Web-based Supplementary Materials for "A general instrumental variable framework for regression analysis with outcome missing not at random" by Eric J Tchetgen Tchetgen and Kathleen Wirth

1 A semiparametric shared parameter model

The following result describes a relatively large class of possible data generating mechanisms for which (IV.3) is shown to hold. Suppose that conditional on $(\mathbf{X}, Z)$, the residual dependence between Y and R can be captured by a single continuous latent variable U, such that conditional on $(\mathbf{X}, U)$, the outcome Y is rendered independent of (R, Z) in the population, that is

$$E(Y|\mathbf{X}, U) = m(\mathbf{X}) + U; \tag{1}$$

$$\operatorname{logit} \Pr(R = 1|\mathbf{X}, Z, U) = \alpha(Z, \mathbf{X}) + \chi(\mathbf{X}) U; \tag{2}$$

$$U|R = 0, \mathbf{X}, Z \sim v(\mathbf{X}, Z) + \zeta, \text{ where } \zeta \text{ is independent of } (\mathbf{X}, Z). \tag{3}$$

This model is a semiparametric IV formulation of the classical shared parameter model of Wu and Carroll (1988) defined here for a single occasion. The model is semiparametric in the sense that the functions $m(\mathbf{X}), \chi(Z, \mathbf{X}), \alpha(\mathbf{X}), v(\mathbf{X}, Z)$ and the density of ζ are unrestricted. The main structural restriction is additivity of the random effect in the outcome regression, and in the propensity score model for the missing data process relative to Z on the logit scale, together with an assumption of homoscedasticity of ζ with respect to Z for nonresponders. The basic idea is to model the dependence between Y and R by incorporating an unobserved shared random effect parameter U in both the nonresponse model and the model generating the outcome Y, so that U represents an unobserved common cause of the outcome and the nonresponse mechanism. Under the shared parameter formulation, below we show that assumption (IV.3) is satisfied essentially provided that the IV does not interact with U in the nonresponse model. This assumption may be more likely to hold if $\mathbf{X}$ captures all possible effect modifiers of the association between Z and R, even if one has failed to capture all common causes of R and Y. This motivates the need to generally enrich $\mathbf{X}$, even if (IV.1) and (IV.2) hold unconditionally.

Result A.1: Assumptions $(1), (2)$ *and* (3) *imply that assumption (IV.3) holds, with* $\delta(\mathbf{X}) = E(Y|R=1,\mathbf{X},Z) - E(Y|R=0,\mathbf{X},Z) = \left.\frac{\partial \log E\{\exp(t\zeta)|\mathbf{X}\}}{\partial t}\right|_{t=\chi(\mathbf{X})}$

The proof is relegated to the Supplemental Materials.

2 Comparison to a Nonparametric Heckman Selection Model

Heckman's selection model is perhaps the most common strategy used in economics and other social sciences to address selection bias in regression analysis (Heckman, 1979). We adopt a nonparametric formulation of the model due to Das et al (2003) to ease a comparison to the proposed approach. This formulation assumes that the selection or missingness mechanism is generated under the latent variable threshold model:

(D.IV.1) $R = 1\{U < \zeta(\mathbf{X}, Z)\}$ where U is a latent random variable.

(D.IV.2) The model further supposes that, $Y = \mu(\mathbf{X}) + \varepsilon$ where ε is a separable residual error with the joint density of (ε, U) assumed to be independent of $(\mathbf{X}, Z)$ but otherwise unrestricted.

