Published in final edited form as:

Stat Med. 2017 March 15; 36(6): 985–997. doi:10.1002/sim.7195.

Statistical inferences for data from studies conducted with an aggregated multivariate outcome-dependent sample design

Tsui-Shan Lu[a], **Matthew P. Longnecker**[b], and **Haibo Zhou**[c,*]

[a]Department of Mathematics, National Taiwan Normal University, Taipei, Taiwan

[b]National Institute of Environmental Health Sciences, National Institute of Health, Research Triangle Park, NC 27709

[c]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Abstract

Outcome-dependent sampling (ODS) scheme is a cost-effective sampling scheme where one observes the exposure with a probability that depends on the outcome. The well-known such design is the case-control design for binary response, the case-cohort design for the failure time data and the general ODS design for a continuous response. While substantial work has been done for the univariate response case, statistical inference and design for the ODS with multivariate cases remain under-developed. Motivated by the need in biological studies for taking the advantage of the available responses for subjects in a cluster, we propose a multivariate outcome dependent sampling (*Multivariate-ODS*) design that is based on a general selection of the continuous responses within a cluster. The proposed inference procedure for the *Multivariate-ODS* design is semiparametric where all the underlying distributions of covariates are modeled nonparametrically using the empirical likelihood methods. We show that the proposed estimator is consistent and developed the asymptotically normality properties. Simulation studies show that the proposed estimator is more efficient than the estimator obtained using only the simple-random-sample portion of the *Multivariate-ODS* or the estimator from a simple random sample with the same sample size. The *Multivariate-ODS* design together with the proposed estimator provides an approach to further improve study efficiency for a given fixed study budget. We illustrate the proposed design and estimator with an analysis of association of PCB exposure to hearing loss in children born to the Collaborative Perinatal Study.

Keywords

Continuous multivariate responses; correlated responses; empirical likelihood; outcome-dependent sampling; semiparametric

*Correspondence to: Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. zhou@bios.unc.edu.

1. Introduction

Epidemiological and other biomedical studies often rely on the observational study to investigate the relationships between a disease outcome and an exposure given other characteristics. Cohort and case-control studies are most commonly used designs. Large cohort study could be prohibitively expensive as observing exposures and response variables could take a long time and costly. Investigators often seek to increase the study power for a given budget by doing retrospective design, especially when the disease is rare. Case-control design is the most famous retrospective study design that increases the study power by allowing investigators to oversample the diseased subjects relative to those free of disease [1–7]. A similar idea of target sampling of the more informative subjects can also be found in the case-cohort design [8], the general outcome-dependent sampling (ODS) design for a continuous outcome [9–12]. Such type of the design, where the sample is dependent on the response, is also referred to as the choice-based sampling in econometrics. The key advantage of those ODS designs is that it allows the researchers to concentrate budgetary resources on observations with the greatest amount of information.

While substantial progress has been done in the univariate response variable case, there is little work for the multivariate response case, especially when the responses are continuous. In practice, multivariate data arise in many contexts, for example, in epidemiological cohort studies where the outcomes are recorded for members within the families, in the animal experiments in which treatments are applied to the samples of littermates, or in most clinical trials where the study subjects are experiencing multiple events. Among these studies, the correlation between the responses from the same cluster cannot be neglected. An increasing number of studies are indeed performed using the multivariate outcome-dependent sampling design (*Multivariate-ODS*), a further generalization of the biased sampling, which is built on the idea of the ODS design with an aggregate of the responses in the multivariate form and at the same time preserves the advantages of the ODS design. The usual statistical methods for analyzing the multivariate data collected from a *Multivariate-ODS* design is no longer appropriate. New statistical inference procedures need to be developed to take advantage of the *Multivariate-ODS* design.

Our research is motivated by the Collaborative Perinatal Project (CPP), a prospective cohort study to identify determinants of neurodevelopmental deficits in children [13–14]. Longnecker *et al.* [15] studied the association in humans between maternal third trimester serum polychlorinated biphenyls (PCBs) levels and audiometry results in offsprings at approximately 8 years old. The sample selected by the investigators was based on the following ODS scheme: an overall simple random sample (SRS) (about 1200 subjects) selected at random from the underlying population and an additional supplemental sample (about 200) from the children whose 8-year audiometric evaluation showed sensorineural hearing loss (SNHL). The SNHL was defined as a hearing threshold ≥ 13.3 dB based on the average of the measurements from the right and left ears at 1000, 2000, and 4000 Hz, in conjunction with no evidence of conductive hearing loss. It is desirable to model the children's left and right ears' hearing abilities in a multivariate framework in relation to PCBs while simultaneously taking the nature of the underlying *Multivariate-ODS* design into account.

In this article we consider statistical inferences on multivariate regression models under a *Multivariate-ODS* design. The innovation of this paper is that we proposed a *Multivariate-ODS* design that is based on the aggregated responses in a cluster. Our proposed estimator enables the investigator to analyze the data done in a multivariate fashion while taking advantage of ODS design. On a theoretical front, the proposed estimator is robust as we leave the underlying distributions of covariates unspecified. We model nonparametrically using the empirical likelihood methods. The simulation results show that the proposed estimator with the multivariate outcome-dependent nature accounted for is more efficient and statistically powerful than alternative estimators. We also explore that the sampling strategies under the *Multivariate-ODS* framework can be used to design a cost-effective study.

The remainder of the paper is organized as follows. In Section 2, we present the notation and the data structure under the *Multivariate-ODS* design with multivariate continuous outcomes. We also present the semiparametric empirical likelihood approach and derive the asymptotic properties. In Section 3, we describe simulation studies and compare the small sample properties of our proposed estimator with other estimators. We apply the proposed method to analyze the data in Collaborative Perinatal Project study in Section 4. We give final remarks in Section 5.

2. The multivariate-ODS design and inference

2.1. The multivariate-ODS data structure and likelihood

Let Y_{ij} be the jth continuous outcome for the subject i, where $i = 1,\ldots, n$ and $j = 1,\ldots, p$ ($p \geq 2$), and $\boldsymbol{X}_i$ be a vector of covariates for the ith subject, which can include both discrete and continuous components. Let $\boldsymbol{a} = \{a_j, j = 1,\ldots, p\}$ and $\boldsymbol{b} = \{b_j, j = 1,\ldots, p\}$, where a_j and b_j are known constants and satisfying $\{a_j > b_j, \forall j\}$, are the fixed cutpoints on the domain of $\boldsymbol{Y}_j = \{Y_{ij}, \forall i\}$. The data structure of the *Multivariate-ODS* design consists of three components: an overall *simple random sample* (SRS) of size n_0 (≥ 0), a *supplemental sample* of size n_1 (≥ 0) conditional on $\{Y_{i1} > a_1,\ Y_{i2} > a_2,\ldots,\ Y_{ip} > a_p\}$, and another *supplemental sample* of size n_2 conditional on $\{Y_{i1} < b_1,\ Y_{i2} < b_2,\ldots,\ Y_{ip} < b_p\}$:

- **i.** SRS Component: $\{\boldsymbol{Y}_i, \boldsymbol{X}_i\}$, $i = 1, \cdots, n_0$;
- **ii.** Supplemental Component 1: $\{\boldsymbol{Y}_i, \boldsymbol{X}_i \mid \{Y_{i1} > a_1,\ Y_{i2} > a_2,\ldots,\ Y_{ip} > a_p\}\}$, $i = 1,\ldots, n_1$ and $j = 1,\ldots, p$;
- **iii.** Supplemental Component 2: $\{\boldsymbol{Y}_i, \boldsymbol{X}_i \mid \{Y_{i1} < b_1,\ Y_{i2} < b_2,\ldots,\ Y_{ip} < b_p\}\}$, $i = 1,\ldots, n_2$ and $j = 1,\ldots, p$;

the total sample size in the *Multivariate-ODS* is $n = \sum_{k=0}^{2} n_k$. Without loss of generality, we assume that $p = 2$, i.e., each individual has two responses, and the cutpoints are set to be a_1, a_2, b_1 and b_2.

