Notes on High–Dimensional Models of Sample Selection

Shuowen Chen

November 15, 2019

1 Model

Many economics data are missing due to sample selection, in which the sampling is non-random. A popular method for sample selection estimation is Heckit (Heckman, 1979) as follows.

$$y = dy^*; \tag{1}$$
$$Prob(d = 1|Z) = \Phi(Z\gamma) \tag{2}$$
$$y^* = X\beta + u \tag{3}$$
$$\mathbb{E}[y|X, D = 1] = X\beta + \rho\sigma_u\lambda(Z\gamma) \tag{4}$$

Heckit assumes that missing data are due to a binary-choice selection equation. Once a consistent estimate of the correction term is obtained, economists add the term into the outcome regression with observed dependent variable only and obtain the estimate of parameter in interest. Heckit relies on three conditions. Identification requires exclusion restrictions, namely that the selection equation has at least one covariate that is not in the main regression equation. Errors in selection and main regression equations are correlated, usually assumed to be jointly normally distributed. Lastly, economists need to pre-specify covariates and functional forms.

I explore a sample selection estimation method in a high-dimensional context without pre-specification of control variables. The following is the model:

$$y_i = d_i y_i^*; \tag{5}$$
$$y_i^* = X_i\theta + g_0(W_i) + \nu_i, \quad \mathbb{E}[\nu_i|X_i, W_i] = 0; \tag{6}$$
$$d_i = \mathbb{1}\{m_0(Z_i, X_i, W_i) + \epsilon_i \geq 0\}, \quad (i = 1, ..., n), \quad \mathbb{E}[\epsilon_i|X_i, W_i, Z_i] = 0 \tag{7}$$
$$X_i = p_0(W_i) + \eta_i, \mathbb{E}[\eta_i|W_i] = 0; \tag{8}$$
$$\nu_i = \frac{\rho}{\sqrt{1-\rho^2}}\Phi^{-1}\Lambda(\epsilon_i) + \zeta_i, \quad \zeta_i \sim \mathcal{N}(0, 1), \quad \rho := corr(\Phi^{-1}\Lambda(\epsilon), \nu). \tag{9}$$

1. Equation (1): d_i is the binary selection indicator and y_i is the sample observation. y_i^* is the unobserved true data.

2. Equation (2): DGP of y_i^*. The economic question in interest is how X affects y^*, and the goal is to conduct consistent estimate and valid inference on θ. W denotes an $n \times p$ control variable matrix in which $p > n$. $g_0(W_i)$ is an unknown function. ν is normally distributed.

3. Equation (3): Selection equation. Z_i, the variable in selection but not in y_i^*, is to satisfy the exclusion restriction for identification[1]. $m_0(Z_i, X_i, W_i)$ is an unknown function. ϵ_i is logistically distributed.

4. Equation (4): Statistical relationship between X and controls as a guard against omitted variable bias. Crucial for tackling post-selection inference problem by Leeb and Pöscher (2005).

5. Equation (5): Imposes a relation between ϵ_i and ν_i, without which missing data is not a concern, implies an analytical form of the sample correction term κ.

Key Assumption: approximate sparsity, namely that m_0, g_0 and p_0 can be well approximated by a few covariates whose identities are ex-ante unknown. In practice I impose a linear structure on the three functions and use LASSO to select important controls.

Assumption 1 (Approximate Sparsity with Exclusion Restriction)**.** *Each of z_i and y_i is well-approximated by a function of $s \geq 1$ covariate terms, whose coefficients γ and β depend on n and P. and the approximation error is no larger than $\sqrt{s/n}$ of the oracle estimator error:*

$$\begin{aligned} m_0(Z_i, X_i, W_i) &= Z_i\alpha + X_i\delta + W_i'\gamma + r_{mi} + \epsilon_i, \quad ||\gamma||_0 \leq s, \quad \sqrt{\{\mathbb{E}(r_{zi}^2)\}} \leq C\sqrt{s/n} \\ g_0(W_i) &= W_i'\beta + r_{gi}, \quad ||\beta||_0 \leq s, \quad \sqrt{\{\mathbb{E}(r_{gi}^2)\}} \leq C\sqrt{s/n}; \\ p_0(W_i) &= W_i'\vartheta + r_{pi}, \quad ||\vartheta||_0 \leq s, \quad \sqrt{\{\mathbb{E}(r_{xi}^2)\}} \leq C\sqrt{s/n} \end{aligned}$$

The sparsity index s obeys $s^2 \log^2(\max\{p, n\})/n \to 0$.