Then, assuming that the CDF of U, $G_u(\cdot)$ is one-to-one, for $V = G_u(U)$, Das et al (2003) establish that $E(\varepsilon|\mathbf{X}, Z, R=1) = E(\varepsilon|\mathbf{X}, Z, U < \zeta(\mathbf{X}, Z)) = \lambda(\pi(\mathbf{X}, Z))$ where $\pi(\mathbf{X}, Z) = \Pr(R = 1|\mathbf{X}, Z) = \Pr(U < \zeta(\mathbf{X}, Z)|\mathbf{X}, Z) = G_u(\zeta(\mathbf{X}, Z))$. Assuming that (ε, U) are joint Gaussian with

$Var(U) = 1$, gives $\lambda(\pi) = \sigma_{\varepsilon U}\ \phi\left(\Phi^{-1}(\pi)\right)/\pi$ where $\sigma_{\varepsilon U} = Cov(\varepsilon, U)$, $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal CDF and $\phi(\cdot)$ is the standard normal density. Assumptions (D.IV.1), (D.IV.2), the Gaussian assumption together with a linear specification for $\zeta(\mathbf{X}, Z)$ and $\mu(\mathbf{X})$ yield Heckman's (1979) standard correction for selection, which is completely identified from the observed data. However, in the larger nonparametric model defined by assumptions (D.IV.1) and (D.IV.2), so that $\zeta(\mathbf{X}, Z)$ and $\mu(\mathbf{X})$ remain unrestricted, Das et al (2003) established that $\mu(\mathbf{X})$ becomes nonparametrically identified up to an additive constant, provided assumption (D.IV.3) below also holds.

(D.IV.3) $\mu(\mathbf{X})$, $\lambda(\pi)$, and $\pi(\mathbf{X}, Z)$ are continuously differentiable with continuous distribution functions almost everywhere and with probability one, $\partial(\pi(\mathbf{X}, Z))/\partial Z \neq 0$.

Similar to (IV.2) assumption (D.IV.3) states that Z must predict R conditional on $\mathbf{X}$, although the latter is restricted to a continuous IV. Note that for the complete-case sample, under the nonparametric model given by assumptions (D.IV.1)-(D.IV.2), one can write

$$\begin{aligned} E(Y|R=1, \mathbf{X}, Z) &= \mu(\mathbf{X}) + \lambda(\pi(\mathbf{X}, Z)) \\ &= \mu(\mathbf{X}) + \delta^*(\mathbf{X}, Z)\{1 - \pi(\mathbf{X}, Z)\} \end{aligned}$$

which using equation (**??**), implies that the model restricts selection bias to be of the following form:

$$\delta^*(\mathbf{X}, Z) = \frac{\lambda(\pi(\mathbf{X}, Z))}{1 - \pi(\mathbf{X}, Z)}$$

Thus, we have learned that the nonparametric version of Heckman's selection model allows dependence of the selection bias function on both $\mathbf{X}$ and Z, but restricts such dependence to operate only through an unrestricted function of the propensity score. In contrast, in this paper, we have allowed under Assumption (IV.3), the selection bias function to be an unrestricted function of $\mathbf{X}$, however restricting it to not further depend on Z. Assumptions (IV.1)-(IV.3) give nonparametric identification of the function $\mu(\mathbf{X})$, while assumptions (D.IV.1)-(D.IV.3) can only identify $\mu(\mathbf{X})+C$ for an unknown constant C. This means that the intercept of the function $\mu(\mathbf{X})$ is not identified under the latter conditions, while it is under the former. The intercept may itself be of interest, in settings such as in the previous DHS example where the outcome level for each value of $\mathbf{X}$ is of primary scientific interest. The intercept will also be key to recover a valid estimate of the average outcome $E(Y) = E[\mu(\mathbf{X})]$. Interestingly, Newey (2009) also notes that, together with (D.IV.1)-(D.IV.3), further restricting $\mu(\mathbf{X})$ to be a linear function of $\mathbf{X}$, and assuming that $\pi(\mathbf{X}, Z)$ is a single index model still does not suffice to identify the intercept of $\mu(\mathbf{X})$ and thus to identify $E(Y)$. This further clarifies that identification of the intercept in the original Heckman model is principally derived from the joint Gaussian assumption of (ε, U), a parametric assumption which together with linearity assumptions, imposes strong restrictions on the observed data distribution, and thus, it should be of no surprise that, as reported in the literature, inferences about the intercept in this framework can be quite sensitive to the underlying identifying assumptions (Arabmazar & Schmidt 1981, Winship and Mare, 1992, Puhani, 2000).