Let $f(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta})$ be the conditional density function of $\boldsymbol{Y}$ given $\boldsymbol{X}$, $\boldsymbol{\theta}$ be a vector of the regression coefficients of interest, and $g_{\boldsymbol{X}}(\boldsymbol{X})$ be the marginal density of $\boldsymbol{X}$, which is independent of $\boldsymbol{\theta}$. Then the joint density of $(\boldsymbol{Y}, \boldsymbol{X})$ can be written as $f(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X})$. The

corresponding unknown cumulative distribution function of $\boldsymbol{X}$ is denoted as $G_{\boldsymbol{X}}(\boldsymbol{X})$. The joint likelihood function for the observed data obtained through the *Multivariate-ODS* design is

$$L(\boldsymbol{\theta}, G_{\boldsymbol{X}}) = \left[\prod_{i=1}^{n_0} f(Y_{i1}, Y_{i2}, \boldsymbol{X}_i; \boldsymbol{\theta})\right] \left[\prod_{i=1}^{n_1} f(Y_{i1}, Y_{i2}, \boldsymbol{X}_i; \boldsymbol{\theta} | Y_{i1} > a_1, Y_{i2} > a_2)\right] \times \left[\prod_{i=1}^{n_2} f(Y_{i1}, Y_{i2}, \boldsymbol{X}_i; \boldsymbol{\theta} | Y_{i1} < b_1, Y_{i2} < b_2)\right], \quad (1)$$

where the first bracket is the likelihood corresponding to the observations from the SRS portion of the *Multivariate-ODS* while the last two parts are contributions from the two supplemental samples. Using Bayes' Law, we can further rewrite the likelihood function (1) as

$$\begin{aligned} L(\boldsymbol{\theta}, G_{\boldsymbol{X}}) &= \left[\prod_{i=1}^{n_0} f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}_i)\right] \left[\prod_{i=1}^{n_1} \frac{f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}_i)}{\Pr(Y_{i1} > a_1, Y_{i2} > a_2)}\right] \times \left[\prod_{i=1}^{n_2} \frac{f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}_i)}{\Pr(Y_{i1} < b_1, Y_{i2} < b_2)}\right] \\ &= \left[\prod_{i=1}^{n_0} f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}_i)\right] \left[\prod_{i=1}^{n_1} \frac{f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}_i)}{\pi_1(\boldsymbol{\theta}, G_{\boldsymbol{X}})}\right] \times \left[\prod_{i=1}^{n_2} \frac{f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}_i)}{\pi_2(\boldsymbol{\theta}, G_{\boldsymbol{X}})}\right] \\ &= \left[\prod_{i=1}^{n} f(Y_{i1}, Y_{i2} | \boldsymbol{X}_i; \boldsymbol{\theta})\right] \left[\left(\prod_{i=1}^{n} g_{\boldsymbol{X}}(\boldsymbol{X}_i)\right) \pi_1^{-n_1} \pi_2^{-n_2}\right] \\ &= L_{GL1}(\boldsymbol{\theta}) \times L_{GL2}(\boldsymbol{\theta}, G_{\boldsymbol{X}}), \end{aligned} \quad (2)$$

where $P_1(\boldsymbol{X}; \boldsymbol{\theta}) = \Pr\{Y_1 > a_1, Y_2 > a_2 | \boldsymbol{X}\} = \int_{a_1}^{\infty}\int_{a_2}^{\infty} f(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta}) dY_1 dY_2$ and $\boldsymbol{\pi_1} = \boldsymbol{\pi_1}(\boldsymbol{\theta}, G_{\boldsymbol{X}}) = \int_{\mathcal{X}} P_1(\boldsymbol{x}; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}) d\boldsymbol{X}$ are the conditional and marginal probabilities that Y_1 and Y_2 satisfy $\{Y_1 > a_1, Y_2 > a_2\}$; $P_2(\boldsymbol{X}; \boldsymbol{\theta}) = \Pr\{Y_1 < b_1, Y_2 < b_2 | \boldsymbol{X}\} = \int_{-\infty}^{b_1}\int_{\infty}^{b_2} f(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta}) dY_1 dY_2$ and $\boldsymbol{\pi_2} = \boldsymbol{\pi_2}(\boldsymbol{\theta}, G_{\boldsymbol{X}}) = \int_{\mathcal{X}} P_2(\boldsymbol{x}; \boldsymbol{\theta}) g_{\boldsymbol{X}}(\boldsymbol{X}) d\boldsymbol{X}$ are the conditional and marginal probabilities for $\{Y_1 < b_1, Y_2 < b_2\}$.

To make inferences about $\boldsymbol{\theta}$, several naive approaches can be used. First, one could simply use the SRS portion of the *Multivariate-ODS* and derive a maximum likelihood estimator for $\boldsymbol{\theta}$. However, this approach will not utilize the information from the supplemental sample. If $G_{\boldsymbol{X}}(\boldsymbol{X})$ is parameterized to a parameter vector, say ξ, one could also maximize the resulting $L(\boldsymbol{\theta}, \hat{G}_{\boldsymbol{X}})$ subject to $(\boldsymbol{\theta}, \xi)$. However, misspecification of $G_{\boldsymbol{X}}$ could lead to erroneous conclusions. A nonparametric modeling of $G_{\boldsymbol{X}}$ is desirable in this case, though $G_{\boldsymbol{X}}$ will be an infinite-dimensional nuisance parameter and will not be factored out of $L(\boldsymbol{\theta}, G_{\boldsymbol{X}})$. Thus, to incorporate all the available information in the *Multivariate-ODS* data without specifying $G_{\boldsymbol{X}}$, one needs a new method that will be tractable both theoretically and computationally. We next describe a semiparametric empirical likelihood estimator, where $G_{\boldsymbol{X}}$ is left unspecified.

2.2. A semiparametric likelihood approach for the multivariate-ODS

To outline our approach for estimating $\boldsymbol{\theta}$, we develop a profile likelihood function for $\boldsymbol{\theta}$ by first maximizing $L(\boldsymbol{\theta}, G_{\boldsymbol{X}})$ with $\boldsymbol{\theta}$ fixed and $G_{\boldsymbol{X}}$ treated as a nonparametric maximum likelihood estimate (NPMLE) [16], which will become a function of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. Then we can obtain $\hat{\boldsymbol{\theta}}$ by maximizing the resulting profile log likelihood function over $\boldsymbol{\theta}$. The procedure is detailed in the following.

We first maximize $L(\boldsymbol{\theta}, G_{\boldsymbol{X}})$, with $\boldsymbol{\theta}$ fixed, over all discrete distributions whose support includes the observed values by considering a discrete distribution function (i.e. a step function) which has all of its probability located at the observed data points [16]. Let $p_i = dG_{\boldsymbol{X}}(\boldsymbol{X}_i) = g_{\boldsymbol{X}}(\boldsymbol{X}_i)$, $i = 1,\ldots, n$, be the probability mass for the ith covariate vector. We want to find values $\{\hat{p}_i\}$, which maximize the log likelihood function corresponding to (2)

$$l(\boldsymbol{\theta}, \{p_i\}) = \sum_{i=1}^{n} \ln f(\boldsymbol{Y}_i | \boldsymbol{X}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln p_i - n_1 \ln \pi_1 - n_2 \ln \pi_2, \quad (3)$$

under the following constraints:

$$\left\{ \{p_i\} \geq 0 \forall i, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \ (P_1(\boldsymbol{X}_i; \boldsymbol{\theta}) - \pi_1) = 0, \sum_{i=1}^{n} p_i \ (P_2(\boldsymbol{X}_i; \boldsymbol{\theta}) - \pi_2) = 0 \right\}. \quad (4)$$

The above conditions reflect the fact that $G_{\boldsymbol{X}}$ is a discrete distribution function. For a fixed $\boldsymbol{\theta}$, there exists a unique maximum for $\{p_i\}$ in (3) subject to the constraints in (4) if 0's are inside the convex hull of the points $\{P_1(\boldsymbol{X}_i; \boldsymbol{\theta}), \forall\ i\}$ and $\{P_2(\boldsymbol{X}_i; \boldsymbol{\theta}), \forall\ i\}$ [17]. We use the Lagrange multiplier argument to obtain $\hat{p}_i$ through maximizing H,

$$H(\boldsymbol{\theta}, \{p_i\}, \delta, \lambda_1, \lambda_2) = \sum_{i=1}^{n} \ln p_i - n_1 \ln \pi_1 - n_2 \ln \pi_2 - \delta \left(\sum_{i=1}^{n} p_i - 1 \right) - n\lambda_1 \sum_{i=1}^{n} p_i \ (P_1(\boldsymbol{X}_i; \boldsymbol{\theta}) - \pi_1) - n\lambda_2 \sum_{i=1}^{n} p_i \ (P_2(\boldsymbol{X}_i; \boldsymbol{\theta}) - \pi_2),$$