2 Equations for Estimation

In the fist step, we have

$$d_i = \mathbb{1}\{Z_i\alpha + X_i\delta + W_i'\gamma + r_{mi} + \epsilon_i \geq 0\}.$$

[1]Without this variable the identification of θ will depend on distribution assumption on errors.

Run LASSO-Logit on d against X, imposing zero penalty on Z; calculate $\hat{\kappa}$.

In the second step, we have

$$\mathbb{E}(y_i|W_i, X_i, Z_i, d_i = 1) = X_i\theta + g_0(W_i) + \omega\kappa,$$
$$X_i = p_0(W_i) + \eta_i.$$

This looks very similar to double selection model in Belloni, Chernozhukov and Hansen (2014). **Double Selection Algorithm:** Regress y on W using LASSO. Denote covariates with non-zero coefficients as $\hat{I}_1$. Regress X on W using LASSO. Denote covariates with non-zero coefficients as $\hat{I}_2$. Denote $\hat{I} = \hat{I}_1 \cup \hat{I}_2$. In practice we only have $\widehat{\kappa}$.

In the third step, use a **plug-in** approach. Now that we have $\hat{\kappa}$ and $\hat{I}$, one thought to obtain $\hat{\theta}$ is to plug the selected control $\hat{I}$ and estimated correction term $\hat{\kappa}$ into the main regression, and get

$$\hat{\theta} = \arg\min_{\theta,\beta,\omega} \sum_{i=1}^{n} d_i(y_i - X_{1i}\theta - \hat{I}\beta - \hat{\kappa}\omega)^2.$$

3 Current Results:

If there were no $\hat{\kappa}$ this estimator is essentially the post-LASSO solution in Belloni, Chernozhukov & Hansen (2014)[2]. However, the presence of $\hat{\kappa}$brings in two complications. Firstly, $\hat{\kappa}$ is a generated regressor, so approximation error in the LASSO-Logit regression will carry into the main regression. What's more, my model doesn't take a stand whether X enters the selection equation in an economically significant way. It turns out the coefficients of X in the selection equation greatly affects the performance of the plug-in estimate.

4 Analytical Form of Correction Terms

By writing the expectation conditional on observed y, we have the following:

$$\mathbb{E}(y_i|W_i, X_i, Z_i, d_i = 1) = X_i\theta + g_0(W_i) + \omega\mathbb{E}(\nu_i|W_i, X_i, Z_i, \epsilon_i \geq -m_0(Z_i, X_i, W_i)),$$

where the last term on the right hand side is the sample selection correction term. Denote it as κ. Due to the assumptions on ν and ϵ, the analytical forms of κ and its coefficients ω are

[2]In fact, the paper allows researchers to add a small amount of covariates that are not selected by LASSO in the regression. The authors call them amelioration covariates, and show that their presence doesn't affect consistency and asymptotic distributions as long as the cardinality is smaller than that of $\hat{I}$

as follows:

$$\kappa := \mathbb{E}(\nu_i | \epsilon_i \geq -m_0(Z_i, X_i, W_i)) = \frac{\int_{-\infty}^{J(m_0(Z_i,X_i,W_i))} J(\epsilon) f(J(\epsilon)) dJ(\epsilon)}{\Lambda(m_0(Z_i, X_i, W_i))}$$

$$\omega = \frac{\rho}{\sqrt{1-\rho^2}}$$

where $J = \Phi^{-1}\Lambda$ and $f(J(\epsilon))$ denotes the PDF of $J(\epsilon)$.