Inference with the log link

In this section, we consider regression analysis with the log link function, and define the model of interest as $\mu(\mathbf{X}) = \log E(Y|\mathbf{X})$. We may proceed as with the identity link and first derive the multiplicative selection bias for the observed complete-case regression $E(Y|\mathbf{X}, R = 1)$,

$$
\begin{aligned}
\frac{E(Y|\mathbf{X},R=1)}{E(Y|\mathbf{X})} &= \frac{E(Y|\mathbf{X},R=1)}{E(Y|\mathbf{X},R=0)} \times \left\{ \sum_{r=1} \frac{E(Y|\mathbf{X},R=r)}{E(Y|\mathbf{X},R=0)} \Pr(R=r|\mathbf{X}) \right\}^{-1} \\
&= \widetilde{\nu}\left(\mathbf{X}\right) \left\{ \widetilde{\nu}\left(\mathbf{X}\right) \Pr(R=1|\mathbf{X}) + \Pr(R=0|\mathbf{X}) \right\}^{-1}
\end{aligned}
$$

where $\widetilde{\nu}\left(\mathbf{X}\right) = E(Y|\mathbf{X},R=1)/E(Y|\mathbf{X},R=0)$ encodes the degree of association between Y and R given $\mathbf{X}$ on the mean ratio scale, and quantifies the amount of selection bias. Naturally, as before, $E(Y|\mathbf{X},R=1) = E(Y|\mathbf{X})$ if and only if $\widetilde{\nu}\left(\mathbf{X}\right) = 1$ or $\Pr(R=1|\mathbf{X}) = 1$, that is if and only if there is no selection bias or no missing data. We say that Z is a valid IV for a log regression analysis with nonignorable missing outcome, if Z satisfies assumptions (IV.1) and (IV.2) and the following additional assumption,

(IV.3') Homogeneous multiplicative selection bias : $E(Y|R=1,\mathbf{X},Z)/E(Y|R=0,\mathbf{X},Z) = \nu\left(\mathbf{X}\right)$ does not depend on Z.

Similar to assumption (IV.3), the new assumption (IV.3') states that the IV essentially behaves as if it were randomized relative to the magnitude of selection bias which is now encoded on the multiplicative scale. We show below, that similar to Result 3, an analogous shared parameter model can be shown to imply (IV.3'), upon replacing the identity link function with the log link function, and strengthening the model for U such that $U|R,\mathbf{X},Z \sim v(R,\mathbf{X},Z) + \zeta$, where ζ is independent of $Z|R,\mathbf{X}$. Thus, assumption (IV.3') essentially states that there is no unobserved common correlate of R and the mean of Y (on the loglinear scale), which interacts with Z in predicting R (on the logit scale). The assumption is thus made more credible if one can account in $\mathbf{X}$ for all predictors of R that may also interact with Z even if one cannot with certainty rule out the presence of U.

Our identification result for the multiplicative scale is given next.

Result A.2: Under assumptions (IV.1)-(IV.3'), the regression function $\mu(\mathbf{X})$ *is nonparametrically identified from the observed data* $(\mathbf{X}, RY, R, Z)$, *and the complete-case regression curve* $m(\mathbf{X}, Z) = E(Y|Z, \mathbf{X}, R = 1)$ *can be expressed as a function of* $\mu(\mathbf{X}), \nu(\mathbf{X})$ *and* $\pi(\mathbf{X}, Z)$ *as followed:*

$$\log m(\mathbf{X}, Z) = \log \nu(\mathbf{X}) - \overline{\nu}(\mathbf{X}, Z) + \mu(\mathbf{X}) \tag{4}$$

$$\text{where } \overline{\nu}(\mathbf{X}, Z) = \log\{\nu(\mathbf{X})\pi(\mathbf{X}, Z) + 1 - \pi(\mathbf{X}, Z)\} \tag{5}$$

