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample
  - $V = (\widetilde{X}, U) \in \mathcal{V}$ observable in phase 1 (for all)

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
  - $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample
  - $V = (\widetilde{X}, U) \in \mathcal{V}$ observable in phase 1 (for all)
- Phase 1: $\{W_1, \ldots, W_N\}$ i.i.d. $P = P_W$

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
  - $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample
  - $V = (\widetilde{X}, U) \in \mathcal{V}$ observable in phase 1 (for all)
- Phase 1: $\{W_1, \ldots, W_N\}$ i.i.d. $P = P_W$
  - but observe only $\{V_1, \ldots, V_N\}$

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
  - $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample
  - $V = (\widetilde{X}, U) \in \mathcal{V}$ observable in phase 1 (for all)
- Phase 1: $\{W_1, \ldots, W_N\}$ i.i.d. $P = P_W$
  - but observe only $\{V_1, \ldots, V_N\}$
- Phase 2: Sampling indicators $\{\xi_1, \ldots, \xi_N\}$

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
  - $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample
  - $V = (\widetilde{X}, U) \in \mathcal{V}$ observable in phase 1 (for all)
- Phase 1: $\{W_1, \ldots, W_N\}$ i.i.d. $P = P_W$
  - but observe only $\{V_1, \ldots, V_N\}$
- Phase 2: Sampling indicators $\{\xi_1, \ldots, \xi_N\}$
  - observe W_i (all of X_i) if $\xi_i = 1$

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

independent

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - Phase 1: Observe $N_j = \sum_{j=1}^{N} 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - Phase 1: Observe $N_j = \sum_{j=1}^N 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j
  - Phase 2: Sample n_j of N_j without replacement:

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - Phase 1: Observe $N_j = \sum_{j=1}^{N} 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j
  - Phase 2: Sample n_j of N_j without replacement:
  - Result: sampling indicators $\xi_{j,i}$ for subject i in stratum j

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - Phase 1: Observe $N_j = \sum_{j=1}^{N} 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j
  - Phase 2: Sample n_j of N_j without replacement:
  - Result: sampling indicators $\xi_{j,i}$ for subject i in stratum j
  - $(\xi_{j,1}, \ldots, \xi_{j,N_j})$ exchangeable with

$$Pr(\xi_{ji} = 1|V_1, \ldots, V_N) = n_j/N_j.$$

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

$$Pr(\xi_i = 1|W_i) = Pr(\xi_i = 1|V_i) = \pi_0(V_i)$$

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - Phase 1: Observe $N_j = \sum_{j=1}^N 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j
  - Phase 2: Sample n_j of N_j without replacement:
  - Result: sampling indicators $\xi_{j,i}$ for subject i in stratum j
  - $(\xi_{j,1}, \ldots, \xi_{j,N_j})$ exchangeable with

$$Pr(\xi_{ji} = 1|V_1, \ldots, V_N) = n_j/N_j.$$

  - The vectors $(\xi_{j,1}, \ldots, \xi_{j,N_j})$, $j = 1, \ldots, J$ are independent

2. Horovitz-Thompson (or IPW Likelihood) Estimators

- Define inverse probability weighted (IPW) empirical measure:

$$\mathbb{P}_N^{\pi} = \frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i}{\pi_i}\delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

2. Horovitz-Thompson (or IPW Likelihood) Estimators

- Define inverse probability weighted (IPW) empirical measure:

$$\mathbb{P}_N^{\pi} = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

- Jointly solve the finite - (for θ) and infinite (for η) dimensional equations

$$\begin{aligned} \mathbb{P}_N^{\pi} \dot{l}_{\theta} &= 0 \qquad \text{in } \mathbb{R}^d \\ \mathbb{P}_N^{\pi} \dot{l}_{\eta} &= 0 \qquad \forall h \in \mathcal{H} \end{aligned}$$

