Empirical Likelihood Methods for Two-sample Problems with Data Missing-by-Design

Changbao Wu
Department of Statistics and Actuarial Science
University of Waterloo

(Joint work with Min Chen and Mary Thompson)

November 13, 2015

1. Empirical Likelihood for Two-sample Problems
2. Survey Samples for Observational Studies
3. Concluding Remarks

(1) Two Independent Samples

- Two groups: treatment vs control
- Response Y: Y_1 for treatment and Y_0 for control
- Sample data on Y:

$$\{Y_{11}, \cdots, Y_{1n_1}\} \quad \text{and} \quad \{Y_{01}, \cdots, Y_{0n_0}\}$$

- Covariates might be involved

- $\mu_1 = E(Y_1)$ and $\mu_0 = E(Y_0)$
- $F_1(t) = P(Y_1 \leq t)$ and $F_0(t) = P(Y_0 \leq t)$
- $S_1(t) = P(Y_1 > t) = 1 - F_1(t)$
  $S_0(t) = P(Y_0 > t) = 1 - F_0(t)$

(1) Two Independent Samples

- Inference on treatment effect: $\theta = \mu_1 - \mu_0$
- Inference on survival functions (distribution functions): Y_1 is stochastically larger than Y_0 if

$$S_1(t) > S_0(t) \qquad \text{for} \qquad t > 0\,.$$

- Empirical likelihood methods for two-sample problems (Wu and Yan, 2012):
  - Two independent samples with no missing values
  - Two independent samples with responses missing at random
  - Two independent survey samples with no missing values

(2) Pretest-Posttest Studies

- A very popular approach in medical and social sciences
- Measure changes resulting from a treatment or an intervention
- A version of two-sample problems

- Design I: Paired comparison (less popular one)

  Select a random sample of n units from the target population; Measure the response Y on all units BEFORE and AFTER the treatment/intervention.

(2) Pretest-Posttest Studies

Design II: (commonly used method)

- A random sample of n units is taken from the target population
- Measures on certain baseline variables $\boldsymbol{Z}$ are obtained for ALL n individuals (pretest measures)
- Among the n units, n_1 are randomly selected and assigned to "treatment";
  Values of the response variable Y are obtained (posttest measures)
- The other $n_0 = n - n_1$ units are assigned to "control";
  Values of the response variable Y are also obtained

(2) Pretest-Posttest Studies

- Y_1: response under treatment
  Y_0: response under control
- The available data ($n = n_1 + n_0$)

i	1	2	$\cdots$	n_1	$n_1 + 1$	$n_1 + 2$	$\cdots$	n
$\boldsymbol{Z}$	$\boldsymbol{Z}_1$	$\boldsymbol{Z}_2$	$\cdots$	$\boldsymbol{Z}_{n_1}$	$\boldsymbol{Z}_{n_1+1}$	$\boldsymbol{Z}_{n_1+2}$	$\cdots$	$\boldsymbol{Z}_n$
Y_1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$*$	$*$	$\cdots$	$*$
Y_0	$*$	$*$	$\cdots$	$*$	$Y_{0(n_1+1)}$	$Y_{0(n_1+2)}$	$\cdots$	Y_{0n}

- Two distinct features of the design:
  - Response Missing-by-Design
  - Availability of baseline information for all n units
  - The $\boldsymbol{Z}$ variables follow the same distributions for both groups (due to randomization)

(2) Pretest-Posttest Studies

- A two-sample problem with unique features
- Parameter of primary interest: $\theta = \mu_1 - \mu_0$
- Test H_0: $\theta = 0$ vs H_1: $\theta \neq 0$ (or $\theta > 0$)
- Test H_0: $F_1 = F_0$ vs H_1: $F_1 < F_0$
  (OR H_0: $S_1 = S_0$ vs H_1: $S_1 > S_0$)
- Question: How to effectively use the baseline information and the feature of missing-by-design?

(3) Two-sample Problems for Observational Studies

- Baseline information (Z) collected for all n units
- Each unit is assigned to either treatment or control (Missing-by-Design)
- Assignments to treatment or control are **not randomized**:

Example 1. Patients self-selection of treatment among two alternative choices.

Example 2. Voluntary participation in a school smoking-intervention education program.

Example 3. Modes of survey data collection: Web versus telephone interview.

