Relaxing Conditional Independence in an Endogenous Binary Response Model

Alyssa Carlson[*]

March 13, 2020

Abstract

For binary response models, the literature primarily addresses endogeneity by a control function approach with conditional independence (CF-CI). However, as the literature also notes, CF-CI implies conditions like homoskedasticity (of the latent error with respect to the instruments) that fail in most empirical settings. I propose an alternative approach that allows for heteroskedasticity, achieving identification with a conditional mean restriction. These identification results apply to a latent Gaussian error term with flexibly parameterized heteroskedasticity. I propose a two step conditional maximum likelihood estimator and derive its asymptotic distribution. Simulations show the benefit of the new estimator when CF-CI fails and is observed to be fairly robust against distributional misspecification. As an empirical illustration, I apply the new estimator to a model of married women's labor force participation.

1 Introduction

This paper considers a binary response model with an endogenous regressor, where the binary response can be modeled from a latent variable exceeding a threshold. For such models, a series of papers (Smith and Blundell (1986), Rivers and Vuong (1988), Blundell and Powell (2004), and Rothe (2009)) have proposed using a control approach. To gain identification, these papers essentially impose an exclusion restriction such that the conditional distribution

[*] Department of Economics, University of Missouri. Email: carlsonah@missouri.edu

of the latent error is independent of the instruments. These control function conditional independence assumptions (CF-CI) are unlikely to hold in an empirical setting.

To address this concern, I propose an alternative framework that does not rely on CF-CI for identification. This generalization has been explored in other settings, such as the case of endogenous random coefficients for a linear model in Wooldridge (2005) and demand estimation where the unobserved product characteristics do not enter the utility function additively in Gandhi et al. (2011). More generally, Kim and Petrin (2011) set up a framework for the "general control function," permitting the unobserved error to be a function of the instruments, in the case of additively separable triangular equation models.

In the Non-Parametric Control Function (NP-CF) literature (Newey et al. (1999), Pinkse (2000), Su and Ullah (2008), and Florens et al. (2008)), a control function assumption - usually exclusion of the instrument in the conditional mean of the error - was used for identification. This meant that the Non-Parametric Instrumental Variables (NP-IV) approach as in Ai and Chen (2003), Newey and Powell (2003), Hall and Horowitz (2005), and Newey (2013), appeared to be a superior approach that only requires a generally weaker Conditional Mean Restriction (CMR). Kim and Petrin (2011) shows that a control function approach is still valid under the CMR as long as their general control function is specified.

Like Kim and Petrin (2011), I use CMR for identification, which can hold even if CF-CI fails. I provide identification results when the latent error is Gaussian with flexibly parametrized heteroskedasticity. This leads to an easily implementable estimator that I show to be consistent and asymptotically normal in settings where no other existing estimator can produce both consistent parameter estimates and interpretation through the Average Structural Function (as defined in Blundell and Powell (2003)).

There are alternative estimators that do not require the CF-CI assumption in estimating endogenous binary response models. Lewbel (2000) and Dong and Lewbel (2015) require a "special regressor" that has large support and is fully independent of the latent error. Hong and Tamer (2003) and Krief (2014) extend the maximum score and smoothed maximum score methods of Manski (1985) and Horowitz (1992) to estimate the structural parameters. But for identification they would still require conditional median of the latent error to be independent of the instruments. Although this is weaker than independence from the entire conditional distribution, this assumption still prohibits a large range of likely data generating cases that the propose approach can address. Moreover the maximum score approaches do not recover the latent distribution and therefore cannot provide predicted outcomes or partial effects, statistical objects needed for most applied analysis.

Section 2 introduces a motivating example that illustrates a simple case in which CF-CI fails to hold while identification still seems attainable. Section 3 presents the model set-up while Section 4 shows identification in the considered setting. Section 5 explains implementation of the proposed estimator and provides asymptotic properties. All proofs of theorems are provided in the appendix. Because coefficients in a latent threshold binary response model bear little interpretative value, Section 6 discusses the derivation of the Average Structural Function, a statistical object useful for economic analysis in cases of endogeneity. CF-CI plays a role not only consistent parameter estimation but also in constructing the Average Structural Function, so imposing independence when it does not truly hold will have a doubly detrimental effect. A comprehensive simulation study comparing the proposed approach to other estimators in the literatures is presented in Section 7. In Section 8, I examine the effect of non-wage income on a married woman's probability of labor force participation. Using 1991 CPS data, I find both the semi-parametric estimator of Rothe (2009) and the proposed approach produce similar results that are interpretively different from the more restrictive approaches.

2 Motivating Example

Consider the latent variable triangular system where y_{1i} is a binary response variable, $\mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$ is a $1 \times (k_1 + k_2)$ vector of non-endogenous[1] included and excluded instruments, y_{2i} is a single continuous endogenous regressor, and $\mathbf{x}_i$ is a $1 \times k$ vector of regressors where each element is a function of $(\mathbf{z}_{1i}, y_{2i})$ and includes a constant:

$$
\begin{aligned}
y_{1i} &= \begin{cases} 1 & y_{1i}^* \geq 0 \\ 0 & y_{1i}^* < 0 \end{cases} \\
y_{1i}^* &= \mathbf{x}_i \beta_o - u_{1i} \\
y_{2i} &= m_o(\mathbf{z}_i, \pi_o) + v_{2i}.
\end{aligned} \tag{2.1}
$$

The last line specifies a first stage that is not a structural model, but merely a conditional mean decomposition of y_{2i}.