2. Horovitz-Thompson (or IPW Likelihood) Estimators

- Define inverse probability weighted (IPW) empirical measure:

$$\mathbb{P}_N^\pi = \frac{1}{N}\sum_{i=1}^N \frac{\xi_i}{\pi_i}\delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

- Jointly solve the finite - (for θ) and infinite (for η) dimensional equations

$$\begin{aligned} \mathbb{P}_N^\pi \dot{l}_\theta &= 0 \qquad \text{in } \mathbb{R}^d \\ \mathbb{P}_N^\pi \dot{l}_\eta &= 0 \qquad \forall h \in \mathcal{H} \end{aligned}$$

- MLE for complete data solves same equations with $\mathbb{P}_N$ instead of $\mathbb{P}_N^\pi$.

Our Main Result:

- $\widehat{\theta}_N$ solving the IPW estimating equations is asymptotically linear

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \widetilde{l}_{\theta_0}(X_i) + o_p(1)$$

$$= \mathbb{G}_N^{\pi}(\widetilde{l}_{\theta_0}) + o_p(1)$$

where $\widetilde{l}_{\theta}(x)$ is the semparametric efficient influence function for θ (complete data)

$$\mathbb{G}_N^{\pi} = \sqrt{N}(\mathbb{P}_N^{\pi} - P).$$

3. Finite sampling empirical processes, stratified sampling

- Finite sampling empirical measure for stratum $j \in \{1, \ldots, J\}$:

$$\mathbb{P}^{\xi}_{j,N_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} \xi_{ji} \delta_{X_{ji}}$$

3. Finite sampling empirical processes, stratified sampling

- Finite sampling empirical measure for stratum $j \in \{1, \ldots, J\}$:

$$\mathbb{P}^{\xi}_{j,N_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} \xi_{ji} \delta_{X_{ji}}$$

- Closely related to *exchangeably weighted bootstrap empirical measure*
of Praestgaard and Wellner (1993),
van der Vaart and Wellner (1996), section 3.6

- Finite sampling empirical process

$$\mathbb{G}^{\xi}_{j,N_j} = \sqrt{N_j}\left(\mathbb{P}^{\xi}_{j,N_j} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right),$$

where

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\delta_{X_{ji}}.$$

- Finite sampling empirical process

$$\mathbb{G}^{\xi}_{j,N_j} = \sqrt{N_j}\left(\mathbb{P}^{\xi}_{j,N_j} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right),$$

where

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\delta_{X_{ji}}.$$

- Suppose that $\mathcal{F}$ is a P_0-Donsker class of functions containing all the scores $\dot{l}_\theta$ and $\dot{l}_\eta$ corresponding to parameter values in a neighborhood of the (θ_0, η_0) (guaranteed by Assumption 1).

- Finite sampling empirical process

$$\mathbb{G}^{\xi}_{j,N_j} = \sqrt{N_j}\left(\mathbb{P}^{\xi}_{j,N_j} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right),$$

where

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\delta_{X_{ji}}.$$

- Suppose that $\mathcal{F}$ is a P_0-Donsker class of functions containing all the scores $\dot{l}_\theta$ and $\dot{l}_\eta$ corresponding to parameter values in a neighborhood of the (θ_0, η_0) (guaranteed by Assumption 1).
- $\nu_j \equiv P_0(\mathcal{V}_j)$

- Finite sampling empirical process

$$\mathbb{G}^{\xi}_{j,N_j} = \sqrt{N_j}\left(\mathbb{P}^{\xi}_{j,N_j} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right),$$

where

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\delta_{X_{ji}}.$$

- Suppose that $\mathcal{F}$ is a P_0-Donsker class of functions containing all the scores $\dot{l}_\theta$ and $\dot{l}_\eta$ corresponding to parameter values in a neighborhood of the (θ_0, η_0) (guaranteed by Assumption 1).
- $\nu_j \equiv P_0(\mathcal{V}_j)$
- $n_j/N_j \to_p p_j$