Result A.2. states that the regression curve $E(Y|\mathbf{X}) = \exp\{\mu(\mathbf{X})\}$ is identified from data $(RY, Z, \mathbf{X}, R)$ provided that Z is an IV satisfying assumptions (IV.1)-(IV.3'). Equation (4) gives an explicit representation of the complete-case regression $E(Y|Z, \mathbf{X}, R = 1)$ as a function of the regression of interest $\mu(\mathbf{X})$, the selection bias function $\nu(\mathbf{X})$ and the propensity score $\pi(\mathbf{X}, Z)$. Crucially, we note that $\overline{\nu}(\mathbf{X}, Z)$ in equation (4) is not a free parameter, but corresponds to a carefully crafted offset fully determined by the selection bias function and the missingness mechanism as displayed in equation (5).

Equation (4) suggests a simple strategy for estimating $\mu(\mathbf{X})$ in practice. To illustrate, suppose that Y is a count, and interest lies in the familiar log-linear model $\mu(\mathbf{X}, \psi) = (1, \mathbf{X}')\psi$. Further suppose that one specifies a similar log-linear model to encode selection bias $\log \nu(\mathbf{X}; \eta) = (1, \mathbf{X}')\eta$. Then, assuming that Y follows a Poisson distribution with mean computed using formula (4) under the above model,

$$m(\mathbf{X}, Z; \eta, \alpha, \psi) = \exp((1, \mathbf{X}')\eta - \overline{\nu}(\mathbf{X}, Z; \eta, \alpha) + (1, \mathbf{X}')\psi)$$

$$\text{where } \overline{\nu}(\mathbf{X}, Z; \alpha, \eta) = \log\{\exp[(1, \mathbf{X}')\eta]\pi(\mathbf{X}, Z; \alpha) + 1 - \pi(\mathbf{X}, Z; \alpha)\}$$

The maximum likelihood estimator of $\phi = (\psi, \alpha, \eta)$ maximizes equation (6) upon replacing f_1 with the Poisson density with mean given in the previous display. Maximum likelihood inference then proceeds as previously described. A two-stage estimation approach similar to the one proposed for the identity link can likewise be used for the log link and is easily inferred from the presentation.

Shared parameter model for the log link function: Suppose that $(\mathbf{X}, Z, Y, U)$ follow the model:

$$\log E(Y|\mathbf{X}, U) = m(\mathbf{X}) + U \tag{6}$$

$$\operatorname{logit}\Pr(R = 1|\mathbf{X}, Z, U) = \alpha(Z, \mathbf{X}) + \chi(\mathbf{X})U \tag{7}$$

$$U|R, \mathbf{X}, Z \sim v(R, \mathbf{X}, Z) + \zeta, \text{ where } \zeta \text{ is independent of } Z|\mathbf{X}, R \tag{8}$$

then,

$$E(Y|\mathbf{X}, R = 1, Z) / E(Y|\mathbf{X}, R = 0, Z) = \nu(\mathbf{X})$$

does not depend on Z.

Proof: Under the assumptions,

$$\begin{aligned}
&E(Y|\mathbf{X}, R = 1, Z) \\
&= \exp\{m(\mathbf{X})\} E[\exp\{U\} |\mathbf{X}, R = 1, Z] \\
&= \exp\{m(\mathbf{X})\} \exp\{E(U|\mathbf{X}, R = 1, Z)\} E[\exp(\zeta) |\mathbf{X}, R = 1, Z] \\
&= \exp\{m(\mathbf{X})\} \exp\left\{\frac{E(U\exp(\chi(\mathbf{X})U) |R = 0, \mathbf{X}, Z)}{E(\exp(\chi(\mathbf{X})U) \ |R = 0, \mathbf{X}, Z)}\right\} E[\exp(\zeta) |\mathbf{X}, R = 1] \\
&= \exp\{m(\mathbf{X})\} \exp\left\{v(0, \mathbf{X}, Z) + \left.\frac{\partial \log E[\exp(t\zeta) |\mathbf{X}, R = 0]}{\partial t}\right|_{t=\chi(\mathbf{X})}\right\} \\
&\times E[\exp(\zeta) |\mathbf{X}, R = 1]
\end{aligned}$$