In this framework, Smith and Blundell (1986), Rivers and Vuong (1988), Blundell and Powell (2004), and Rothe (2009) developed estimators to address endogeneity using a control function approach. The control function approach supposes that there is a particular function (or variable) that when included as an additional covariate, is able to control for the endogeneity of the other regressors. For example, Rivers and Vuong (1988) (hereafter RV) show that if u_{1i} and v_{2i} are bivariate normal and independent of the instruments z_i, then

$$
u_{1i} \mid v_{2i}, \mathbf{z}_i \sim \mathrm{N}\left(\rho \frac{\sigma_1}{\sigma_2} v_{2i}, (1-\rho^2)\sigma_1^2\right) \tag{2.2}
$$

where σ_1 and σ_2 are the standard deviations of u_{1i} and v_{2i} respectively and ρ is their correlation. The RV control function approach (commonly called ivprobit) is a conditional maximum likelihood estimator where v_{2i} is estimated in the first stage and included as an additional regressor in the second stage.

[1] Non-endogenous is used rather than exogenous to emphasize that $\mathbf{z}_i$ will not be fully independent of the latent error.

To relax the distributional assumptions in RV, Blundell and Powell (2004) (hereafter BP) and Rothe (2009) propose alternative semi-parametric estimators. They relax RV's independence assumption of $(u_{1i}, v_{2i}) \perp \mathbf{z}_i$ to a conditional independence assumption. The CF-CI assumption of BP and Rothe (2009) requires the the instruments $\mathbf{z}_i$ to be independent of the latent error u_{1i} after conditioning on the reduced form error v_{2i}:

$$u_{1i} \mid v_{2i}, \mathbf{z}_i \sim u_{1i} \mid v_{2i}. \tag{2.3}$$

This means that any source of endogeneity must be fully captured through the control variate v_{2i}, or in terms of an exclusion restriction, the conditional CDF $F_{u_1|v_2,\mathbf{z}}(u_{1i} \mid v_{2i}, \mathbf{z}_i)$ is only a function of v_{2i} (the instruments $\mathbf{z}_i$ are excluded). This exclusion restriction must also hold for all moments which, as will be shown shortly, may be hard to justify.

As a slight relaxation of CF-CI, Rothe (2009) also proposes a Linear Index (CF-LI) sufficiency assumption that, after conditioning on the first stage error and the linear index $\mathbf{x}_i\beta_o$, the latent error is independent of the instruments:

$$u_{1i} \mid v_{2i}, \mathbf{x}_i\beta_o, \mathbf{z}_i \sim u_{1i} \mid v_{2i}, \mathbf{x}_i\beta_o. \tag{2.4}$$

Now the instruments can be a part of the conditional distribution but only through the linear index. The linear index restricts the relative direction and magnitudes of the regressors in the conditional distribution. So, although it allows for a more relaxed relationship between the instruments and the latent error, it is hard to justify in a general setting.

However both CF-CI and CF-LI are too stringent in many empirical contexts. To give a motivating parametric example, consider a slight variation of the RV setup where u_{1i} and v_{2i} are still bivariate normal but are allowed to be heteroskedastic in the instruments; i.e.,

$$\begin{pmatrix} u_{1i} \\ v_{2i} \end{pmatrix} \Bigg| \, \mathbf{z}_i \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} [\sigma_1(\mathbf{z}_i)]^2 & \rho(\mathbf{z}_i)\sigma_1(\mathbf{z}_i)\sigma_2(\mathbf{z}_i) \\ \rho(\mathbf{z}_i)\sigma_1(\mathbf{z}_i)\sigma_2(\mathbf{z}_i) & [\sigma_2(\mathbf{z}_i)]^2 \end{pmatrix} \right). \tag{2.5}$$

Heteroskedasticity is commonly found in empirical data, whether it is caused by variability in the latent error over the regressors, or by heterogeneity in the slopes as in a random coefficients setting.[2] Even in linear regression, heteroskedasticity has been accepted as endemic in empirical settings, and heteroskedasticity-robust inference is always employed.

By the properties of the bivariate normal distribution,

$$u_{1i} \mid v_{2i}, \mathbf{z}_i \sim \mathrm{N}\left(\rho(\mathbf{z}_i)\frac{\sigma_1(\mathbf{z}_i)}{\sigma_2(\mathbf{z}_i)}v_{2i}, (1-[\rho(\mathbf{z}_i)]^2)[\sigma_1(\mathbf{z}_i)]^2\right). \tag{2.6}$$

This is a fairly small variation to the framework considered in RV, but ignoring heteroskedasticity in the context of binary response models produces inconsistent parameter estimates. Similarly, applying the ivprobit estimator from RV also produces inconsistent parameter estimates.

Now compare the distribution in equation (2.6) to the CF-CI and CF-LI assumptions for the semi-parametric estimators. CF-CI clearly does not hold since the exclusion restriction does not hold: both the conditional mean and conditional variance depend on the instruments. For example, it is commonly noted that the control function methods are only valid for continuous and unbounded endogenous variables. This is because there would be inherent heteroskedasticity in the first stage which will violate CF-CI. For the CF-LI assumption to hold, the heteroskedastic functions $\sigma_1(\cdot)$, $\sigma_2(\cdot)$, and $\rho(\cdot)$ could only be functions of the linear index $\mathbf{x}_i\beta_o$. This is quite restrictive and would not generally hold. This means the semi-parametric estimators from Rothe (2009) and BP are not valid, even in this simple parametric setting. The proposed estimator is motivated from this simple example by incorporating the instruments into the conditional distribution.