Step one: Weak convergence of Finite Sampling Empirical Process

- Define:

$$
\begin{aligned}
\mathbb{G} &= \text{a } P_0\text{–Brownian bridge process indexed by } \mathcal{F} \\
\mathbb{G}_j &= \text{a } P_{0|j}\text{-Brownian bridge process indexed by } \mathcal{F} \\
\mathbb{G}_j(f) &= \frac{1}{\sqrt{\nu_j}} \mathbb{G}\{(f - P_{0|j}(f)|1_{\mathcal{V}_j}\}, \quad f \in \mathcal{F} \\
P_{0|j}(f) &= E\left(f(X)|V \in \mathcal{V}_j\right).
\end{aligned}
$$

Step one: Weak convergence of Finite Sampling Empirical Process

- Define:

$$
\begin{aligned}
\mathbb{G} &= \text{a } P_0-\text{Brownian bridge process indexed by } \mathcal{F} \\
\mathbb{G}_j &= \text{a } P_{0|j}\text{-Brownian bridge process indexed by } \mathcal{F} \\
\mathbb{G}_j(f) &= \frac{1}{\sqrt{\nu_j}}\mathbb{G}\{(f - P_{0|j}(f)|1_{\mathcal{V}_j}\}, \quad f \in \mathcal{F} \\
P_{0|j}(f) &= E\left(f(X)|V \in \mathcal{V}_j\right).
\end{aligned}
$$

- By the exchangeably weighted bootstrap limit theorem (Praestgaard & Wellner, 1993), with $\{\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_J\} \in UC(\mathcal{F})$ independent

$$(\mathbb{G}_N, \mathbb{G}^{\xi}_{1,N_1}, \ldots, \mathbb{G}^{\xi}_{J,N_J}) \rightsquigarrow (\mathbb{G}, \sqrt{p_1(1-p_1)}\mathbb{G}_1, \ldots, \sqrt{p_J(1-p_J)}\mathbb{G}_J),$$

$$\mathbb{G}^{\pi}_N = \mathbb{G}_N + \sum_{j=1}^{J} \frac{N_j}{N}\left(\frac{N_j}{n_j}\right)\mathbb{G}^{\xi}_{j,N_j} \rightsquigarrow \mathbb{G} + \sum_{j=1}^{J}\sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j$$

Step two: Apply van der Vaart's general Z-theorem (vdvW, 3.3.1):

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightsquigarrow N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

Step two: Apply van der Vaart's general Z-theorem (vdvW, 3.3.1):

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightsquigarrow N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling is centering of efficient scores

Step two: Apply van der Vaart's general Z-theorem (vdvW, 3.3.1):

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightsquigarrow N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.

Step two: Apply van der Vaart's general Z-theorem (vdvW, 3.3.1):

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightsquigarrow N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.
  - Select strata via covariates so that $\tilde{\ell}$ has small conditional variances on the strata

Step two: Apply van der Vaart's general Z-theorem (vdvW, 3.3.1):

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightsquigarrow N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.
  - Select strata via covariates so that $\tilde{\ell}$ has small conditional variances on the strata
  - Alternatively: Bernoulli sampling, but model the selection probabilities $\pi_\alpha(V)$ and estimate the α's (Norm's talk on Monday)

Norm's talk on Monday:

- Application to Cox regression

Different finish here: back to the math!

Norm’s talk on Monday:

- Application to Cox regression
- Bernoulli sampling with estimated weights

Different finish here: back to the math!

Norm’s talk on Monday:

- Application to Cox regression
- Bernoulli sampling with estimated weights
  - Correspondence with finite population sampling

Different finish here: back to the math!

Norm’s talk on Monday:

- Application to Cox regression
- Bernoulli sampling with estimated weights
  - Correspondence with finite population sampling
  - Further possible efficiency gains

Different finish here: back to the math!