An EL Approach to Pretest-Posttest Studies (Huang , Qin and Follmann, JASA, 2008)

- Parameter of interest: $\theta = \mu_1 - \mu_0$
- Find the EL estimators of μ_1 and μ_0 separately
- Estimate θ by $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_0$
- Estimate the standard error of $\hat{\theta}$ through a bootstrap method
- Inference on $\theta = \mu_1 - \mu_0$ using $(\hat{\theta} - \theta)/SE(\hat{\theta})$

- Finding $\hat{\mu}_1$ (and $\hat{\mu}_0$) is the main focus of the HQF paper

An Imputation-based Two-Sample EL Approach

- Why imputation? More efficient use of baseline information!

i	1	2	$\cdots$	n_1	n_1+1	n_1+2	$\cdots$	n
$\boldsymbol{Z}$	$\boldsymbol{Z}_1$	$\boldsymbol{Z}_2$	$\cdots$	$\boldsymbol{Z}_{n_1}$	$\boldsymbol{Z}_{n_1+1}$	$\boldsymbol{Z}_{n_1+2}$	$\cdots$	$\boldsymbol{Z}_n$
Y_1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$*$	$*$	$\cdots$	$*$
Y_0	$*$	$*$	$\cdots$	$*$	$Y_{0(n_1+1)}$	$Y_{0(n_1+2)}$	$\cdots$	Y_{0n}

- Regression modelling:

$$Y_{1i} = \boldsymbol{Z}_i^T \boldsymbol{\beta}_1 + \epsilon_{1i}, \quad i = 1, \cdots, n, \tag{1}$$

$$Y_{0i} = \boldsymbol{Z}_i^T \boldsymbol{\beta}_0 + \epsilon_{0i}, \quad i = 1, \cdots, n, \tag{2}$$

An Imputation-based Two-Sample EL Approach

- Let $R_i = 1$ if i is under treatment, $R_i = 0$ otherwise,

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} R_i \mathbf{Z}_i \mathbf{Z}_i^T\right)^{-1} \sum_{i=1}^{n} R_i \mathbf{Z}_i Y_{1i},$$
$$\hat{\beta}_0 = \left(\sum_{i=1}^{n} (1-R_i) \mathbf{Z}_i \mathbf{Z}_i^T\right)^{-1} \sum_{i=1}^{n} (1-R_i) \mathbf{Z}_i Y_{0i}$$

- Regression imputation:

$$Y_{1i}^{*} = \mathbf{Z}_i^T \hat{\beta}_1\,, \quad i = n_1 + 1, \cdots, n$$
$$Y_{0i}^{*} = \mathbf{Z}_i^T \hat{\beta}_0\,, \quad i = 1, \cdots, n_1$$

Imputation-basede EL Inference on $\theta = \mu_1 - \mu_0$

- Two augmented samples after imputation:

$$\{\tilde{Y}_{1i} = R_i Y_{1i} + (1 - R_i) Y_{1i}^*, \quad i = 1, \cdots, n\}$$

$$\{\tilde{Y}_{0i} = (1 - R_i) Y_{0i} + R_i Y_{0i}^*, \quad i = 1, \cdots, n\}.$$

- Each sample has an enlarged sample size at $n = n_1 + n_0$
- The imputed samples are no longer independent
- Inferences are under the assumed regression models (the imputation model)

Two Independent Surveys

- Two independent survey samples S_1 and S_0 from the same population
- Two (possibly different) designs: $d_{1i}, i \in S_1$; $d_{0i}, i \in S_0$

$$\tilde{d}_{1i} = \frac{d_{1i}}{\sum_{k \in S_1} d_{1k}} \quad \text{and} \quad \tilde{d}_{0i} = \frac{d_{0i}}{\sum_{k \in S_0} d_{0k}}$$

- Response variables Y_1 and Y_0; Survey sample data:

$$\left\{ (y_{1i}, z_{1i}), i \in S_1 \right\} \quad \text{and} \quad \left\{ (y_{0i}, z_{0i}), i \in S_0 \right\}$$

- Parameter of interest: $\theta = \mu_{y1} - \mu_{y0}$
- Design-based estimator of θ:

$$\hat{\theta}_1 = \hat{\mu}_{y1} - \hat{\mu}_{y0} = \sum_{i \in S_1} \tilde{d}_{1i} y_{1i} - \sum_{i \in S_0} \tilde{d}_{0i} y_{0i}$$