[2] This is similar to the setup in Kasy (2011) where he provides a counter-example to the control function approach proposed by Imbens and Newey (2009). Imbens and Newey (2009) propose using the control variable $V_i = F_{y_1|z}(y_{1i}, z_i)$ which would satisfy CF-CI when the heterogeneity is only one-dimensional, as pointed out in Kasy (2011). In his example, a linear random coefficient model is used to show the failure of CF-CI using the Imbens and Newey control variable. Note that the random coefficient model can be rewritten as a linear model with heteroskedasticity as suggested here.

3 Model Set Up

Return to the set up described in equation (2.1). The distributional assumptions for u_{1i} and v_{2i} determine the consistent estimation procedure. Although most of the assumptions in the literature are based on a specification of the joint distribution of u_{1i} and v_{2i} (e.g., RV and Petrin and Train (2010)), the following assumption specifies the conditional distribution.[3]

Assumption 3.1. *Consider the setup in equation (2.1), where $\{y_{1i}, \boldsymbol{z}_i, y_{2i}\}_{i=1}^{n}$, is iid. In the first stage, the true conditional mean is*

$$\mathrm{E}(y_{2i} \mid \boldsymbol{z}_i) = m(\boldsymbol{z}_i, \pi_o)$$

and the control variate is defined as $v_{2i} = y_{2i} - m(\boldsymbol{z}_i, \pi_o)$. The unobserved latent error has the following conditional distribution

$$u_{1i} \mid \boldsymbol{z}_i, v_{2i}, y_{2i} = u_{1i} \mid \boldsymbol{z}_i, v_{2i} \sim \mathrm{N}\left(\boldsymbol{h}(v_{2i}, \boldsymbol{z}_i)\gamma_o, \exp(2 \times \boldsymbol{g}(y_{2i}, \boldsymbol{z}_i)\delta_o)\right)$$

where $\boldsymbol{z}_i = (\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})$ and $\boldsymbol{h}(v_{2i}, \boldsymbol{z}_i)$ and $\boldsymbol{g}(y_{2i}, \boldsymbol{z}_i)$ are known vectors while $m(\boldsymbol{z}_i, \pi)$ is an known function up to a vector of unknown parameters π.

The first part of the assumption breaks up the endogenous variable into its conditional mean and what will be referred to as the control variate v_{2i}. By construction, the control variate is mean independent of the instruments. To make this assumption reasonable, the conditional mean is allowed to be non-linear in its specification. This is going to allow for the case of discrete and bounded endogenous variables. This is an important distinction from the previous control literature whose methods could only be applied to continuous endogenous variables.

The second part of Assumption 3.1 specifies a conditional distribution that allows for the violation of CF-CI. Both the conditional mean and the conditional variance are functions of the instruments, so the exclusion restriction implied by CF-CI is violated. Under this

[3] The normality assumption could be easily generalized to any known distribution with CDF $G(\cdot)$ which allows for a Logit specification.

assumption, the conditional mean of y_{1i} is:

$$\mathrm{E}(y_{1i} \mid \mathbf{z}_i, y_{2i}, v_{2i}) = \mathrm{E}(y_{1i} \mid \mathbf{z}_i, v_{2i}) = \Phi\left(\frac{\mathbf{x}_i\beta_o + \boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i))\delta_o}\right)$$

Note that there is a one-to-one mapping between y_{2i} and v_{2i} given the instruments $\mathbf{z}_i$. This implies the mean is preserved regardless of which term is included in the conditioning argument. This result should be unsurprising as the conditional mean appears to be a heteroskedastic probit model that adjusts for endogeneity using the control function approach, both of which have been discussed extensively in the literature. But in this case, the control function ($\boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o$) is a function of both the control variate v_{2i} and the instruments $\mathbf{z}_i$. This was first introduced in Wooldridge (2005) where he suggests using the following control function,

$$\boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o = v_{2i}\gamma_{1o} + v_{2i}\mathbf{z}_i\gamma_{2o}$$

in a linear regression with random coefficients.[4] Gandhi et al. (2011) adopted a similar generalization for demand estimation and Kim and Petrin (2011) provides a general control function framework for the case of non-linear but additively separable triangular equation models. As in Kim and Petrin (2011), this generalization will be referred to as the "general control function," as opposed to a more traditional control function that upholds the exclusion restriction (not a function of the instruments) as in Rivers and Vuong (1988) and Petrin and Train (2010).

This assumption does not take a stand on the true data generating process of the endogenous variable. In the more general setting of non-separable triangular equation models, Imbens and Newey (2009) consider the case of a structural first stage

$$y_{2i} = d(\mathbf{z}_i, \eta_i) \tag{3.1}$$

[4] He also discusses in this paper the implementation of the control function approach in a binary response setting such as probit. But in that example, he does not propose interaction with the instruments and instead only suggests including higher order moments of the reduced form error. So his analysis stops short of what is proposed in this paper.

where z_i are the instruments, η_i is unobserved heterogeneity independent of the instruments, and $d(\cdot,\cdot)$ is the unknown and true data generating process in the first stage. In this setting they suggest using the conditional CDF, $e_{2i} = F_{y_2|\mathbf{z}}(y_{2i}, \mathbf{z}_i)$, as the control variable. They show that their proposed control variable satisfies CF-CI and so their control function method[5] recovers the parameters of the model. But throughout this entire process they require the instruments to be completely independent of any unobserved heterogeneity. In contrast, Assumption 3.1 does not require full independence between the control variable and the instruments. The population residual $v_{2i} = y_{2i} - \mathrm{E}(y_{2i} \mid \mathbf{z}_i)$ is used as a control variable with full knowledge that it does not satisfy CF-CI. Then to make up for the relaxation of CF-CI, the relationship between the unobserved latent error u_{1i}, the control variable v_{2i}, and the instruments $\mathbf{z}_i$ is flexibly modeled using a general control function.