Norm's talk on Monday:

- Application to Cox regression
- Bernoulli sampling with estimated weights
  - Correspondence with finite population sampling
  - Further possible efficiency gains
- Algorithm for computing variances (T. Lumley, R Survey package)

Different finish here: back to the math!

Norm's talk on Monday:

- Application to Cox regression
- Bernoulli sampling with estimated weights
  - Correspondence with finite population sampling
  - Further possible efficiency gains
- Algorithm for computing variances (T. Lumley, R Survey package)
- Extensions (unbiased estimating equations; complex probability sampling)

Different finish here: back to the math!

4. Applying the Praestgaard-Wellner theorem

- Recall the stratum specific empirical measure

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N} \delta_{X_{j,i}} = \frac{1}{N_j}\sum_{i=1}^{N} \delta_{X_i} 1_{\mathcal{V}_j}(V_i)$$

Note “double indexing” versus “single indexing”.

4. Applying the Praestgaard-Wellner theorem

- Recall the stratum specific empirical measure

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N}\delta_{X_{j,i}} = \frac{1}{N_j}\sum_{i=1}^{N}\delta_{X_i}1_{\mathcal{V}_j}(V_i)$$

Note “double indexing” versus “single indexing”.

4. Applying the Praestgaard-Wellner theorem

- Recall the stratum specific empirical measure

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N} \delta_{X_{j,i}} = \frac{1}{N_j}\sum_{i=1}^{N} \delta_{X_i} 1_{\mathcal{V}_j}(V_i)$$

Note “double indexing” versus “single indexing”.

- Need to show: if $\mathcal{F}$ is P_0-Donsker and $\nu_j > 0$, then $\mathcal{F}$ is $P_{0|j}$-Donsker on stratum $\mathcal{V}_j$ in the sense that

$$\mathbb{G}_{j,N_j} \equiv \sqrt{N_j}(\mathbb{P}_{j,N_j} - P_{0|j}) \rightsquigarrow \mathbb{G}_j \quad \text{in} \quad \ell^\infty(\mathcal{F})$$

- where $\mathbb{G}_j$ is a $P_{0|j}$-Brownian bridge process:

$$\{\mathbb{G}_j(f) \overset{d}{=} \nu_j^{-1/2}\mathbb{G}((f - P_{0|j}(f))1_{\mathcal{V}_j}), \qquad f \in \ell^\infty(\mathcal{F})\}.$$

- Exchangeable weights: $W_n = (W_{n1}, \ldots, W_{nn})$ satisfying

- Exchangeable weights: $W_n = (W_{n1}, \ldots, W_{nn})$ satisfying
    - W1 $\sup_n \|W_{n1} - \overline{W}_n\|_{2,1} < \infty$

- Exchangeable weights: $W_n = (W_{n1}, \ldots, W_{nn})$ satisfying
  - W1 $\sup_n \|W_{n1} - \overline{W}_n\|_{2,1} < \infty$
  - W2 $n^{-1/2} E \max_{1 \le i \le n} |W_{ni} - \overline{W}_n| \to 0$

- Exchangeable weights: $W_n = (W_{n1}, \ldots, W_{nn})$ satisfying
  - W1 $\sup_n \|W_{n1} - \overline{W}_n\|_{2,1} < \infty$
  - W2 $n^{-1/2} E \max_{1 \le i \le n} |W_{ni} - \overline{W}_n| \to 0$
  - W3 $n^{-1} \sum_{i=1}^{n} (W_{ni} - \overline{W}_n)^2 \to_p c^2$

- Exchangeable weights: $W_n = (W_{n1}, \ldots, W_{nn})$ satisfying
  - W1 $\sup_n \|W_{n1} - \overline{W}_n\|_{2,1} < \infty$
  - W2 $n^{-1/2} E \max_{1 \le i \le n} |W_{ni} - \overline{W}_n| \to 0$
  - W3 $n^{-1} \sum_{i=1}^{n} (W_{ni} - \overline{W}_n)^2 \to_p c^2$
- Exchangeably weighted bootstrap empirical measure

$$\widehat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^{n} W_{ni} \delta_{X_i}$$