(5)

where δ, λ_1 and λ_2 are the Lagrange multipliers corresponding to the normalized restriction on the $\{\hat{p}_i\}$. With $\boldsymbol{\theta}$ fixed, taking the derivative of H with respect to p_i, solving the score equation and applying the constraints in (4), we obtain $\hat{\delta} = n$ and

$$\hat{p}_i = \{ n \ [1 + \lambda_1 \ (P_1(\boldsymbol{X}_i; \boldsymbol{\theta}) - \pi_1) + \lambda_2 \ (P_2(\boldsymbol{X}_i; \boldsymbol{\theta}) - \pi_2)] \}^{-1}. \quad (6)$$

Substituting $\{\hat{p}_i\}$ back into (3), we then have the resulting profile log likelihood function,

$$l(\boldsymbol{\phi})=\sum_{i=1}^{n}\ln f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\boldsymbol{\theta})-\sum_{i=1}^{n}\ln n\ [1+\lambda_1\ (P_1(\boldsymbol{X}_i;\boldsymbol{\theta})-\pi_1)+\lambda_2\ (P_2(\boldsymbol{X}_i;\boldsymbol{\theta})-\pi_2)]-n_1\ln\pi_1-n_2\ln\pi_2. \quad (7)$$

where $\boldsymbol{\phi}^T=(\boldsymbol{\theta}^T, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \lambda_1, \lambda_2)$ represents a combined parameter vector; λ_1, λ_2, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are treated as the parameters independent of $\boldsymbol{\theta}$. We refer $\hat{\boldsymbol{\phi}}$, a maximizer for (7), as the *semiparametric empirical maximum likelihood estimator* (SEMLE). The Newton-Raphson iterative algorithm is used to solve the score equation from (7).

2.3. Asymptotic properties of the SEMLE

The main results for $\boldsymbol{\phi}$ regarding the existence and consistency, asymptotic normality, and a consistent estimator for the asymptotic variance-covariance matrix are demonstrated as three theorems, respectively. Outlines of the proofs of the main results are provided in the Appendix 1.

We indicate $\boldsymbol{\phi}^0$ as the true parameter vector of interest containing $\boldsymbol{\theta}^0$, π_1^0, π_2^0, λ_1^0 and λ_2^0, where π_1^0 and π_2^0 are the true marginal probabilities that $\{Y_1 > a_1,\ Y_2 > a_2\}$ and $\{Y_1 < b_1,\ Y_2 < b_2\}$, respectively; λ_1^0 and λ_2^0 are the true Lagrange multipliers.

Theorem 1 (Consistency of the SEMLE)—With probability going to 1 as $n \to \infty$, there exists a sequence $\{\hat{\boldsymbol{\phi}}\}$ of solutions to the score equations (7) such that $\hat{\phi} \xrightarrow{p} \phi^0$, where $\boldsymbol{\phi}^0$ is the true parameter vector of interest. If another sequence $\{\bar{\boldsymbol{\phi}}\}$ of solutions to the score equations exists such that $\overline{\phi} \xrightarrow{p} \phi^0$, then $\bar{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}$ with probability going to 1 as $n \to \infty$.

Theorem 2 (Asymptotic Normality of the SEMLE)—The SEMLE has the following asymptotic normal distribution: $\sqrt{n}(\hat{\phi}-\phi^0) \xrightarrow{D} N_{(p+2)}(\mathbf{0}, \sum(\phi^0))$, with the asymptotic variance-covariance matrix

$$\sum=\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}, \quad (8)$$

where $\boldsymbol{J}=-\dfrac{\partial^2\tilde{l}(\phi^0)}{\partial\phi\partial\phi^T}$ and $\boldsymbol{V}=Var\left[\dfrac{\partial l(\boldsymbol{Y},\boldsymbol{X};\phi^0)}{\partial\phi}\right]$, where $\tilde{l}$ is the limiting form of l.

Theorem 3 (A Consistent Estimator for the Asymptotic Variance-Covariance Matrix)—A consistent estimator for the variance-covariance matrix shown in Equation (8)

is $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\phi}}) = \hat{\boldsymbol{J}}^{-1}(\hat{\boldsymbol{\phi}})\hat{\boldsymbol{V}}(\hat{\boldsymbol{\phi}})\hat{\boldsymbol{J}}^{-1}(\hat{\boldsymbol{\phi}})$, where $\hat{\boldsymbol{J}}(\phi)=-\frac{1}{n}\frac{\partial^2 l(\phi)}{\partial\phi\partial\phi^T}$ and $\hat{\boldsymbol{V}}(\phi)=\frac{1}{n}\widehat{Var}_{\{i\}}\left[\frac{\partial l(\boldsymbol{Y}_i,\boldsymbol{X}_i;\phi^0)}{\partial\phi}\right]$.

3. Simulation studies

In this section, we evaluate the performance of the proposed estimator in the small sample settings. We compare our proposed estimator, $\hat{\boldsymbol{\theta}}_P$, to the following competitive estimators under each setting in our simulation study: (i) the maximum likelihood estimator from the SRS portion of the *Multivariate-ODS* data ($\hat{\boldsymbol{\theta}}_R$), (ii) the maximum likelihood estimator by maximizing the conditional likelihood, L_{GL1}, based on the complete *Multivariate-ODS* data ($\hat{\boldsymbol{\theta}}_C$), and (iii) the maximum likelihood estimator obtained from a random sample of the same size as the *Multivariate-ODS* sample ($\hat{\boldsymbol{\theta}}_S$). Comparing $\hat{\boldsymbol{\theta}}_P$ with $\hat{\boldsymbol{\theta}}_R$ and $\hat{\boldsymbol{\theta}}_C$ will give us an insight of the impact on ignoring the part of the information from the *Multivariate-ODS* sample. The comparison between $\hat{\boldsymbol{\theta}}_P$ and $\hat{\boldsymbol{\theta}}_S$ will demonstrate the efficiency gain of the *Multivariate-ODS* design over the simple random sample of the same size.

We consider the following bivariate normal model to generate the simulated data:

$$Y_{i1}, Y_{i2}|X_{i1}, X_{i2}\sim N\left(\boldsymbol{\mu}_i=\begin{bmatrix}\mu_{i1}\\ \mu_{i2}\end{bmatrix}, \sum\nolimits_i=\begin{bmatrix}\sigma_1{}^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2{}^2\end{bmatrix}\right),$$

where $\mu_{i1} = \alpha_1 + \beta_1 X_{i1}$ and $\mu_{i2} = \alpha_2 + \beta_2 X_{i2}$; i.e., the conditional distributions of Y_{i1} given X_{i1} and Y_{i2} given X_{i2} are normally distributed with means $\alpha_1 + \beta_1 X_{i1}$ and $\alpha_2 + \beta_2 X_{i2}$, variances σ_1^2 and σ_2^2, respectively, and the correlation coefficient is ρ. Our goal is to estimate the parameter vector, $\boldsymbol{\theta}_P = (\alpha_1, \beta_1, \alpha_2, \beta_2, \sigma_1, \sigma_2, \rho)^T$. We will investigate the behavior of $\hat{\beta}_1$ and $\hat{\beta}_2$ under a variety of configurations of the parameter values. In particular, we choose $\alpha_1 = 0.5$, $\alpha_2 = -0.8$, $\sigma_1^2=\sigma_2^2=1$ and $\rho = 0.5$ and 0.85 with $X_1 = X_2 \sim N(0, 1)$.

The study *Multivariate-ODS* sample sizes for investigation were $n = 200$ and $n = 800$. The *Multivariate-ODS* design consisted an overall SRS of size n_0 supplemented with two additional samples of sizes n_1 and n_2 separately from individuals whose outcome values fall in the two tails of the outcome distributions. We also considered two settings of the cutpoints: (i) the upper tails of the 90th percentiles from the distributions of $\{Y_{i1}, \forall i\}$ and $\{Y_{i2}, \forall i\}$ and the lower tails of the 10th percentiles of the distributions, and (ii) the upper tails of the 70th percentiles and the lower tails of the 30th percentiles. The cutpoints chosen for each setting were fixed through all the simulation runs. For all simulation studies, we generated 1,000 simulated data sets, each with an underlying population of size $N = 10,000$ as a basis and then followed the *Multivariate-ODS* design to obtain n_0, n_1 and n_2. The mean of the parameter estimates, the sample standard deviation (SSD) of the 1,000 estimates and the mean of the estimated standard errors (ESE) were computed for the proposed method and other competing methods, and the nominal 95% confidence intervals were calculated based on their asymptotic normal distributions.