Finally, Assumption 3.1 makes a linear-in-parameters and known distribution assumption.[6] These specifications are flexible but simple and are useful in showing identification and deriving asymptotic properties but general enough for most applications. With the linear-in-parameters and known distribution assumption, the proposed approach will be easily implementable with existing commands in a variety of statistical packages.

The distribution assumption is particularly pertinent in contrast to estimators from BP and Rothe (2009) that make no distributional assumptions. In simulation, this paper finds that the consequence of distributional misspecification is not as severe on Average Structural Function (ASF) estimates as presuming CF-CI when it does not hold. Intuition for this result can be attributed to a recent paper, Khan (2013), which shows a heteroskedastic probit model sufficient in producing consistent parameter estimates for a range of distributional

[5] They also require that the non-separable first stage needs to be monotonic in the unobserved heterogeneity and for there to only be a single source of unobserved heterogeneity in the first stage. A major critique to the approach of Imbens and Newey (2009) is the caveat to their framework brought up in Kasy (2011) noting their method only allows for one source of heterogeneity (independent of the instruments) in the first stage. This would prohibit the simple example of random coefficients in the first stage: $y_{2i} = \eta_{1i} + \eta_{2i}\mathbf{z}_i$. The approach in this paper allows for this possibility since the above equation can be rewritten in terms of a linear conditional mean with heteroskedasticity in the first stage error.

[6] The distributional assumption can be altered to accommodate a logistic distribution as commonly used for Logit.

misspecification in the latent error. This stems from an observational equivalence result concerning binary response models: a heteroskedastic probit model with a non-parametric heteroskedasticity function is observationally equivalent to a "distribution free" model with only a conditional median restriction. Consequently, introducing heteroskedasticity not only allows for a violation of CF-CI, but also provides a layer of protection against distributional misspecification.

Khan (2013) proves the observational equivalence result for the case of all exogenous regressors. An extension to the case of an endogenous regressors is included in Appendix A. It is left to future research to establish asymptotic properties of this extension with a fully non-parametrically specified heteroskedastic function and control function.

4 Identification

This section will show identification under the CMR and then discuss how this identifying restriction relates to the other control functions assumptions in the literature. An advantage of the CMR compared to the CF-CI is the interpretability of the assumption. While the CF-CI is useful to have in showing identification, it is difficult to understand in what cases the CF-CI should hold. In contrast, the CMR reinforces a more common exogeneity assumption: mean independence. For applied researchers, this exogeneity assumption is already well understood and utilized therefore valid instruments will be more easily attainable.

Recently, there has been growing interest in identification for the control function approach in non-parametric non-separable triangular simultaneous equation models.[7] However, the discussion usually starts with independence assumptions between the instruments and the unobservables. Then one searches for a control function that will satisfy strong identification assumptions such as CF-CI in BP or CF-CI and monotonicity in Imbens and Newey (2009). In this setting, the relationship between the instruments and the latent error is

[7] Imbens and Newey (2009), Kasy (2011), Hahn and Ridder (2011), Blundell and Matzkin (2014), Chen et al. (2014), Torgovitsky (2015), and D'Haultfœuille and Février (2015)

allowed to be flexible, which necessitates the general control function, $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o$, that can address endogeneity in this flexible manner. Because the binary response model is separable in the latent variable equation, identification is possible without CF-CI.

Throughout this section, identification of the first stage is assumed. It is up to the applied researcher to verify identification via usual rank conditions following Rothenberg (1971).

For the second stage, the main concern for identification is separately identifying the mean effect $\mathbf{x}_i\beta_o$ and the general control function $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o$. Because both of these terms are perfectly determined by $\mathbf{z}_i$ and v_{2i}, without any additional assumptions on the construction of $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)$, perfect multicollinearity is possible such that the parameters β_o and γ_o are not identified.[8] When linearity of the general control function is imposed, as in Assumption 3.1, identification requires $\mathrm{E}((\mathbf{x}_i, \boldsymbol{h}(v_{2i}, \mathbf{z}_i))'(\mathbf{x}_i, \boldsymbol{h}(v_{2i}, \mathbf{z}_i)))$ is non-singular. The following assumption provides sufficient conditions that will be used to show identification.