therefore

$$\begin{aligned}
&\frac{E\left(Y|\mathbf{X},R=1,Z\right)}{E\left(Y|\mathbf{X},R=0,Z\right)} \\
&= \frac{\exp\left\{m\left(\mathbf{X}\right)\right\}\exp\left\{v(0,\mathbf{X},Z)\right\}\exp\left\{\left.\frac{\partial\log E[\exp(t\zeta)|\mathbf{X},R=0]}{\partial t}\right|_{t=\chi(\mathbf{X})}\right\}E\left[\exp\left(\zeta\right)|\mathbf{X},R=1\right]}{\exp\left\{m\left(\mathbf{X}\right)\right\}\exp\left\{v(0,\mathbf{X},Z)\right\}E\left[\exp\left(\zeta\right)|\mathbf{X},R=0,Z\right]} \\
&= \frac{\exp\left\{\left.\frac{\partial\log E[\exp(t\zeta)|\mathbf{X},R=0]}{\partial t}\right|_{t=\chi(\mathbf{X})}\right\}E\left[\exp\left(\zeta\right)|\mathbf{X},R=1\right]}{E\left[\exp\left(\zeta\right)|\mathbf{X},R=0,Z\right]}
\end{aligned}$$

which does not depend on Z, thus establishing the result.

Two Stage Estimation: *An alternative estimation strategy follows a two stage approach, whereby in a first stage one computes $\widehat{\alpha}_2$ by maximizing the log partial likelihood function $\sum_i \log f_2\left(R_i|Z_i,\mathbf{X}_i;\alpha\right)$, followed by a second stage, in which one uses $\pi\left(\mathbf{X},Z;\widehat{\alpha}\right)$ to estimate $m(X,Z;\theta)$ via complete case ordinary least square regression of Y on $\left(1,\mathbf{X}',\left(1,\mathbf{X}\right)\left(1-\pi\left(\mathbf{X},Z;\widehat{\alpha}\right)\right)\right)$. For inference under the two stage approach, we recommend the nonparametric bootstrap.*

Proof of Result A.1: Note that, similar to Tchetgen Tchetgen (2013)

$$\begin{aligned}
E(Y|R=1,\mathbf{X},Z) &= m\left(\mathbf{X}\right)+E(U|R=1,\mathbf{X},Z) \\
&= m\left(\mathbf{X}\right)+\frac{E(U\exp\left(\chi\left(\mathbf{X}\right)U\right)|R=0,\mathbf{X},Z)}{E(\exp\left(\chi\left(\mathbf{X}\right)U\right)\ |R=0,\mathbf{X},Z)} \\
&= m\left(\mathbf{X}\right)+\partial\log E(\exp\left(tU\right)|R=0,\mathbf{X},Z)/\partial t|_{t=\chi(\mathbf{X})} \\
&= m\left(\mathbf{X}\right)+\partial\log\left\{E\left(e^{t\zeta}|\mathbf{X},Z,R=0\right)\exp\left(tv(\mathbf{X},Z)\right)\right\}/\partial t|_{t=\chi(\mathbf{X})} \\
&= m\left(\mathbf{X}\right)+v(\mathbf{X},Z)+\left.\frac{\partial\log M^{**}(t;\mathbf{X})}{\partial t}\right|_{t=\chi(\mathbf{X})}
\end{aligned}$$

where $M^{**}(t;X) = E\left(e^{t\zeta}|\mathbf{X},Z,R=0\right)$ *by assumption. Therefore, it follows that*

$$\begin{aligned}
&E\left(Y|R=1,\mathbf{X},Z\right)-E\left(Y|R=0,\mathbf{X},Z\right)\\
&=m\left(\mathbf{X}\right)+v(\mathbf{X},Z)+\left.\frac{\partial\log M^{**}(t;\mathbf{X})}{\partial t}\right|_{t=\chi(\mathbf{X})}\\
&-m\left(\mathbf{X}\right)-v(\mathbf{X},Z)\\
&=\left.\frac{\partial\log M^{**}(t;\mathbf{X})}{\partial t}\right|_{t=\chi(\mathbf{X})}
\end{aligned}$$

