Inverse probability weighting

Motivation and assumptions

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ represent the finite population with N units and $\{(x_i, y_i), i \in \mathcal{S}_{\mathrm{A}}\}$ and $\{(x_i, d_i^B), i \in \mathcal{S}_{\mathrm{B}}\}$ be the datasets from non-probability and probability samples respectively. Following assumptions are required for this model:

1. The selection indicator R_i and the response variable y_i are independent given the set of covariates x_i.
2. All units have a nonzero propensity score, that is, $\pi_i^A > 0$ for all i.
3. The indicator variables R_i^A and R_j^A are independent for given x_i and x_j for $i \neq j$.

Maximum likelihood estimation

Suppose that propensity score can be modelled parametrically as $\mathbb{P}\left(R_i = 1 \mid x_i\right) = \pi(x_i, \theta_0)$. The maximum likelihood estimator is computed as $\hat{\pi}_i^A = \pi(x_i, \hat{\theta}_0)$, where $\hat{\theta}_0$ is the maximizer of the following log-likelihood function:

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{N} \left\{ R_i \log \pi_i^{\mathrm{A}} + (1 - R_i) \log\left(1 - \pi_i^{\mathrm{A}}\right) \right\} \\
&= \sum_{i \in \mathcal{S}_{\mathrm{A}}} \log \left\{ \frac{\pi\left(x_i, \theta\right)}{1 - \pi\left(x_i, \theta\right)} \right\} + \sum_{i=1}^{N} \log\left\{1 - \pi\left(x_i, \theta\right)\right\}
\end{aligned}
$$

Since we do not observe x_i for all units, Yilin Chen, Pengfei Li & Changbao Wu presented following log-likelihood function is subject to data integration basing on samples S_A and S_B. They proposed logistic regression model with $\pi(x_i, \theta) = \frac{\exp(x_i^\top \theta)}{\exp(x_i^\top \theta) + 1}$ in order to estimate θ. We expanded this approach on probit regression and complementary log-log model. For the sake of accuracy, let us recall that the probit and cloglog models are based on the assumption that model takes the form $\pi(x_i, \theta) = \Phi(x_i^\top \theta)$ and $\pi(x_i, \theta) = 1 - \exp(-\exp(x_i^\top \theta))$ respectively.