Assumption 4.1. *Let $\theta_o = (\beta_o', \gamma_o', \delta_o')' \in \Theta$ where Θ denotes the parameter spaces.*

(i) $\mathrm{E}(\boldsymbol{x}_i'\boldsymbol{x}_i)$ *is non-singular and the variance-covariance matrix of* $\mathrm{E}(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ *has full rank,*

(ii) $\mathrm{E}(\boldsymbol{h}(v_{2i}, \boldsymbol{z}_i)'\boldsymbol{h}(v_{2i}, \boldsymbol{z}_i))$ *is non-singular,*

(iii) *(CMR)* $\mathrm{E}(u_{1i} \mid \boldsymbol{z}_i) = 0$*,*

(iv) $\boldsymbol{g}(y_{2i}, \boldsymbol{z}_i)$ *consists of polynomial functions of the elements in* $\boldsymbol{x}_i$ *and* $\boldsymbol{h}(v_{2i}, \boldsymbol{z}_i)$*, does not include a constant, and* $\mathrm{E}(\boldsymbol{g}(y_{2i}, \boldsymbol{z}_i)'\boldsymbol{g}(y_{2i}, \boldsymbol{z}_i))$ *is non-singular,*

(v) $\theta_o = (\beta_o', \gamma_o')'$ *is a non-zero vector.*

The first three conditions are used to show no perfect multicollinearity. The second half of 4.1(i) is essentially a relevancy assumption that insures that no element of $\mathrm{E}(\mathbf{x}_i \mid \mathbf{z}_i)$ is non-random. If this part of the assumption would not hold, the the variation provided by the instruments cannot identify the endogenous effect. This part of the assumption can be replaced with an assumption of relevancy within the first stage for a specific $m(\mathbf{z}_,\pi)$.

The CMR is the identification assumption used in Kim and Petrin (2011). This assumption will be discussed in more detail later. The last two conditions, 4.1(iv)-(v), help in

[8] For example, if $\mathbf{x}_i = (1, \mathbf{z}_{1i}, y_{2i})$ then a general control function of the form $\boldsymbol{h}(v_{2i}, \mathbf{z}_i) = (\mathbf{z}_{1i}, \mathbf{z}_{2i}, v_{2i})$ creates perfect multicollinearity. Even when $\mathbf{z}_{1i}$ is excluded from the general control function (so $\mathbf{x}_i$ and $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)$ do not include the same terms) there is multicollinearity when $y_{2i} = \pi_0 + \mathbf{z}_{1i}\pi_1 + \mathbf{z}_{2i}\pi_2 + v_{2i}$.

showing identification in the highly non-linear heteroskedastic probit model. The following theorem states the identification result.

Theorem 4.1. *In the set-up described by equation (2.1), under Assumption 3.1, if Assumption 4.1 holds, then the parameters* $(\beta_o, \gamma_o, \delta_o)$ *are identified.*

The CMR approach to show identification using a control function is adopted from Kim and Petrin (2011) where they show non-parametric identification in a triangular system with an additively separable error.[9] The CMR requires $\mathbf{z}_i$ is mean independent of u_{1i} which is a fairly weak exogeneity condition on the controls and instruments compared to CF-CI. The CMR distinguishes between the endogeneity of y_{2i} and the "non-endogeneity" of $\mathbf{z}_i$. By the law of iterated expectations,

$$\mathrm{E}(u_{1i} \mid \mathbf{z}_i) = \mathrm{E}(\mathrm{E}(u_{1i} \mid \mathbf{z}_i, v_{2i}) \mid \mathbf{z}_i) = \mathrm{E}(\boldsymbol{h}(v_{2i}, \mathbf{z}_i) \mid \mathbf{z}_i)\gamma_o = 0.$$

The middle equality holds by the specification provided in Assumption 3.1 and the last equality holds by the CMR. Because this equality should hold for any possible $\gamma_o \in \Gamma$, the CMR requires $\mathrm{E}(\boldsymbol{h}(v_{2i}, \mathbf{z}_i) \mid \mathbf{z}_i)$ be a zero vector.

In practice, this means the CMR instructs on the construction of the general control function $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)$. To satisfy the CMR, each candidate element of $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)$ must be conditionally demeaned. This means no element can only be a function of $\mathbf{z}_i$ alone – the instruments can only enter as an interaction with functions of v_{2i}. This prevents any issues of linear dependence between elements of $\mathbf{x}_i$ and $\boldsymbol{h}(v_{2i}, \mathbf{z}_i)$. Wooldridge (2005) explains that identification holds given exclusion restriction on the instruments $\mathbf{z}_{2i}$ in the structural equa-

[9] Hahn and Ridder (2011) show that a "Conditional Mean Restriction" is insufficient for identifying the ASF in a general non-parametric non-separable model. However I would like to be clear that the CMR they consider is

$$\mathrm{E}(y_{1i} - \Psi(\mathbf{x}_i) \mid \mathbf{z}_i) = 0$$

where $\Psi(\mathbf{x}_i)$ is the unknown ASF. This differs from the CMR consider here which is on the latent error. Although the binary response model is non-separable, since the latent error is additively separable from the mean component $\mathbf{x}_i\beta_o$ within the indicator function, identification follows analogously from Kim and Petrin (2011) tion that creates variation in the control variate unexplained by $\mathbf{x}_i$.[10] Moreover any higher order functions need to be conditionally demeaned. For instance, v_{2i}^2 could not be an element of the general control function, but $v_{2i}^2 - \mathrm{E}(v_{2i}^2 \mid \mathbf{z}_i)$ could be.