proving the result. $\square$

Proof of Result 1: *The proof relies on the following decomposition:*

$$\begin{aligned}
&E\left(Y|Z,\mathbf{X},R\right)\\
&=E\left(Y|Z,\mathbf{X},R\right)-E\left(Y|Z,\mathbf{X},R=0\right)\\
&-\sum_{r=0}^{1}\left\{E\left(Y|Z,\mathbf{X},r\right)-E\left(Y|Z,\mathbf{X},R=0\right)\right\}\Pr\left(R=r|\mathbf{X},Z\right)\\
&+E\left(Y|Z,\mathbf{X}\right)
\end{aligned}$$

Thus, under assumptions (IV.1)-(IV.3), we obtain for $R=1$

$$m(\mathbf{X},Z)=\delta\left(\mathbf{X}\right)-\delta\left(\mathbf{X}\right)\pi\left(\mathbf{X},Z\right)+\mu\left(\mathbf{X}\right)$$

Next, since $\pi\left(\mathbf{X},Z\right)$ *is identified from the partial likelihood of* R *given* (X,Z)*, we may take it as known. Then, we obtain the identification result by noting that for all* $\pi\left(\mathbf{X},Z\right)$ *that satisfy (IV.2),* $m^{*}(X,Z)=m(X,Z)$ *if and only if* $\delta^{*}\left(\mathbf{X}\right)=\delta\left(\mathbf{X}\right)$ *and* $\mu^{*}\left(\mathbf{X}\right)=\mu\left(\mathbf{X}\right)$*, where* $m^{*}(X,Z)=$

$\delta^*(\mathbf{X}) - \delta^*(\mathbf{X})\,\pi(\mathbf{X}, Z) + \mu^*(\mathbf{X})$. $\square$

Proof of Result 2: *Note that the odds function*

$$ODDS\ (Z, \mathbf{X}, R = 1) = P\,(Y = 1|Z, \mathbf{X}, R = 1)\,/\,\Pr(Y = 0|Z, \mathbf{X}, R = 1)$$

can be decomposed nonparametrically as followed:

$$\begin{aligned} &ODDS\ (Z, \mathbf{X}, R = 1) \\ &= \frac{ODDS\ (Z, \mathbf{X}, R = 1)}{ODDS\ (Z, \mathbf{X}, R = 0)} \times \left\{ \sum_r \frac{ODDS\ (Z, \mathbf{X}, R = r)}{ODDS\ (Z, \mathbf{X}, R = 0)} \Pr(R = r|Z, \mathbf{X}, Y = 0) \right\}^{-1} \\ &\times ODDS\ (Z, \mathbf{X}) \end{aligned}$$

Then, under assumption (IV.3†), we have that

$$\log ODDS\ (Z, \mathbf{X}, R = 1) = \omega(\mathbf{X}) - \overline{\omega}(\mathbf{X}, Z) + \mu(\mathbf{X})$$

Identification of $\mu(\mathbf{X})$ and $\omega(\mathbf{X})$ *then follows by Corollary 1 and Example 2 of Sun et al (2016). Finally, we note that*

$$
\begin{aligned}
&\Pr(R=1|Z,\mathbf{X}) \\
&\quad = \Pr(R=1|Z,\mathbf{X},Y=0)\Pr(Y=0|\mathbf{X},Z)+\Pr(R=1|Z,\mathbf{X},Y=1)\Pr(Y=1|\mathbf{X},Z) \\
&\quad = \Pr(R=1|Z,\mathbf{X},Y=0)\Pr(Y=0|\mathbf{X})+\Pr(R=1|Z,\mathbf{X},Y=1)\Pr(Y=1|\mathbf{X}) \\
&\quad = \{1-t(\mathbf{X})\}\lambda(\mathbf{X},Z)+t(\mathbf{X})\lambda(\mathbf{X},Z)\exp\{\omega(\mathbf{X})\}/[\lambda(\mathbf{X},Z)\exp\{\omega(\mathbf{X})\}+\{1-\lambda(\mathbf{X},Z)\}]
\end{aligned}
$$

which implies that $\Pr(R=1|Z,X,Y=0)$ *is identified from the observed data likelihood of* R *given* (X,Z)*.*