- Exchangeable weights: $W_n = (W_{n1}, \ldots, W_{nn})$ satisfying
  - W1 $\sup_n \|W_{n1} - \overline{W}_n\|_{2,1} < \infty$
  - W2 $n^{-1/2} E \max_{1 \le i \le n} |W_{ni} - \overline{W}_n| \to 0$
  - W3 $n^{-1} \sum_{i=1}^n (W_{ni} - \overline{W}_n)^2 \to_p c^2$
- Exchangeably weighted bootstrap empirical measure

$$\widehat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^n W_{ni} \delta_{X_i}$$

- Exchangeably weighted bootstrap empirical measure

$$\begin{aligned} \widehat{\mathbb{G}}_n &= \sqrt{n}(\widehat{\mathbb{P}}_n - \overline{W}_n \mathbb{P}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - \overline{W}_n) \delta_{X_i} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{ni} (\delta_{X_i} - \mathbb{P}_n). \end{aligned}$$

Theorem (Praestgaard & Wellner, 1993) Suppose that $\mathcal{F}-$ is $P-$Donsker and W1-W3 hold. Then

$$\sup_{h \in BL_1} |E_W h(\widehat{\mathbb{G}}_n) - Eh(c\mathbb{G})| \to_P 0.$$

If $\mathcal{F}$ has a square integrable envelope F, $PF^2 < \infty$, then the convergence is also (outer) almost sure.

- Apply this for each fixed strata $j \in \{1, \dots, J\}$ with:

Theorem (Praestgaard & Wellner, 1993) Suppose that $\mathcal{F}-$ is $P-$Donsker and W1-W3 hold. Then

$$\sup_{h \in BL_1} |E_W h(\widehat{\mathbb{G}}_n) - Eh(c\mathbb{G})| \to_P 0.$$

If $\mathcal{F}$ has a square integrable envelope F, $PF^2 < \infty$, then the convergence is also (outer) almost sure.

- Apply this for each fixed strata $j \in \{1, \ldots, J\}$ with:
  - $n = N_j$ random

Theorem (Praestgaard & Wellner, 1993) Suppose that $\mathcal{F}-$ is $P-$Donsker and W1-W3 hold. Then

$$\sup_{h \in BL_1} |E_W h(\widehat{\mathbb{G}}_n) - Eh(c\mathbb{G})| \to_P 0.$$

If $\mathcal{F}$ has a square integrable envelope F, $PF^2 < \infty$, then the convergence is also (outer) almost sure.

- Apply this for each fixed strata $j \in \{1, \ldots, J\}$ with:
  - $n = N_j$ random
  - $P = P_{0|j} = P_0(\cdot 1_{\mathcal{V}_j})/P_0(1_{\mathcal{V}_j})$

Theorem (Praestgaard & Wellner, 1993) Suppose that $\mathcal{F}-$ is $P-$Donsker and W1-W3 hold. Then

$$\sup_{h \in BL_1} |E_W h(\widehat{\mathbb{G}}_n) - Eh(c\mathbb{G})| \to_P 0.$$

If $\mathcal{F}$ has a square integrable envelope F, $PF^2 < \infty$, then the convergence is also (outer) almost sure.

- Apply this for each fixed strata $j \in \{1, \ldots, J\}$ with:
  - $n = N_j$ random
  - $P = P_{0|j} = P_0(\cdot 1_{\mathcal{V}_j})/P_0(1_{\mathcal{V}_j})$
  - $W = (\xi_{j,i}, \ldots, \xi_{j,N_j})$

Theorem (Praestgaard & Wellner, 1993) Suppose that $\mathcal{F}-$ is $P-$Donsker and W1-W3 hold. Then

$$\sup_{h \in BL_1} |E_W h(\widehat{\mathbb{G}}_n) - Eh(c\mathbb{G})| \to_P 0.$$

If $\mathcal{F}$ has a square integrable envelope F, $PF^2 < \infty$, then the convergence is also (outer) almost sure.