It should be noted that CMR is not strictly weaker in the technical sense (i.e., CF-CI does not imply CMR) even though it is more plausible. Here is a simple example where CF-CI does hold but CMR does not hold:

$$u_{1i} \mid \mathbf{z}_i, v_{2i} \sim N(v_{2i} + v_{2i}^2 - \mathrm{Var}(v_{2i}), 1),$$

and suppose there is heteroskedasticity in the first stage, $\mathrm{Var}(v_{2i} \mid \mathbf{z}_i) = \exp(2\mathbf{z}_i\delta)$. Then the conditional expectation is non-zero,

$$\mathrm{E}(v_{2i} + v_{2i}^2 - \mathrm{Var}(v_{2i}) \mid \mathbf{z}_i) = \exp(2\mathbf{z}_i\delta) - \mathrm{Var}(v_{2i}),$$

and the CMR does not hold. In this example, the source of endogeneity (i.e., the conditional mean of u_{1i}) is quadratic in the first stage residual, but the quadratic term is deviated from the unconditional variance even though there is heteroskedasticity in the first stage. Therefore, the endogeneity depends on $(v_{2i}^2 - \mathrm{Var}(v_{2i}))$ instead of $(v_{2i}^2 - \exp(2\mathbf{z}_i\delta))$. If the quadratic term is deviated from the conditional variance, which would be more plausible, then CMR is satisfied and CF-CI is violated.

5 The Estimator

5.1 Estimation Procedure

The proposed estimation procedure is simple and easy to implement. In the first stage, the conditional mean function $\mathrm{E}(y_{2i} \mid \mathbf{z}_i) = m(\mathbf{z}_i, \pi_o)$ is estimated using (possibly a non-

[10] An alternative identification strategy is used in Escanciano et al. (2016) that does not require an exclusion restriction on the instruments z_{2i}. But in their setting, identification is dependent on non-linearity in the reduced form and they still impose CF-CI as a control function assumption.

linear) least squares. The control variable is constructed from the reduced form residuals, $\hat{v}_{2i} = y_{2i} - m(\mathbf{z}_i, \hat{\pi})$, and plugged into the second step. In the second stage, one would maximize the following log-likelihood

$$\begin{aligned}\mathcal{L}_n(\hat{\pi}, \beta, \gamma, \delta) = &\sum_{i=1}^{n} y_{1i} \log \left[\Phi\left(\frac{\mathbf{x}_i\beta + \boldsymbol{h}(\hat{v}_{2i}, \mathbf{z}_i)\gamma}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i)\delta)}\right)\right] \\ &+ (1 - y_{1i}) \log \left[1 - \Phi\left(\frac{\mathbf{x}_i\beta + \boldsymbol{h}(\hat{v}_{2i}, \mathbf{z}_i)\gamma}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i)\delta)}\right)\right]\end{aligned} \tag{5.1}$$

with respect to β, γ and δ to obtain estimates of the parameters. In addition to relaxing assumptions in the literature, the proposed estimation procedure is quite simple to implement using commands from standard statistical packages.[11] However, the estimated standard errors need to be adjusted to account for the variation from using the residual from the first stage as an approximation for the control variate. I will refer to the proposed estimator as the general control function probit estimator (or simply gcfprobit). Asymptotic variance formulas that account for the multi-step approach are given in the next section.

5.2 Asymptotic Properties

This section presents the asymptotic properties of the gcfprobit estimator. All proofs are provided in the appendix. Let $\theta' = (\beta', \gamma', \delta')$ fall within the parameter space Θ. Consistency follows from Theorem 2.1 of Newey and McFadden (1994).

Theorem 5.1. *In the set-up described by equation (2.1), under Assumptions 3.1 and 4.1, if the following holds,*

(i) *$\theta_o \in \Theta$ where Θ is compact,*

(ii) *the first stage estimator is consistent, $\hat{\pi} - \pi_o = o_p(1)$,*

(iii) *the control function, $\boldsymbol{h}(v, \boldsymbol{z}_i)$ is continuous in v and the first stage conditional mean, $m(\boldsymbol{z}_i, \pi)$ is continuous in π,*

then the second step estimator that maximizes the log likelihood in equation (5.1) is consistent, or $\hat{\theta} - \theta_o = o_p(1)$.

[11] For example, the parameter estimates can be obtained using `reg` and `hetprobit` commands in Stata.

First stage consistency is assumed. Deriving asymptotic normality follows from Theorem 6.1 in Newey and McFadden (1994).

Theorem 5.2. *In the set-up described by equation (2.1) under Assumption 3.1, Assumption 4.1 and Theorem 5.1(i)-(iii), if the following holds,*

(i) $\theta_o \in \operatorname{int}(\Theta)$,

(ii) $v_{2i}, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{h}(v_{2i}, \boldsymbol{z}_i), \boldsymbol{g}(y_{2i}, \boldsymbol{z}_i)$, *and* $\bigtriangledown_\pi m(\boldsymbol{z}_i, \pi_o)$ *have finite second moments,*

(iii) $\mathrm{E}(\bigtriangledown_\pi m(\boldsymbol{z}_i, \pi_o)(\bigtriangledown_\pi m(\boldsymbol{z}_i, \pi_o))')$ *is non-singular,*

then the second step estimator that maximizes the log likelihood in equation (5.1) is asymptotically normal, or $\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, V)$ *where*

$$V = G_{2\theta}^{-1}\,\mathrm{E}\left(B(\boldsymbol{w}_i, \pi_o, \theta_o)B(\boldsymbol{w}_i, \pi_o, \theta_o)'\right)G_{2\theta}^{-1\prime}$$

where $B(\boldsymbol{w}_i, \pi_o, \theta_o) = S(\boldsymbol{w}_i, \pi_o, \theta_o) + G_{2\pi}G_{1\pi}^{-1}(y_{2i} - m(\boldsymbol{z}_i, \pi_o))\bigtriangledown_\pi m(\boldsymbol{z}_i, \pi_o)$ *where* $S(\cdot)$ *is the score from the second stage likelihood and*

$$\mathrm{E}(\bigtriangledown_{(\pi', \theta')'} M(\boldsymbol{w}_i, \pi_o, \theta_o)) = \begin{pmatrix} G_{1\pi} & G_{1\theta} \\ G_{2\pi} & G_{2\theta} \end{pmatrix}$$

where $M(\cdot)$ *is the stacked fist stage and second stage moment conditions. All terms are defined in detail in the appendix.*