The simulation results were presented in Tables 1 and 2. Within each table, the sampling specifications and the covariate distribution were fixed. Table 1 provided the results for the sample size of $n = 200$ and the correlation coefficients of $\rho = 0.5$ and 0.85, and included the small sample properties of the proposed estimator $\hat{\boldsymbol{\theta}}_P$ and the competing estimators from $\hat{\boldsymbol{\theta}}_C$, $\hat{\boldsymbol{\theta}}_R$ and $\hat{\boldsymbol{\theta}}_S$. Table 2 presented the relative efficiencies (ratios of variances) to evaluate the amount of information gained by implementing the *Multivariate-ODS* design.

For $\rho = 0.5$ in Table 1, we observed that three methods yielded unbiased means of the estimates compared with the "true" parameter values for all four settings. The proposed estimator $\hat{\boldsymbol{\theta}}_P$ is the most efficient one among all the estimators compared. For $\hat{\boldsymbol{\theta}}_P$, the means of the estimated standard errors (ESE) were close to the "true" simulated sample standard deviations (SSD), meaning that $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_P)$ provided a very good estimate of the true variability. The confidence intervals based on the proposed estimator provided good coverage close to the nominal 95% level. The same findings were also observed for $\hat{\boldsymbol{\theta}}_R$ and $\hat{\boldsymbol{\theta}}_C$. Within the same sampling design, the standard errors of $\hat{\boldsymbol{\theta}}_P$ decreased as the percentiles of the cutpoints increased. For example, the SSD was dropped from 0.066 when U = 70% and L = 30% for $n_1 = n_2 = 20\%$ to 0.062 when U = 90% and L = 10%, indicating that our proposed method was even more efficient and favored when the supplemental samples included more extreme observations. With the cutpoints fixed, the SSD for $\hat{\boldsymbol{\theta}}_P$ decreased as the proportion of the supplemental samples increased. These observations were true for both $\hat{\beta}_1$ and $\hat{\beta}_2$.

In Table 1, we also presented the results for a higher correlation coefficient (the case of $\rho = 0.85$). First, the observations seen for $\rho = 0.5$ above held for $\rho = 0.85$. Secondly, when comparing the results for $\hat{\boldsymbol{\theta}}_P$ across $\rho = 0.5$ to 0.85, we observed that the SSDs for $\hat{\boldsymbol{\theta}}_P$ were smaller for $\rho = 0.85$ than those for $\rho = 0.5$ in all cases considered. For example, the SSD for $\hat{\beta}_2$ when U = 70%, L = 30%, and $n_1 = n_2 = 5\%$ was 0.71 for $\rho = 0.5$ whereas the corresponding SSD for $\rho = 0.85$ was 0.68. This suggested that the proposed estimator be more efficient with high-correlated outcomes.

Table 2 presented the results from a relative efficiency study by comparing the *Multivariate-ODS* design with the SRS portion only and the design of a simple random sample with the same sample size, under the same settings studied in Table 1 with various sampling fractions for $n = 200$ and $n = 800$. We calculated the asymptotic relative efficiencies (*ARE*) of $\hat{\boldsymbol{\theta}}_P$ to $\hat{\boldsymbol{\theta}}_S$ and $\hat{\boldsymbol{\theta}}_P$ to $\hat{\boldsymbol{\theta}}_R$, $\widehat{ARE}_S\,(=Var_{\hat{\boldsymbol{\theta}}_S}/Var_{\hat{\boldsymbol{\theta}}_P})$ and $\widehat{ARE}_R\,(=Var_{\hat{\boldsymbol{\theta}}_R}/Var_{\hat{\boldsymbol{\theta}}_P})$, respectively. We observed that most of the *AREs* were greater than 1, suggesting that $\hat{\boldsymbol{\theta}}_P$ was more efficient than $\hat{\boldsymbol{\theta}}_R$ and $\hat{\boldsymbol{\theta}}_S$ under all the circumstances. It is clear that $\hat{\boldsymbol{\theta}}_P$ led to more efficiency gains over $\hat{\boldsymbol{\theta}}_S$ as the proportion of the supplemental data in the *Multivariate-ODS* increased with the cutpoints held the same. We also observed more efficiency gains when the two outcomes were more correlated. Combining these two observations, we can see that the greatest efficiency gain in Table 2 was when $\rho = 0.85$, U = 90%, L = 10%, and $n_1 = n_2 = 25\%$.

To investigate the effect of changing the supplemental sampling fractions on the improvement of the *Multivariate-ODS* design over the simple-random-sample design as the CPP study employed, we conducted several simulation experiments using the same simulation model in Tables 1 with $n = 200$, U = 90%, and L = 10%. Figure 1 illustrated the

standard errors (dotted line) of $\hat{\theta}_P$ for $\hat{\beta}_1$ with $\rho = 0.5$ and the relative efficiencies (solid line) of the *Multivariate-ODS* design to a simple-random-sample design across various supplemental sampling fractions (the number on the horizontal axis represents $(n_1 + n_2)/n$). It is clear that $\hat{\theta}_P$ was more efficient than $\hat{\theta}_S$ as all the *AREs* were greater than 1. Furthermore, the most *AREs* gain was around the proportions of the supplemental samples were between 0.3 ad 0.6.

4. Analysis of the Collaborative Perinatal Project data set

We applied the proposed method to analyze the Collaborative Perinatal Project (CPP) data to study the effect of the third trimester maternal pregnancy serum level of polychlorinated biphenyls (PCBs) on hearing loss children. Nearly 56,000 pregnant women were recruited into the CPP study from 1959 through 1966 through 12 study centers across the United States. Women were enrolled, usually at their first prenatal visit; it resulted in 55,908 pregnancies (9,161 women contributed multiple pregnancies to the study). Data were collected on the mothers at each prenatal visit and at delivery and when the children were 24 hours, 4 and 8 months, and 1, 3, 4, 7, and 8 years. Among all the measures, we were interested in audiometric evaluation, which was done when the children were approximately 8 years old. In our selection of the subjects, we follow the selection criteria and the sampling scheme used in Longnecker et al. [15]. There were 44,075 eligible children who met the following criteria: (1) live born singleton, and (2) a 3-ml third trimester maternal serum specimen was available. The audiometric evaluations showed sensorineural hearing loss (SNHL) was defined by a hearing threshold ≥ 13.3 dB according to the average across both ears at 1000, 2000, and 4000 Hz, without any evidence of conductive hearing loss. Evidence of conductive hearing loss exists when the air-bone difference in hearing threshold is ≥ 10 dB again based on the average across both ears.

We considered the subjects who did not have missing observations for the variables selected into the model fitting and we assumed that missing data were missing completely at random. Of the 44,075 eligible children, 1,200 subjects were selected at random, of which 729 had complete data for the variables mentioned above and will then represent the study population in our data analysis. In order to adjust for our selection criterion described in the previous section, we considered the first and third quartiles of the distributions of hearing levels for each ear as the cutpoints. Hence, 100 out of 729 subjects were those whose hearing level measurements were both above the third quartiles, and 122 children had hearing measurements both below the first quartiles. To illustrate our proposed method with the application of real data, we considered the following design with the total sample size $n = 200$ under the *Multivariate-ODS* design: an overall simple random sample of size $n_0 = 100$ from 729 supplemented with additional samples of $n_1 = 50$ and $n_2 = 50$ separately drawn from the remaining subjects in each group. The exposure variable of interest was PCB measured in $\mu g/L$. Additional factors considered potentially confounding included, for the mother, age (AGE), the socioeconomic index score (SEI) and the highest education level attained when giving birth (EDUC), and the race (RACE) and the gender (SEX) for children. After examining the distributions of the hearing levels across three frequencies for each ear, we transformed the outcome variables on the natural log scale in order to exploit the normal properties. We therefore fitted the following linear model to the CPP *Multivariate-ODS* data,

$$\ln(Hearing_{ij}) = \beta_{0j} + \beta_{1j} PCB_i + \beta_{2j} SEX_{ij} + \beta_{3j} RACE_{ij} + \beta_{4j} AGE_{ij} + \beta_{5j} EDUC_{ij} + \beta_{6j} SEI_{ij} + \varepsilon_j, \quad (9)$$

where $\varepsilon_j \sim N(0, \sigma_j^2)$, $i = 1, \ldots, 200$ and $j = 1$ representing the hearing level across three frequencies from the left ear and $j = 2$ from the right ear; $\rho = Corr(\varepsilon_1, \varepsilon_2)$. We assumed that $f(\boldsymbol{Y} | \boldsymbol{X}; \boldsymbol{\theta})$ is bivariate normal, where $\boldsymbol{\theta}^T = (\beta_1^T, \beta_2^T, \sigma_1^2, \sigma_2^2)$ and $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{6j})$ and $j = 1, 2$. We estimated the parameters using the methods considered in the simulation studies: the proposed estimator $\boldsymbol{\theta}_P$ and the competing estimators, $\boldsymbol{\theta}_R$ and $\boldsymbol{\theta}_S$.