- Apply this for each fixed strata $j \in \{1, \ldots, J\}$ with:
  - $n = N_j$ random
  - $P = P_{0|j} = P_0(\cdot 1_{\mathcal{V}_j})/P_0(1_{\mathcal{V}_j})$
  - $W = (\xi_{j,i}, \ldots, \xi_{j,N_j})$
  - If $n_j/N_j \to_p p_j$, then

$$\frac{1}{N_j}\sum_{i=1}^{N_j}(\xi_{j,i} - \overline{\xi}_{j,i})^2 \to_p p_j(1-p_j)$$

- $\mathcal{F}$ is $P_{0|j}$ - Donsker: if $X_{j,1}, \ldots, X_{j,n}$ are i.i.d. $P_{0|j}$ and $\mathbb{P}_{j,n} = n^{-1} \sum_{i=1}^{n} \delta_{X_{j,i}}$, then

$$\mathbb{G}_{j,n} \equiv \sqrt{n}(\mathbb{P}_{j,n} - P_{0|j}) \rightsquigarrow \mathbb{G}_j \qquad \text{in} \;\; \ell^{\infty}(\mathcal{F})$$

where $\mathbb{G}_j$ is a $P_{0|j}$ - Brownian bridge process.

- $\mathcal{F}$ is $P_{0|j}$ - Donsker: if $X_{j,1}, \ldots, X_{j,n}$ are i.i.d. $P_{0|j}$ and $\mathbb{P}_{j,n} = n^{-1} \sum_{i=1}^{n} \delta_{X_{j,i}}$, then

$$\mathbb{G}_{j,n} \equiv \sqrt{n}(\mathbb{P}_{j,n} - P_{0|j}) \rightsquigarrow \mathbb{G}_j \qquad \text{in } \ \ell^\infty(\mathcal{F})$$

  where $\mathbb{G}_j$ is a $P_{0|j}$ - Brownian bridge process.
- Starting hypothesis: $\mathcal{F}$ is P_0 - Donsker

- $\mathcal{F}$ is $P_{0|j}$ - Donsker: if $X_{j,1}, \ldots, X_{j,n}$ are i.i.d. $P_{0|j}$ and $\mathbb{P}_{j,n} = n^{-1} \sum_{i=1}^{n} \delta_{X_{j,i}}$, then

$$\mathbb{G}_{j,n} \equiv \sqrt{n}(\mathbb{P}_{j,n} - P_{0|j}) \rightsquigarrow \mathbb{G}_j \qquad \text{in } \ell^\infty(\mathcal{F})$$

  where $\mathbb{G}_j$ is a $P_{0|j}$ - Brownian bridge process.
- Starting hypothesis: $\mathcal{F}$ is P_0 - Donsker
- Connecting link: double indexing (or “conditional sampling”) representation lemma for sampling from P_0.

Lemma. (Conditional sampling representation)

- Suppose $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_J$, $\mathcal{X}_j$'s disjoint

Lemma. (Conditional sampling representation)

- Suppose $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_J$, $\mathcal{X}_j$'s disjoint
- Let $\Delta \sim \text{Multinomial}_J(1, (\nu_1, \ldots, \nu_J))$ $\nu_j = P_0(X \in \mathcal{X}_j)$

Lemma. (Conditional sampling representation)

- Suppose $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_J$,
  $\mathcal{X}_j$'s disjoint
- Let $\Delta \sim$ Multinomial$_J(1, (\nu_1, \ldots, \nu_J))$
  $\nu_j = P_0(X \in \mathcal{X}_j)$
- $X_j^\dagger \sim P_{0|j}$ for $j \in \{1, \ldots, J\}$

Lemma. (Conditional sampling representation)