The asymptotic variance takes into account the variation introduced from the first stage. A consistent estimator for the asymptotic variance would be the method of moments estimator that replaces all the unknown parameters with their consistent estimates and then use sample averages in place of expectations. Although this section provides consistency and $\sqrt{n}$-asymptotic normality for the second stage parameter estimates, the parameters themselves bear very little interpretative value. Next, I will discuss the derivation of the ASF and its importance for economic interpretation. This structural object is useful for empirical researchers to discuss the effectiveness of a particular policy or the average probability of a successful outcome for an individual with a particular set of characteristics.

6 Average Structural Function

Researchers are often interested in using model estimates to infer the average predicted probability of success at possible values of the covariates (counterfactuals). In the absence of endogeneity, this quantity can be easily described by the conditional mean, which in the case of binary response, is equivalent to the propensity score. But when endogeneity is present, the conditional mean is unable to capture the structural relationship between the endogenous variable and the outcome. This section clarifies the role of endogeneity and CF-CI, in deriving BP's ASF for binary response models.

6.1 Deriving ASF in a Linear Model

For clarification, first consider a simple linear structural equation,

$$y_i = \mathbf{x}_i \beta_o + u_i. \tag{6.1}$$

Without endogeneity, $\mathrm{E}(u_i \mid x_i) = 0$ and the interpretation of the average outcome for a given observation $\mathbf{x}^o$ is simply the conditional mean: $\mathbf{x}^o \beta_o$. The corresponding partial effect is the slope parameter β_o. But when endogeneity is introduced, $\mathrm{E}(u_i \mid \mathbf{x}_i) \neq 0$, the conditional mean is composed of two parts:

$$\mathrm{E}(y_i \mid \mathbf{x}_i = \mathbf{x}^o) = \mathbf{x}^o \beta + \mathrm{E}(u_i \mid \mathbf{x}_i = \mathbf{x}^o).$$

The first component is the structural direct effect of $\mathbf{x}_i$ while the second component is the endogenous indirect effect of $\mathbf{x}_i$ due to the presence of endogeneity.

For instance, consider the ubiquitous example of returns to education where education is endogenous due to unobserved ability. Then the structural direct effect is the average wage for particular education level (independent of ability) and the endogenous indirect effect is the contribution of average ability for that given education level on wages. BP argues that one should only be interested in the structural direct effect because if one were to consider a policy intervention on the level of education (ie: mandatory schooling), there would be no changes in the distribution of ability. Consequently, only the structural direct effect will capture the effect a policy maker should be interested in.

The ASF is derived from integrating over the unconditional distribution of the unobserved error in the structural equation. In the above example, this is essentially enforcing the belief that any structural changes to eduction levels should not alter the distribution of ability (i.e., the unconditional distribution of the error). If the structural equation (6.1) includes an intercept, then $\mathrm{E}(u_i) = 0$ and the ASF is $\mathbf{x}^o\beta_o$, not equal to the conditional mean under endogeneity, but equivalent to the conditional mean in the case of no endogeneity.

6.2 Deriving ASF in a Binary Response Model

Next is to extend the analysis (same as the derivations in Lin and Wooldridge (2015)) to the binary response model,

$$y_i = 1\{\mathbf{x}_i\beta_o + u_i > 0\}. \tag{6.2}$$

When there is independence between the latent error u_i and the regressors $\mathbf{x}_i$, the conditional mean – equivalent to the propensity score – is

$$\mathrm{E}(y_i \mid \mathbf{x}_i = \mathbf{x}^o) = F_{-u}(\mathbf{x}^o\beta_o)$$

which calculates the probability of success for an individual with characteristics $\mathbf{x}^o$.

Now consider the case when there is no longer independence between the latent error and the regressors so the unconditional CDF is not equal to the conditional CDF; i.e., $F_{-u}(-u) \neq F_{-u|\mathbf{x}}(-u; \mathbf{x})$ where $F_{-u|\mathbf{x}}(\cdot;\cdot)$ is the conditional CDF in which the first argument is the point of evaluation and the second argument is the conditioning argument. One can understand the violation of independence either through the standard interpretation of endogeneity, $\mathrm{E}(u_i \mid \mathbf{x}_i) \neq 0$, or possible due to "endogeneity" at higher moments such as heteroskedasticity, $\text{Var}(u_i \mid \mathbf{x}_i) \neq \text{Var}(u_i)$. In this case, the propensity score is

$$\text{E}(y_i \mid \mathbf{x}_i = \mathbf{x}^o) = F_{-u|\mathbf{x}}(\mathbf{x}^o \beta_o; \mathbf{x}^o)$$

in which the first argument in $F_{-u|\mathbf{x}}(\mathbf{x}^o \beta_o; \mathbf{x}^o)$ is the point of evaluation which, corresponds to the structural direct effect, and the second argument is the conditioning argument, which corresponds to the endogenous indirect effect.

