Web-Based Supplementary Materials for "Statistical Inference for a Two-Stage Outcome-Dependent Sampling Design with a Continuous Outcome" by Haibo Zhou, Rui Song, Yuanshan Wu, and Jing Qin

Web Appendix A. Likelihood Function without W

The likelihood function without auxiliary information can be derived similarly as follows

$$\begin{aligned} L_n(\text{no } W) &= \prod_{i=1}^{m_0} f(y_{i0}|x_{i0};\beta)f(x_{i0}) \times \prod_{i=m_0+1}^{n_0} \int_{\mathcal{X}} f(y_{i0}|x;\beta)f(x)dx \\ &\times \prod_{k=1}^{K}\prod_{i=1}^{m_k} f(y_{ik}|x_{ik};\beta)f(x_{ik})/\pi_k \times \prod_{k=1}^{K}\prod_{i=m_k+1}^{n_k} \int_{\mathcal{X}} f(y_{ik}|x;\beta)f(x)dx/\pi_k, \end{aligned}$$

where $\pi_k = \Pr(Y \in C_k) = \int (F(c_k|x;\beta) - F(c_{k-1}|x;\beta))f(x)dx$.

The estimator that maximizes the likelihood $L_n(\text{no } W)$ using the our proposed semiparametric empirical likelihood method is denoted by '$\beta_{P:\text{ no } W}$.'

Web Appendix B. Sample Size Calculations

The sample size calculation can be determined using the asymptotic normal properties of the proposed estimator. For example, one wants to test the hypothesis

$$H_0:\ \beta_1 = a \qquad vs. \qquad H_1:\ \beta_1 = b,$$

where a and b are prespecified constants. Suppose an estimator for β_1, denoted by $\hat{\beta}_1$, satisfies that $(\hat{\beta}_1 - \beta_1)$ converges in distribution to a normal $N(0, \sigma^2)$ and σ can be consistently estimated by $\hat{\sigma}$. To emphasize the dependence of estimators $\hat{\beta}_1$ and $\hat{\sigma}$ on sample size n, we further denote them by $\hat{\beta}_1(n)$ and $\hat{\sigma}(n)$, respectively. Hence, given the significant level $\alpha = 0.05$, the power can be calculated as

$$\begin{aligned} \text{Power} &= 1 - \Pr\left(-1.96 \le \frac{\hat{\beta}_1(n) - a}{\hat{\sigma}(n)} \le 1.96 \mid H_1\right) \\ &= 1 - \Phi(1.96 - \frac{b-a}{\hat{\sigma}(n)_{\beta_1=b}}) + \Phi(-1.96 - \frac{b-a}{\hat{\sigma}(n)_{\beta_1=b}}), \end{aligned}$$

where $\Phi(x)$ is cumulative distribution function of standard normal variable. We used the above formula to calibrate the sample size for a given power and significant level.

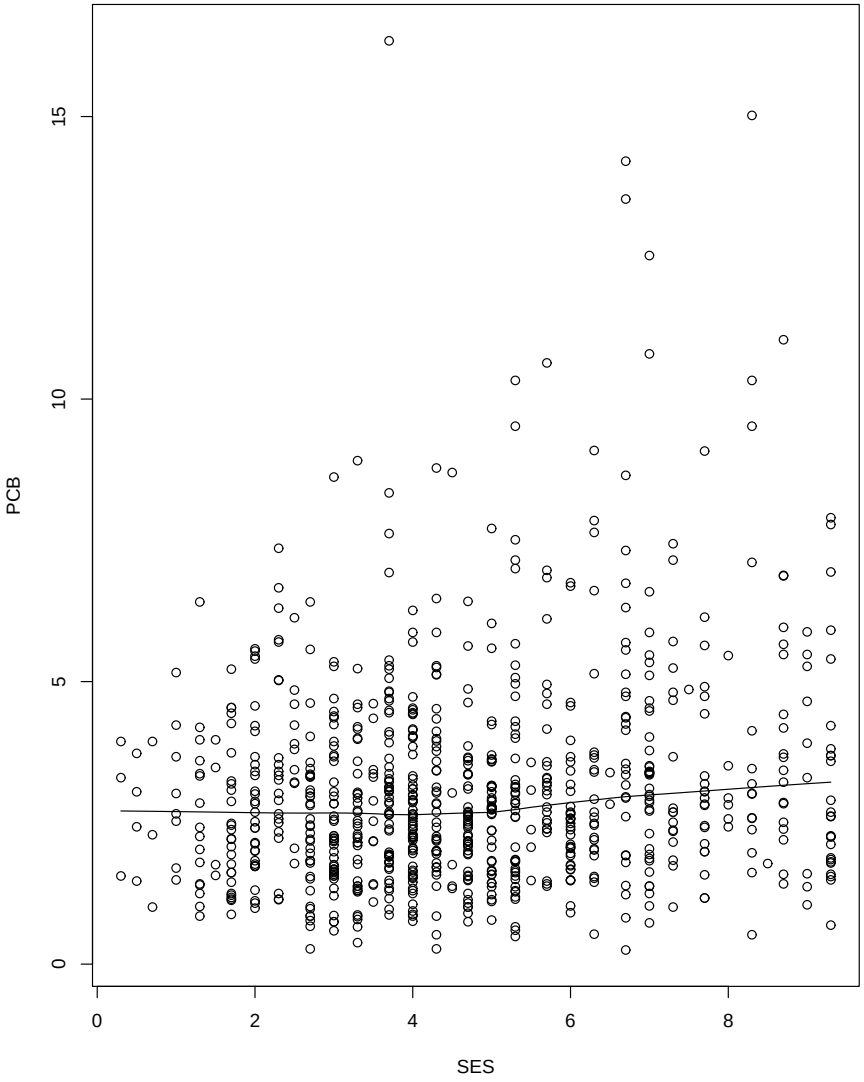

Figure 1: The plot of SES vs. PCB in CPP data set.

Web Appendix C. Investigations of the Relationship of PCB Given SES

To detect the relationship between the discretized SES and PCB, we plot SES versus PCB in Figure 1. The lowess curve indicates a linear association between SES and PCB, which is further verified by a linear model fit with the estimate of slope as 0.154 and a very significant $p < 0.0001$. Hence the relationship between PCB and the discretized SES is set to be $PCB = \xi_0 + \xi_1 W + \varepsilon_0$, and the error term ε_0 is normally distributed.