- Suppose $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_J$,
  $\mathcal{X}_j$'s disjoint
- Let $\Delta \sim \text{Multinomial}_J(1, (\nu_1, \ldots, \nu_J))$
  $\nu_j = P_0(X \in \mathcal{X}_j)$
- $X_j^\dagger \sim P_{0|j}$ for $j \in \{1, \ldots, J\}$
- Δ, $X_1^\dagger, \ldots, X_J^\dagger$ all independent

Lemma. (Conditional sampling representation)

- Suppose $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_J$, $\mathcal{X}_j$'s disjoint
- Let $\Delta \sim$ Multinomial$_J(1, (\nu_1, \ldots, \nu_J))$ $\nu_j = P_0(X \in \mathcal{X}_j)$
- $X_j^\dagger \sim P_{0|j}$ for $j \in \{1, \ldots, J\}$
- Δ, $X_1^\dagger, \ldots, X_J^\dagger$ all independent
- Then $X \sim P_0$ satisfies $X \stackrel{d}{=} \sum_{j=1}^J \Delta_j X_j^\dagger$.

5. Summary; problems and open questions

- Basic tools

5. Summary; problems and open questions

- Basic tools
  - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)

5. Summary; problems and open questions

- Basic tools
    - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)
    - Z- estimator theorem, van der Vaart (1995), vdV-W (1996)

5. Summary; problems and open questions

- Basic tools
  - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)
  - Z- estimator theorem, van der Vaart (1995), vdV-W (1996)
- Basic idea: separate calculations for sampling design and for model

5. Summary; problems and open questions

- Basic tools
  - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)
  - Z- estimator theorem, van der Vaart (1995), vdV-W (1996)
- Basic idea: separate calculations for sampling design and for model
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^\pi$

5. Summary; problems and open questions

- Basic tools
  - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)
  - Z- estimator theorem, van der Vaart (1995), vdV-W (1996)
- Basic idea: separate calculations for sampling design and for model
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_{\theta}$ for θ

5. Summary; problems and open questions

- Basic tools
  - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)
  - Z- estimator theorem, van der Vaart (1995), vdV-W (1996)
- Basic idea: separate calculations for sampling design and for model
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_\theta$ for θ
- Finite population theory seems closer to practice.

5. Summary; problems and open questions

- Basic tools
  - Exchangeably weighted bootstrap empirical process, Praestgaard-Wellner (1993)
  - Z- estimator theorem, van der Vaart (1995), vdV-W (1996)
- Basic idea: separate calculations for sampling design and for model
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_\theta$ for θ
- Finite population theory seems closer to practice.
- Extensions possible ? for other designs, other estimating equations.

- Other, more complex sampling designs?

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...

- Other, more complex sampling designs?
    - Hájek (1964), Rosen (1972a,b), ...
    - Lin (2000)

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!
  - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!
  - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).
- Can we handle both estimating the π's and and finite popl'n sampling?

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!
  - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).
- Can we handle both estimating the π's and and finite popl'n sampling?
- Application to Cox model via $Z-$theorem

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!
  - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).
- Can we handle both estimating the π's and and finite popl'n sampling?
- Application to Cox model via $Z-$theorem
  - van der Vaart (1985), (1998) imposes extra hypotheses

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!
  - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).
- Can we handle both estimating the π's and and finite popl'n sampling?
- Application to Cox model via $Z-$theorem
  - van der Vaart (1985), (1998) imposes extra hypotheses
  - compare vdV with Andersen and Gill (1982)

- Other, more complex sampling designs?
  - Hájek (1964), Rosen (1972a,b), ...
  - Lin (2000)
- Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$? Not easily!
  - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).
- Can we handle both estimating the π's and and finite popl'n sampling?
- Application to Cox model via $Z-$theorem
  - van der Vaart (1985), (1998) imposes extra hypotheses
  - compare vdV with Andersen and Gill (1982)
  - need very sharp / good $Z-$ theorem result