Table 3 presented the results of the parameter estimates, the estimated standard errors and the 95% confidence intervals calculated based on the asymptotic normal distributions for the proposed method $\hat{\boldsymbol{\theta}}_P$ and the competing methods $\hat{\boldsymbol{\theta}}_R$ and $\boldsymbol{\theta}_S$. Three methods all showed that the corresponding 95% confidence intervals for the PCB effect included 0. Thus, we concluded that *in utero* PCB exposure did not have a significant effect on hearing levels for both ears. Observing the confidence intervals for other confounding parameters for the left ear, the covariate RACE showed a significant effect at the nominal level of 0.05, agreed by the three methods; however, for the right ear, the significance was detected only in $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_P$. The results suggested that white children had negative impact on hearing loss; in other words, white children were more likely to have better hearing ability than black and other children. Observing the confidence intervals for other covariates, AGE showed a significance on the borderline for the right ear with $\hat{\boldsymbol{\theta}}_R$.

Although PCB was not significant, we could still see some efficiency gains from the results; the observed 95% confidence intervals for PCB provided by the proposed estimator $\hat{\boldsymbol{\theta}}_P$ were narrower for both ears, compared with the CIs obtained by $\hat{\boldsymbol{\theta}}_R$; for example, for the left ear in Table 3, the CI was (−0.037, 0.067) for $\hat{\boldsymbol{\theta}}_P$ versus (−0.063, 0.084) for $\hat{\boldsymbol{\theta}}_R$ and (−0.058, 0.073) for $\hat{\boldsymbol{\theta}}_S$. It indicated that the proposed estimator provides more precise estimates. Moreover, $\hat{\boldsymbol{\theta}}_P$ obtained relatively smaller standard error estimates for all the variables in the model for both ears than those from $\hat{\boldsymbol{\theta}}_R$. Hence, there were observable benefits of using the proposed method and taking the advantage of the *Multivariate-ODS* design.

5. Discussion

Much research has been discussed for multivariate continuous data, of which is a common and important form; nevertheless, the methods accounting for the *Multivariate-ODS* design are lacking. Throughout previous sections, we have demonstrated the need for developing the statistical inferences on the *Multivariate-ODS* and proposed a semiparametric empirical likelihood method for multivariate continuous outcomes. The proposed estimator is semiparametric in nature that the underlying distributions of the covariates are modeled nonparametrically using the empirical likelihood methods. We have shown that the proposed estimator is consistent and asymptotically normally distributed and a consistent estimator for the asymptotic variance-covariance exists, by incorporating additional information into such

Multivariate-ODS design process. We used simulated data generated from a standard linear regression model with Normal errors to examine the performance and the small-sample properties of our proposed estimator. Our limited simulation results indicated that the proposed estimator, $\boldsymbol{\theta}_P$, holds well for all the properties and is more efficient than $\boldsymbol{\theta}_R$, which only takes the simple random sample into consideration, and $\boldsymbol{\theta}_C$, the conditional estimator, using the complete *Multivariate-ODS* data but ignoring additional information in the supplemental sample. For the relative efficiency studies, we observed that $\boldsymbol{\theta}_P$ exhibits more efficiency gains than $\boldsymbol{\theta}_S$, using a simple random sample of the same size as the *Multivariate-ODS* from the underlying population, in terms of different correlation coefficients between the outcomes, the allocations of the cutpoints and the the supplemental fractions. The proposed method under unequal variances resulted in consistent performance with what we obtained from the equal-variance case (Table 4, Appendix 2). We conclude that the *Multivariate-ODS* design, combined with an appropriate analysis, can provide a cost-effective approach to further improve study efficiency, for a given sample size. Finally, we applied the proposed method to the Collaborative Perinatal Project data, where the researchers are interested in studying the association between a child's hearing loss and *in utero* exposure to PCBs as well as other covariates of interest. Our results showed that the estimator obtained using the proposed method produced substantially smaller standard errors for both ears than those from the competing methods; moreover, the estimator obtained by $\boldsymbol{\theta}_P$ clearly gained more efficiency and was more precise than the other competing estimators, $\boldsymbol{\theta}_R$ and $\boldsymbol{\theta}_S$, although PCBs could not be concluded as a significant effect.

Our simulated studies also suggest that the higher proportion of the sample sizes of the supplemental samples over the *Multivariate-ODS* sample, the greater the gains of efficiency are, which was similar to the guidance suggested by Zhou *et al.* [9] in using the ODS design concerning these issues under one continuous outcome variable. Further investigation for the sample size determination, the optimal sample allocations, the optimal correlation coefficient between the outcomes and power analyses aimed at multivariate outcomes under the *Multivariate-ODS* is required. We considered two-dimensional multivariate data in this paper; the future work may include the flexibility of incorporating the covariance structures for higher-dimensional data. Our proposed method can also be applied to the quantitative genetics studies, in which the quantitative trait is modeled as a continuous variable; in fact, more and more studies in order to limit the expenses on the DNA analysis are actually adopting the form of the ODS design. We believe that the proposed methods can be a useful tool toward such studies.

References

1. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute. 1951; 11:1269–1275. [PubMed: 14861651]
2. White H. Maximum likelihood estimation of misspecified models. Econometrica. 1982; 50:1–25.
3. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988; 75:11–20.
4. Wang X, Zhou H. A semiparametric empirical likelihood model for biased sampling schemes with auxiliary covariates. Biometrics. 2006; 62:1149–1160. [PubMed: 17156290]

5. Wang X, Wu Y, Zhou H. Outcome and auxiliary-dependent subsampling and its statistical inference. Journal of Biopharmaceutical Statistics. 2009; 19:1132–1150. [PubMed: 20183468]
6. Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. Biometrics. 1994; 50:350–357. [PubMed: 8068835]
7. Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. Statistical Medicine. 1992; 11:769–782.
8. Prentice RL. A case-cohort design for epidemiologic studies and disease prevention trials. Biometrika. 1986; 73:1–11.
9. Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 2002; 58:413–421. [PubMed: 12071415]
10. Weaver MA, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. Journal of the American Statistical Association. 2005; 100:459–469.
11. Zhou H, Chen W, Rissanen T, Korrick S, Hu H, Salonen J, Longnecker MP. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. Epidemiology. 2007; 18:461–468. [PubMed: 17568219]
12. Zhou H, Wu W, Zeng D, Cai J. Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76:197–215. [PubMed: 24737947]
13. Niswander, KR., Gordon, M. The women and their pregnancies: the collaborative perinatal study of the National Institute of Neurological Diseases and Stroke. Saunders; Philadelphia: 1972.
14. Gray KA, Klebanoff MA, Brock JW, Zhou H, Needham L, Longnecker MP. *In Utero* exposure to background levels of polychlorinated biphenyls and cognitive functioning among school-aged children. American Journal of Epidemiology. 2005; 162:17–26. [PubMed: 15961582]
15. Longnecker MP, Hoffman H, Klebanoff MA, Brock JW, Zhou H, Needham L, Adera T, Guo X, Gray KA. *In Utero* exposure to polychlorinated biphenyls and sensorineural hearing loss in 8-year-old children. Neurotoxicology and Teratology. 2004; 26:629–637. [PubMed: 15315812]
16. Vardi Y. Empirical distributions in selection bias models. The Annals of Statistics. 1985; 13:178–203.
17. Qin J, Lawless JF. Empirical likelihood and general estimating equations. Annals of Statistics. 1994; 22:300–325.
18. Foutz RV. On the unique consistent solution to the likelihood equations. Journal of the American Statistical Association. 1977; 72:147–148.
19. Cosslett SR. Maximum likelihood estimator for choice-based samples. Econometrika. 1981b; 49:1289–1316.
20. Amemiya T. Regression analysis when the dependent variable is truncated Normal. Econometrica. 1973; 41:997–1016.