Two-sample Pseudo Empirical Likelihood

- Two-sample pseudo EL function

$$\ell(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2} \sum_{i \in S_1} \tilde{d}_{1i} \log(p_{1i}) + \frac{1}{2} \sum_{i \in S_0} \tilde{d}_{0i} \log(q_{0i})$$

- Normalization constraints

$$\sum_{i \in S_1} p_{1i} = 1 \quad \text{and} \quad \sum_{i \in S_0} q_{0i} = 1$$

- Parameter constraint

$$\sum_{i \in S_1} p_{1i} y_{1i} - \sum_{i \in S_0} q_{0i} y_{0i} = \theta$$

- Additional constraints on $\boldsymbol{z}_1$ and $\boldsymbol{z}_0$ depending on what's available

The ITC Four Country Survey (ITC4)

- The International Tobacco Control Policy Evaluation Project (The ITC Project)
- ITC Four Country Survey: Waves 1 -7 by telephone interview
- ITC Four Country Survey: Wave 8, respondents chose to complete the survey either by telephone interview or self-administered web survey
  - Total number of respondents at wave 8: 4507
  - Number of respondents by telephone: 2709
  - Number of respondents by web: 1798
- The research question: Examine the effect of two different modes for data collection
- Y_1: response by web (treatment)
  Y_0: response by telephone (control)

Mode Effect in Survey Data Collection

- Survey sample S of size n; design weights d_i, $i = 1, \cdots, n$.
- Units $i = 1, \cdots, n_1$ chose "treatment"
  Units $i = n_1 + 1, \cdots, n$ chose "control"
- Let $R_i = 1$ if unit i chose "treatment", $R_i = 0$ if unit i chose "control, $i = 1, \cdots, n$
- Baseline information z_i observed for all $i = 1, \cdots, n$
- Response variable **missing-by-design**:
  Y_{1i} observed for $i = 1, \cdots, n_1$
  Y_{0i} observed for $i = n_1 + 1, \cdots, n$
- Point estimate and confidence interval for $\theta = \mu_{y1} - \mu_{y0}$

Mode Effect

- Under randomization to treatment and control:

$$E(Y_1 \mid R = 1) - E(Y_0 \mid R = 0) = E(Y_1) - E(Y_0) = \theta$$

- With self-selection of treatment and control:

$$E(Y_1 \mid R = 1) \neq E(Y_1)\,, \qquad E(Y_0 \mid R = 0) \neq E(Y_0)$$

- Ignorable treatment assignment (Rosenbaum and Rubin, 1983):

$$(Y_1, Y_0) \;\text{ and }\; R \;\text{ are independent given }\; \boldsymbol{Z}$$

- Test ignobility using a two-phase sampling technique? (Chen and Kim, 2014)

Mode Effect: Propensity Score Adjustment (PSA)

- Treatment assignment (self-selection) depends only on $\boldsymbol{Z}$

$$P(R=1 \mid Y_1, Y_0, \boldsymbol{Z}=z) = P(R=1 \mid \boldsymbol{Z}=z) = r(z)$$

- Fit a feasible model to obtain the propensity scores

$$\hat{r}_i = \hat{r}(z_i)\,, \quad i=1,\cdots,n$$

- Available data

$$\Big\{(y_{1i}, z_i), i=1,\cdots,n_1\Big\} \quad \text{and} \quad \Big\{(y_{0i}, z_i), i=n_1+1,\cdots,n\Big\}$$

- Point estimator for $\theta = \mu_{y1} - \mu_{y0}$ using PSA:

$$\hat{\theta}_2 = \sum_{i=1}^{n_1} \frac{\tilde{d}_i}{\hat{r}_i} y_{1i} - \sum_{i=n_1+1}^{n} \frac{\tilde{d}_i}{1-\hat{r}_i} y_{0i}$$