Proof of Result A.2: *Note that the regression function* $E(Y|Z,\mathbf{X},R)$ *can be decomposed nonparametrically as followed:*

$$
\begin{aligned}
&E(Y|Z,\mathbf{X},R=1) \\
&= \frac{E(Y|Z,\mathbf{X},R=1)}{E(Y|Z,\mathbf{X},R=0)} \times \left\{\sum_r \frac{E(Y|Z,\mathbf{X},R=r)}{E(Y|Z,\mathbf{X},R=0)}\Pr(R=r|Z,\mathbf{X})\right\}^{-1} \\
&\times E(Y|Z,\mathbf{X})
\end{aligned}
$$

Then under our assumptions, we have that

$$
\log E(Y|Z,\mathbf{X},R=1) = \log\nu(\mathbf{X}) - \overline{\nu}(\mathbf{X},Z) + \mu(\mathbf{X})
$$

Finally, we obtain the identification result upon noting that $\Pr(R=1|Z,X)$ *is nonparametrically identified from the partial likeihood for the missingness mechanism, and thus* $\log E\left(Y|Z,\mathbf{X},R=1\right)=\log E^*\left(Y|Z,\mathbf{X},R=1\right)$ *if and only if*

$$\nu^*\left(\mathbf{X}\right)=\nu\left(\mathbf{X}\right)$$

$$\mu\left(\mathbf{X}\right)=\mu^*\left(\mathbf{X}\right)$$

where

$$\log E^*\left(Y|Z,\mathbf{X},R=1\right)=\log \nu^*\left(\mathbf{X}\right)-\overline{\nu}^*\left(\mathbf{X},Z\right)+\mu^*\left(\mathbf{X}\right)$$

$$\overline{\nu}^*\left(\mathbf{X},Z\right)=\log\left\{\left[\exp\left\{\nu^*\left(\mathbf{X}\right)\right\}\right]\pi\left(\mathbf{X},Z\right)+1-\pi\left(\mathbf{X},Z\right)\right\}.$$

References

[1] Arabmazar A. and Schmidt P. (1981), Further evidence on the robustness of the Tobit estimator to heteroscedasticity. J. of Econometrics 17: 253-258

[2] Das M, Newey W K, Vella F (2003). Nonparametric estimation of sample selection models. Review of Economic Studies 70, 33-58.

[3] Newey W & James, L., 2003. Instrumental Variable Estimation of Nonparametric Models, Econometrica, Econometric Society, vol. 71(5), pages 1565-1578.

[4] HeckmanJ, . J. (1979). Samples election bias as a specification error. Econometrica 47:153-61.

[5] Heckman, J. J. (1997), Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, Journal of Human Resources, 32, 441–462.

[6] Newey, W. K., Powell, J. L., Walker, J. R. (1990). Semiparametric estimation of selection models: some empirical results. Am. Econ. Rev. 80:324-28.

[7] Puhani PA.(2000) The Heckman correction for sample selection and its critique. J Econ Surv;14:53– 68.

[8] Sun B, Liu L, Miao W, Wirth K, Robins J, Tchetgen Tchetgen EJ. Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable. eprint arXiv:1607.03197.

[9] Tchetgen Tchetgen, Eric J. (2013), A General Regression Framework for a Secondary Outcome in Case-control Studies. Biostatistics. 15 (1): 117-128.

[10] Winship C. and Mare R. Models for sample selection bias (1992). Annu Rev Sociol;18:327–350.

[11] Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics, 44, 175–188.