A. Appendix 1: Proofs of Theorems

For any function $h(\boldsymbol{Y}, \boldsymbol{X})$, let $\mathrm{E}_1[h(\boldsymbol{Y}, \boldsymbol{X})]$ and $\mathrm{E}_2[h(\boldsymbol{Y}, \boldsymbol{X})]$ denote expectations conditional on $\{Y_1 > a_1,\ Y_2 > a_2\}$ and $\{Y_1 < b_1,\ Y_2 < b_2\}$, respectively, that

$$\mathrm{E}_1\ [h(\boldsymbol{Y}, \boldsymbol{X})] = \int_{\mathbb{X}} \frac{1}{\pi_1^0} \int_{a_1}^{\infty} \int_{a_2}^{\infty} h(\boldsymbol{y}, \boldsymbol{x}) f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}^0) d\boldsymbol{y} dG_{\boldsymbol{X}}(\boldsymbol{x})$$

and

$$\mathrm{E}_2\ [h(\boldsymbol{Y},\boldsymbol{X})] = \int_{\mathbb{X}} \frac{1}{\pi_2^0} \int_{\infty}^{b_1} \int_{\infty}^{b_2} h(\boldsymbol{y},\boldsymbol{x}) f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}^0) d\boldsymbol{y} dG_{\boldsymbol{X}}(\boldsymbol{x}).$$

We assume the following regularity conditions:

A1 As $n \to \infty$, $\frac{n_1}{n} \to \gamma_1 > 0$, $\frac{n_2}{n} \to \gamma_2 > 0$ and $\frac{n_0}{n} \to 1-\gamma_1-\gamma_2$, where γ_1 is the sampling fraction of the supplemental sample drawn conditional on $\{Y_1 > a_1,\ Y_2 > a_2\}$ and γ_2 represents the allocation of the supplemental sample conditional on $\{Y_1 < b_1,\ Y_2 < b_2\}$ to the *Multivariate-ODS* sample.

A2 The parameter space, $\boldsymbol{\Theta}$, is a compact subset of $\mathbb{R}^p$; $\boldsymbol{\theta}^0$ lies in the interior of $\boldsymbol{\Theta}$; the covariate space, $\mathscr{X}$, is a compact subset of $\mathbb{R}^q$, for some $q \geq 1$.

A3 $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ is continuous in both $\boldsymbol{y}$ and $\boldsymbol{\theta}$ and is strictly positive for all $\boldsymbol{y} \in \mathscr{Y}$, $\boldsymbol{x} \in \mathscr{X}$, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Furthermore, the partial derivatives, $\partial f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})/\partial\theta_i$ and $\partial^2 f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$, for $i, j = 1,\ldots, p$, exist and are continuous for all $\boldsymbol{y} \in \mathscr{Y}$, $\boldsymbol{x} \in \mathscr{X}$, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

A4 Interchanges of differentiation and integration of $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ are valid for the first and second partial derivatives with respect to $\boldsymbol{\theta}$.

A5 The expected value matrix, $\mathrm{E}\left[-\frac{\partial^2 \ln f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]$, is finite and positive definite at $\boldsymbol{\theta}^0$.

A6 There exists a $\delta > 0$ such that for the set $A = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}^0| \leq \delta\}$,

$$\mathrm{E}\left[\sup_A \left|\frac{\partial^2 \ln f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right|\right] < \infty,$$

for $i, j = 1, \ldots, p$.

A7 The derivatives, $\frac{\partial P_1(\boldsymbol{x};\boldsymbol{\theta}^0)}{\theta_j}$ and $\frac{\partial P_2(\boldsymbol{x};\boldsymbol{\theta}^0)}{\theta_j}$, $j = 1,\ldots, p$, are linearly independent. That is, suppose $\boldsymbol{t}$ and $\boldsymbol{s}$ are any $(p \times 1)$ vectors such that

$$\sum_{j=1}^{p} t_j \frac{\partial P_1(\boldsymbol{x};\boldsymbol{\theta}^0)}{\theta_j} = 0$$

and

$$\sum_{j=1}^{p} s_j \frac{\partial P_2(\boldsymbol{x};\boldsymbol{\theta}^0)}{\theta_j} = 0$$

for almost all $\boldsymbol{x} \in \mathscr{X}$ if $\boldsymbol{t} = \boldsymbol{0}$ and $\boldsymbol{s} = \boldsymbol{0}$.

Sketch Proof of Theorem 1 (Consistency)

Using Assumption A1 and the Law of Large Numbers, we have

$$\frac{1}{n}\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \xrightarrow{p} \frac{\partial \tilde{l}(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}},$$

where

$$\frac{\partial \tilde{l}(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} = \mathrm{E}\left[\frac{\partial \ln f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\lambda_1 \frac{\partial P_1(\boldsymbol{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \lambda_2 \frac{\partial P_2(\boldsymbol{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{1+\lambda_1\ (P_1(\boldsymbol{X};\boldsymbol{\theta})-\pi_1)+\lambda_2\ (P_2(\boldsymbol{X};\boldsymbol{\theta})-\pi_2)}\right].$$

Since it is straightforward to see that

$$\frac{\partial \tilde{l}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathbf{0}$$

at the true parameter values, we know that the profile log-likelihood function converges in probability to a continuous, vector-valued function and a root of the likelihood equations exists; i.e.,

$$\frac{1}{n}\frac{\partial l(\boldsymbol{\phi}^0)}{\partial \boldsymbol{\phi}} \xrightarrow{p} \mathbf{0}.$$

Again using the Law of Large Numbers, we can demonstrate that the convergence in probability of

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} \xrightarrow{p} \frac{\partial^2 \tilde{l}_S(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}}$$

is uniform for $\boldsymbol{\phi}$ in an open neighborhood for $\boldsymbol{\phi}^0$, and at the true parameter values,

$$-\frac{\partial^2 \tilde{l}(\boldsymbol{\phi}^0)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} = \boldsymbol{J},$$

which can be shown to be invertible. Finally, by applying Theorem 2 in Foutz' [18] which showed the existence of a consistent solution to the likelihood equations and its uniqueness by using the Inverse Function Theorem, and weakening the requirement of the matrix of

second derivatives of the log likelihood function to be negative definite, the result in Theorem follows.

Sketch Proof of Theorem 2 (Asymptotic Normality)

We first start from a Taylor series expansion of the estimated score function around the true parameter $\boldsymbol{\phi}^0$ evaluated at $\hat{\boldsymbol{\phi}}$,

$$\frac{\partial l(\hat{\boldsymbol{\phi}})}{\partial \boldsymbol{\phi}} = \frac{\partial l(\boldsymbol{\phi}^0)}{\partial \boldsymbol{\phi}} + \frac{\partial^2 l(\tilde{\boldsymbol{\phi}})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0),$$

where $\tilde{\boldsymbol{\phi}} = \kappa \boldsymbol{\phi}^0 + (1 - \kappa)\hat{\boldsymbol{\phi}}$ for some $\kappa \in [0, 1]$, as in Cosslett [19]. The left-hand side of the above equation is equal to zero since our estimator $\hat{\boldsymbol{\phi}}$ has been shown to be a consistent solution to $\partial l(\boldsymbol{\phi})/\partial \boldsymbol{\phi} = \mathbf{0}$; after rearranging,

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) = \left[-\frac{1}{n} \frac{\partial^2 l(\tilde{\boldsymbol{\phi}})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \right]^{-1} \left[\frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\phi}^0)}{\partial \boldsymbol{\phi}} \right].$$

To prove the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)$, it is sufficient to show that $-(1/n)\partial^2 l(\tilde{\boldsymbol{\phi}})/\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T$ converges to an invertible matrix in probability and $(1/\sqrt{n})\partial l(\boldsymbol{\phi}^0)/\partial \boldsymbol{\phi}$ has an asymptotic normal distribution.