Pseudo EL Under Propensity Score Adjustment

- Design weights $d_i = P(i \in S)$, $i = 1, \cdots, n$; $\tilde{d}_i = d_i / \sum_{k \in S} d_k$
- Pseudo empirical likelihood function and constraints

$$\ell(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n_1} \frac{\tilde{d}_i}{\hat{r}_i} \log(p_i) + \sum_{j=n_1+1}^{n} \frac{\tilde{d}_j}{1 - \hat{r}_j} \log(q_j) \quad (3)$$

$$\sum_{i=1}^{n_1} p_i = 1, \quad \sum_{j=n_1+1}^{n} q_j = 1 \quad (4)$$

$$\sum_{i=1}^{n_1} p_i y_{1i} - \sum_{j=n_1+1}^{n} q_j y_{0j} = \theta \quad (5)$$

- $\hat{\boldsymbol{p}}(\theta)$ and $\hat{\boldsymbol{q}}(\theta)$: Maximizer of (3) under constraints (4) and (5)
- The maximum pseudo EL estimator $\hat{\theta}_3$:
  Maximize $\ell\Big(\hat{\boldsymbol{p}}(\theta), \hat{\boldsymbol{q}}(\theta)\Big)$ w.r.t. θ

A Simulation Study

- $N = 20,000$; $n = 200$; Single stage PPS sampling
- $x_{i1} \sim Bernoulli(0.5)$; $x_{i2} \sim U[0,1]$; $x_{i3} \sim 0.5 + 2\exp(1)$
- $\pi_i \propto x_{i3}$; $\max \pi_i / \min \pi_i = 45$
- Linear models for the responses y_{i0} and y_{i1}:

$$y_{ik} = \beta_{0k} + \beta_{1k} x_{i1} + \beta_{2k} x_{i2} + \beta_{3k} x_{i3} + \varepsilon_i\,, \quad i = 1, 2, \ldots, N$$

- Logistic regression model for R_i: $r_i = P(R_i = 1 \mid \boldsymbol{x}_i)$

$$\log\left(\frac{r_i}{1 - r_i}\right) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3}\,, \quad i = 1, 2, \ldots, N$$

- Three point estimators of $\theta = \mu_{y1} - \mu_{y0}$: $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$
- $B = 1000$ repeated simulation runs

A Simulation Study

Table : **Absolute Relative Bias (ARB, %) and Mean Square Error (MSE)**

		$\theta = \mu_{y1} - \mu_{y0} = 1$			$\theta = \mu_{y1} - \mu_{y0} = 2$		
$E(R)$		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
0.50	ARB	3.2	2.4	2.8	1.4	0.3	0.6
	MSE	0.2	0.1	0.1	0.3	0.1	0.1
0.60	ARB	42.9	17.9	10.6	26.7	8.7	4.9
	MSE	0.4	0.5	0.2	0.6	0.5	0.2
0.70	ARB	104.8	67.2	14.1	61.6	33.8	6.4
	MSE	1.5	12.1	0.7	2.0	12.1	0.7

Post-stratification by Propensity Score

- With self-selection of "treatment" and "control", the distributions of $\boldsymbol{Z} \mid R = 1$ and $\boldsymbol{Z} \mid R = 0$ tend to be different
- Matching by propensity score is an effective way to balance the distribution of $\boldsymbol{Z}$ between the "treated" and the "untreated" (Rosenbaum and Rubin, 1983, 1984)
- Order the units based on fitted propensity scores

$$\hat{r}_{(1)} \leq \hat{r}_{(2)} \leq \cdots \leq \hat{r}_{(n)}$$

- Form K strata based on suitable cut-off of the propensity scores (Popular choice: $K = 5$)
- Units within the same stratum have similar values of propensity scores

Post-stratification by Propensity Score

- Post-stratified samples: $S = Q_1 \cup \cdots \cup Q_K$
- Estimated stratum weights

$$\hat{W}_k = \Big(\sum_{i \in Q_k} d_i\Big) \Big/ \Big(\sum_{i \in S} d_i\Big), \quad k = 1, \cdots, K$$

- Population means

$$\mu_{y1} = \sum_{k=1}^{K} W_k \mu_{y1k} \quad \text{and} \quad \mu_{y0} = \sum_{k=1}^{K} W_k \mu_{y0k}$$

- Treatment and control groups within $Q_k = S_{1k} \cup S_{0k}$:

$$R_i = 1 \quad \text{if} \quad i \in S_{1k} \quad \text{and} \quad R_i = 0 \quad \text{if} \quad i \in S_{0k}$$