From Theorem 1, we have known that $\hat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}^0$, which implies that $\tilde{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}^0$. And we also have shown that

$$\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \xrightarrow{p} \frac{\partial^2 \tilde{l}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}}$$

uniformly for $\boldsymbol{\phi} \in \boldsymbol{U}$. According to Lemma 4 in Amemiya [20], we can see that

$$-\frac{1}{n} \frac{\partial^2 l(\tilde{\boldsymbol{\phi}})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \xrightarrow{p} -\frac{\partial^2 \tilde{l}(\boldsymbol{\phi}^0)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}} = \boldsymbol{J}.$$

Since $\boldsymbol{J}$ is shown to be positive definite, it follows that its inverse exists. By the Central Limit Theorem, we have

$$\frac{1}{\sqrt{n}} \frac{\partial l(\boldsymbol{\phi}^0)}{\partial \boldsymbol{\phi}} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{V}),$$

where

$$\boldsymbol{V}=\mathrm{Var}\left[\frac{\partial l(\boldsymbol{Y},\boldsymbol{X};\phi^0)}{\partial \boldsymbol{\phi}}\right].$$

Finally, we can apply Slutsky's Theorem to conclude that $\sqrt{n}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}^0) \xrightarrow{D} N(\mathbf{0}, \sum(\boldsymbol{\phi}^0))$, where $\boldsymbol{\Sigma} = \boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}$, the asymptotic covariance matrix of $\hat{\boldsymbol{\phi}}$.

Sketch Proof of Theorem 3 (A Consistent Estimator for the Asymptotic Variance-Covariance Matrix)

It is noted that the observations from our *Multivariate-ODS* design are identically-independently-distributed; thus, the sample covariance matrix over the observed values is consistent for $\boldsymbol{\Sigma}(\boldsymbol{\phi})$. Then, it is straightforward to see that

$$\hat{\boldsymbol{V}}(\boldsymbol{\phi})=\frac{1}{n}\widehat{\mathrm{Var}}_{\{i\}}\left[\frac{\partial l(\boldsymbol{Y}_i,\boldsymbol{X}_i;\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right] \xrightarrow{p} \boldsymbol{V}(\boldsymbol{\phi}).$$

By Assumption 3, the components of $\boldsymbol{V}(\boldsymbol{\phi})$ are continuous in $\boldsymbol{\phi}$. We can then use the triangle inequality to obtain that

$$\|\hat{\boldsymbol{V}}(\hat{\boldsymbol{\phi}})-\boldsymbol{V}(\boldsymbol{\phi}^0)\| \leq \|\hat{\boldsymbol{V}}(\hat{\boldsymbol{\phi}})-\boldsymbol{V}(\hat{\boldsymbol{\phi}})\|+\|\boldsymbol{V}(\hat{\boldsymbol{\phi}})-\boldsymbol{V}(\boldsymbol{\phi}^0)\| \xrightarrow{p} 0$$

as n goes to ∞. Furthermore, in the proof of Theorem 2, we have shown that

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\phi}})=-\frac{1}{n}\frac{\partial^2 l(\hat{\boldsymbol{\phi}})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\phi}^0),$$

It then follows that $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\phi}})$ is a consistent estimator of the asymptotic covariance matrix.

B. Appendix 2: Simulation results for unequal variances

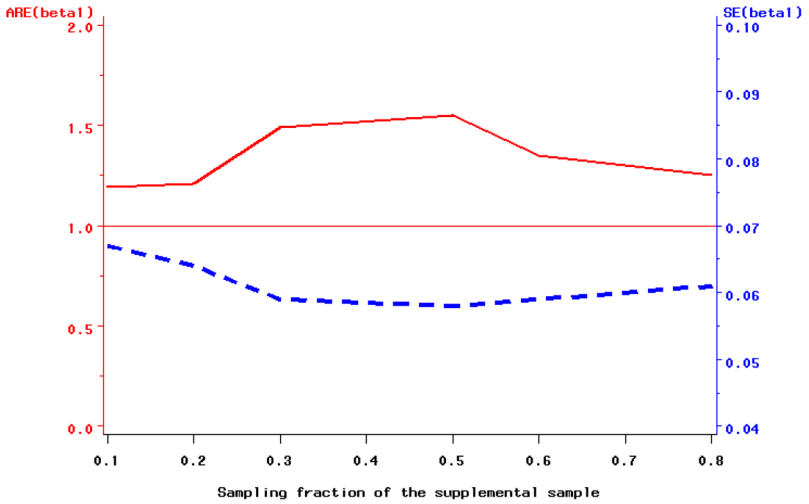

Figure 1.

Relative efficiency of $\hat{\boldsymbol{\theta}}_P$ to $\hat{\boldsymbol{\theta}}_S$ for $\hat{\beta}_1$ across the sampling fraction of the supplemental samples ($n_1/n + n_2/n$), under the bivariate model with $n = 200$ (*Multivariate-ODS*), $\alpha_1 = 0.5$, $\beta_1 = -0.5$, $\alpha_2 = -0.8$, $\beta_2 = \ln(2)$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0.5$, and $X_1 = X_2 \sim N(0, 1)$; U = 90% and L = 10%.

Stat Med. Author manuscript; available in PMC 2018 March 15.

Table 1

Simulation Results: Bivariate normal model with $n = 200$ (*Multivariate-ODS*), $\alpha_1 = 0.5$, $\beta_1 = -0.5$, $\alpha_2 = -0.8$, $\beta_2 = \ln(2)$, $\sigma_1 = \sigma_2 = 1$, and $X_1 = X_2 \sim N(0, 1)$. *

				$\hat{\beta}_1$				$\hat{\beta}_2$			
ρ	Cutpoints (U, L)	Design	Method	Bias	SSD	ESE	95% CI	Bias	SSD	ESE	95% CI
0.5	90%, 10%	$n_1 = n_2 = 5\%$	θ_R	−0.002	0.075	0.075	0.951	−0.004	0.075	0.075	0.953
			θ_C	−0.002	0.074	0.074	0.952	−0.004	0.075	0.074	0.956
			θ_S	0.002	0.073	0.071	0.946	0.000	0.070	0.071	0.961
			θ_P	−0.002	0.067	0.066	0.938	−0.004	0.067	0.067	0.956
		$n_1 = n_2 = 20\%$	θ_R	0.000	0.093	0.092	0.956	−0.003	0.093	0.092	0.947
			θ_C	0.000	0.089	0.089	0.951	−0.003	0.087	0.087	0.948
			θ_S	0.001	0.070	0.071	0.959	0.000	0.068	0.071	0.963
			θ_P	0.001	0.062	0.059	0.941	−0.002	0.061	0.061	0.948
	70%, 30%	$n_1 = n_2 = 5\%$	θ_R	0.001	0.076	0.075	0.952	−0.002	0.076	0.075	0.952
			θ_C	0.002	0.075	0.074	0.958	−0.002	0.075	0.074	0.949
			θ_S	−0.004	0.072	0.071	0.949	−0.001	0.073	0.071	0.947
			θ_P	0.002	0.070	0.069	0.951	−0.001	0.071	0.069	0.944
		$n_1 = n_2 = 20\%$	θ_R	−0.001	0.094	0.092	0.945	0.002	0.092	0.092	0.952
			θ_C	0.000	0.088	0.087	0.946	0.001	0.085	0.085	0.956
			θ_S	0.003	0.070	0.072	0.946	0.000	0.073	0.071	0.945
			θ_P	0.000	0.066	0.066	0.949	0.002	0.068	0.066	0.945
0.85	90%, 10%	$n_1 = n_2 = 5\%$	θ_R	−0.001	0.075	0.075	0.948	0.000	0.074	0.075	0.950
			θ_C	0.000	0.074	0.075	0.945	0.000	0.074	0.075	0.950
			θ_S	0.000	0.072	0.071	0.949	0.001	0.071	0.071	0.950
			θ_P	−0.002	0.066	0.067	0.952	−0.001	0.066	0.067	0.948
		$n_1 = n_2 = 20\%$	θ_R	0.003	0.097	0.093	0.943	0.001	0.100	0.093	0.944
			θ_C	0.002	0.095	0.090	0.944	0.001	0.096	0.089	0.941
			θ_S	−0.005	0.071	0.071	0.945	−0.003	0.071	0.071	0.954

				$\hat{\beta}_1$				$\hat{\beta}_2$			
ρ	Cutpoints (U, L)	Design	Method	Bias	SSD	ESE	95% CI	Bias	SSD	ESE	95% CI
			θ_P	0.000	0.059	0.058	0.954	0.001	0.059	0.059	0.944
	70%, 30%	$n_1 = n_2 = 5\%$	θ_R	−0.001	0.076	0.075	0.952	0.001	0.076	0.075	0.956
			θ_C	−0.001	0.076	0.075	0.953	0.001	0.075	0.074	0.953
			θ_S	−0.003	0.070	0.071	0.956	−0.003	0.068	0.071	0.959
			θ_P	0.000	0.068	0.069	0.956	0.002	0.068	0.069	0.958
		$n_1 = n_2 = 20\%$	θ_R	0.001	0.097	0.092	0.945	0.002	0.095	0.092	0.945
			θ_C	0.000	0.092	0.088	0.948	0.000	0.089	0.088	0.943
			θ_S	0.002	0.069	0.071	0.958	0.002	0.068	0.071	0.950
			θ_P	0.000	0.068	0.065	0.944	0.000	0.065	0.064	0.948

*θ_R denotes the maximum likelihood estimator from the SRS portion only; θ_C denotes the conditional likelihood estimator by maximizing the L_{GL1}, θ_S denotes the regression estimator from a simple random sample of the same size as the *Multivariate-ODS*; and θ_P is the proposed estimator. SSD denotes the sample standard deviation and ESE is the mean of the estimated standard errors.

Table 2

Simulation Results of Relative efficiencies, $\widehat{ARE}_S\,(=Var_{\hat{\theta}_S}/Var_{\hat{\theta}_P})$ and $\widehat{ARE}_R\,(=Var_{\hat{\theta}_R}/Var_{\hat{\theta}_P})$: Bivariate normal model with $\alpha_1 = 0.5$, $\beta_1 = -0.5$, $\alpha_2 = -0.8$, $\beta_2 = \ln(2)$, $\sigma_1 = \sigma_2 = 1$ and $X_1 = X_2 \sim N(0, 1)$.

Cutpoints			Design	$n = 200$				$n = 800$			
				$\widehat{ARE}_R$		$\widehat{ARE}_S$		$\widehat{ARE}_R$		$\widehat{ARE}_S$	
ρ	Upper	Lower	$n_1 = n_2$	$\widehat{ARE}_{\hat{\beta}_1}$	$\widehat{ARE}_{\hat{\beta}_2}$	$\widehat{ARE}_{\hat{\beta}_1}$	$\widehat{ARE}_{\hat{\beta}_2}$	$\widehat{ARE}_{\hat{\beta}_1}$	$\widehat{ARE}_{\hat{\beta}_2}$	$\widehat{ARE}_{\hat{\beta}_1}$	$\widehat{ARE}_{\hat{\beta}_2}$
0.5	90%	10%	5%	1.26	1.27	1.19	1.08	1.27	1.18	1.14	1.13
			10%	1.59	1.59	1.21	1.38	1.63	1.53	1.35	1.17
			20%	2.25	2.34	1.30	1.25	2.52	2.29	1.65	1.28
			25%	2.99	2.74	1.55	1.44	2.87	2.72	1.67	1.34
	70%	30%	5%	1.17	1.14	1.03	1.04	1.16	1.14	0.94	1.12
			10%	1.39	1.35	1.10	1.14	1.32	1.30	0.98	1.02
			20%	1.99	1.81	1.17	1.09	2.02	2.03	1.25	1.25
			25%	2.43	2.14	1.13	1.10	2.49	2.36	1.21	1.07
0.85	90%	10%	5%	1.27	1.29	1.18	1.17	1.33	1.27	1.12	1.07
			10%	1.72	1.68	1.23	1.22	1.61	1.56	1.21	1.15
			20%	2.71	2.85	1.48	1.46	2.51	2.46	1.55	1.55
			25%	3.55	3.48	1.71	1.69	3.15	3.15	1.58	1.50
	70%	30%	5%	1.24	1.24	1.06	1.00	1.18	1.18	1.01	1.02
			10%	1.37	1.38	1.09	1.08	1.55	1.49	1.19	1.10
			20%	2.04	2.09	1.03	1.07	2.17	2.08	1.38	1.35
			25%	2.51	2.71	1.33	1.36	2.54	2.47	1.28	1.23

Table 3

Analysis Results for CPP Data.

		θ_R (n_0 = 100)			θ_S (n = 200)			θ_P (n = 200)		
		$\hat{\beta}$	ESE($\hat{\beta}$)	95% CI	$\hat{\beta}$	ESE($\hat{\beta}$)	95% CI	$\hat{\beta}$	ESE($\hat{\beta}$)	95% CI
Left Ear	Int	1.730	0.543	(0.666, 2.795)	1.618	0.384	(0.865, 2.371)	1.651	0.334	(0.997, 2.305)
	PCB	0.011	0.037	(−0.063, 0.084)	0.008	0.033	(−0.058, 0.073)	0.015	0.027	(−0.037, 0.067)
	SEX	0.069	0.185	(−0.432, 0.294)	0.030	0.126	(−0.216, 0.276)	0.098	0.110	(−0.117, 0.313)
	RACE	0.701	0.225	(−1.142, 0.260)	−0.887	0.140	(−1.161, 0.612)	0.800	0.133	(−1.061, 0.540)
	AGE	0.017	0.014	(−0.011, 0.045)	0.012	0.011	(−0.009, 0.032)	0.003	0.009	(−0.014, 0.020)
	EDUC	0.014	0.043	(−0.099, 0.070)	0.014	0.032	(−0.049, 0.077)	0.007	0.027	(−0.060, 0.047)
	SEI	0.014	0.056	(−0.096, 0.125)	0.019	0.040	(−0.059, 0.097)	0.034	0.035	(−0.035, 0.104)
Right Ear	Int	1.804	0.570	(0.687, 2.922)	1.688	0.403	(0.897, 2.478)	1.840	0.342	(1.169, 2.511)
	PCB	0.009	0.039	(−0.086, 0.068)	0.050	0.035	(−0.118, 0.019)	0.014	0.027	(−0.067, 0.039)
	SEX	0.329	0.194	(−0.710, 0.052)	0.100	0.132	(−0.359, 0.158)	0.070	0.112	(−0.289, 0.150)
	RACE	0.304	0.236	(−0.767, 0.159)	0.852	0.147	(−1.141, 0.564)	0.459	0.138	(−0.730, 0.188)
	AGE	0.031	0.015	(0.001, 0.061)	0.006	0.011	(−0.016, 0.027)	0.007	0.009	(−0.010, 0.025)
	EDUC	0.027	0.045	(−0.116, 0.062)	0.026	0.034	(−0.040, 0.092)	0.011	0.028	(−0.066, 0.044)
	SEI	0.041	0.059	(−0.157, 0.075)	0.033	0.042	(−0.049, 0.115)	0.006	0.036	(−0.066, 0.077)

Table 4

Simulation Results: Bivariate normal model with $n = 200$ (*Multivariate-ODS*), $\alpha_1 = 0.5$, $\beta_1 = -0.5$, $\alpha_2 = -0.8$, $\beta_2 = \ln(2)$, $\sigma_1 = 0.8$, $\sigma_2 = 1.5$, and $X_1 = X_2 \sim N(0, 1)$.

				$\hat{\beta}_1$				$\hat{\beta}_2$			
ρ	**Cutpoints (U, L)**	**Design**	**Method**	**Bias**	**SSD**	**ESE**	**95% CI**	**Bias**	**SSD**	**ESE**	**95% CI**
0.5	90%, 10%	$n_1 = n_2 = 20\%$	θ_R	0.000	0.074	0.074	0.953	0.000	0.140	0.138	0.943
			θ_C	0.001	0.070	0.070	0.945	0.001	0.136	0.133	0.943
			θ_S	−0.001	0.057	0.057	0.953	−0.002	0.105	0.107	0.958
			θ_P	−0.001	0.050	0.048	0.942	−0.002	0.087	0.088	0.956