Post-stratification by Propensity Score

- For each Q_k, the distributions of $\boldsymbol{Z}$ over S_{1k} and S_{0k} are approximately the same
- Balance diagnostics tools are available (Austin, 2008, 2009)
- The post-stratified estimator of θ:

$$\hat{\theta} = \sum_{k=1}^{K} \hat{W}_k \left(\hat{\mu}_{y1k} - \hat{\mu}_{y0k} \right)$$

$$\hat{\mu}_{y1k} = \frac{\sum_{i \in S_{1k}} d_i y_{1i}}{\sum_{i \in S_{1k}} d_i}, \quad \hat{\mu}_{y0k} = \frac{\sum_{i \in S_{0k}} d_i y_{0i}}{\sum_{i \in S_{0k}} d_i}$$

- Post-stratification provides an effective way of using baseline information on $\boldsymbol{Z}$

Using Z in Pseudo EL Through Additional Constraints

- The pseudo EL function for the post-stratified samples

$$\ell = \sum_{k=1}^{K} \hat{W}_k \sum_{i \in S_{1k}} \tilde{d}_{ik} \log(p_{ik}) + \sum_{k=1}^{K} \hat{W}_k \sum_{i \in S_{0k}} \tilde{d}_{ik} \log(q_{ik})$$

- Constraints

$$\sum_{i \in S_{1k}} p_{ik} = 1 \,, \quad \sum_{i \in S_{0k}} q_{ik} = 1 \,, \quad k = 1, \cdots, K$$

$$\sum_{k=1}^{K} \hat{W}_k \sum_{i \in S_{1k}} p_{ik} y_{1i} - \sum_{k=1}^{K} \hat{W}_k \sum_{i \in S_{0k}} q_{ik} y_{0i} = \theta$$

$$\sum_{k=1}^{K} \hat{W}_k \sum_{i \in S_{1k}} p_{ik} \mathbf{z}_i = \sum_{k=1}^{K} \hat{W}_k \sum_{i \in S_{0k}} q_{ik} \mathbf{z}_i$$

Using Z in Pseudo EL Through Imputation

- For each $Q_k = S_{1k} \cup S_{0k}$:

$$(y_{1i}, z_i)\,, \quad i \in S_{1k} \quad \text{plus} \quad z_i\,, \quad i \in S_{0k}$$

$$(y_{0i}, z_i)\,, \quad i \in S_{0k} \quad \text{plus} \quad z_i\,, \quad i \in S_{1k}$$

- Build a model using $\{(y_{1i}, z_i), i \in S_{1k}\}$
  Predict y_{1i} for $i \in S_{0k}$ using $z_i, i \in S_{0k}$
- Build a model using $\{(y_{0i}, z_i), i \in S_{0k}\}$
  Predict y_{0i} for $i \in S_{1k}$ using $z_i, i \in S_{1k}$
- Using the two imputed stratified samples for pseudo EL inference

1. Empirical Likelihood for Two-sample Problems

2. Survey Samples for Observational Studies

3. Concluding Remarks

Concluding Remarks

- Confidence intervals or hypothesis tests using the EL ratio statistic have better performances than the conventional normal theory methods
- Performances of imputation-based EL approach to pretest-posttest studies depend on two crucial conditions:
  - Prediction power of the baseline variables $\boldsymbol{Z}$
  - Reliability of the model used for imputation
- Linear regression models are convenient, but kernel regression models can also be used
- Efficient computational procedures for EL are available for practical implementations
- We are currently conducting further simulation studies on mode effects in survey data collection

Acknowledgement

This talk is partially based on Min Chen's PhD dissertation research at the University of Waterloo

- Chen, M., Wu, C. and Thompson, M.E. (2015). An Imputation Based Empirical Likelihood Approach to Pretest-Posttest Studies. *The Canadian Journal of Statistics*, **43**, 378–402.
- Chen, M., Wu, C. and Thompson, M.E. (2015). Mann-Whitney Test with Empirical Likelihood Methods for Pretest-Posttest Studies. Revised for *Journal of Nonparametric Statistics*.
- Wu, C. and Yan, Y. (2012). Weighted Empirical Likelihood Inference for Two-sample Problems. *Statistics and Its Interface*, **5**, 345–354.