As in the linear case, the conditional mean does not capture the desired structural interpretation. The ASF is derived by integrating over the unconditional distribution of the latent error to obtain: $F_{-u}(\mathbf{x}^o \beta_o)$. Now the ASF only captures the structural direct effect of $\mathbf{x}_i$ and is not clouded by the influence of endogeneity. Calculating the ASF requires knowledge of the unconditional distribution of u_i. But usually only a conditional distribution, like $u_{1i} \mid \mathbf{z}_i, v_{2i}$, is specified when estimating the structural parameters β_o under endogeneity.

Wooldridge (2005) expands on the discussion in BP to derive a formula for the ASF when only a conditional distribution is assumed to be known. Using the same notation as above, the structural model of interest is $\mu_1(\mathbf{x}_i, u_i) \equiv \text{E}(y_i \mid \mathbf{x}_i, u_i)$, where $\mathbf{x}_i$ is observed covariates and u_i is unobserved heterogeneity. Then the ASF is defined as

$$ASF(\mathbf{x}^o) \equiv \text{E}_u(\mu_1(\mathbf{x}^o, u_i))$$

where the subscript of u emphasizes that the expectation is taken with respect to the unconditional distribution of u_i. Using Lemma 2.1 from Wooldridge (2005), which is essentially an application of law of iterated expectations, the ASF can also be calculated from

$$\begin{aligned} ASF(\mathbf{x}^o) &= E_w(\mu_2(\mathbf{x}^o, w_i)) \\ \mu_2(\mathbf{x}^o, w_i) &= \int_{\mathcal{U}} \mu_1(\mathbf{x}^o, u) f_{u|w}(u; w_i) \eta(\text{d}u) \end{aligned}$$

where $\mathcal{U}$ is the support of u_i and $f_{u|w}(\cdot;\cdot)$ is the conditional density of the unobserved heterogeneity u_i given w_i with respect to a σ-finite measure $\eta(\cdot)$. The consequences of this result is that the conditioning argument w_i can be used to help identify the ASF. It is important to note that the evaluation of the ASF requires the ability to distinguish between the point of evaluation $\mathbf{x}^o$ and the conditioning argument w_i.[12]

To apply Lemma 2.1, the following two conditions must hold

(i) *(Ignorability)* $\mathrm{E}(y_i \mid \mathbf{x}_i, u_i, w_i) = \mathrm{E}(y_i \mid \mathbf{x}_i, u_i)$,

(ii) *(Conditional Independence)* $u_i \mid \mathbf{x}_i, w_i \sim u_i \mid w_i$.

The Conditional Independence condition here (used to identify the ASF) should not be confused with the CF-CI assumption of BP and Rothe (used to identify the parameters). This Conditional Independence condition is with respect to the conditioning argument, w_i, which has yet to be specified in our context.

Consider using the control variate, v_{2i}, as the conditioning argument, w_i. Ignorability easily holds since controlling for $\mathbf{x}_i$ and u_i is the same as controlling for $\mathbf{x}_i$,u_i, and v_{2i}. Then the Conditional Independence condition reduces to the CF-CI assumption. This means the CF-CI assumption of BP, used to show identification of parameters, can also be used to obtain identification of the ASF. But identification of the parameters was obtainable without CF-CI, so can we similarly identify the ASF in the absence of CF-CI?

Consider using both the control variate v_{2i} and the instruments $\mathbf{z}_i$ as the conditioning argument, w_i. This easily satisfies the ignorability assumption $\mathrm{E}(y_{1i} \mid \mathbf{x}_i, u_{1i}, v_{2i}, z_i) = \mathrm{E}(y_{1i} \mid \mathbf{x}_i, u_{1i})$ given ignorability of the excluded instruments $\mathbf{z}_{2i}$. It also satisfies the Conditional

[12] Wooldridge (2005) considers the example of the heteroskedastic probit model where in equation (6.2), it is assumed u_i is mean zero, normally distributed with $\mathrm{Var}(u_i \mid \mathbf{x}_i) = \exp(2\mathbf{x}_i\delta)$. Then the covariates $\mathbf{x}_i$ are used as the conditioning argument (i.e., $w_i = \mathbf{x}_i$) such that

$$\begin{aligned} ASF(\mathbf{x}^o) &= E_{\mathbf{x}_i}\left(\int_{\Re} 1\{\mathbf{x}^o\beta_o + u > 0\} f_{u|\mathbf{x}}(u;\mathbf{x}_i)\mathrm{d}u\right) \\ &= E_{\mathbf{x}_i}\left(\Phi\left(\frac{\mathbf{x}^o\beta}{\exp(\mathbf{x}_i\delta)}\right)\right) \end{aligned}$$

where the expectation is taken with respect to the $\mathbf{x}_i$ in the heteroskedastic function (part of the conditioning argument) and not with respect to the structural direct effect of $\mathbf{x}^o$. Therefore, even when the conditioning argument is the same as the covariates in the structural equation, it is necessary to be able to distinguish between the two sources when composing the ASF.

Independence condition, $u_{1i} \mid \mathbf{x}_i, v_{2i}, \mathbf{z}_i \sim u_{1i} \mid v_{2i}, \mathbf{z}_i$, since $\mathbf{x}_i$ is composed of functions of v_{2i} and $\mathbf{z}_i$. This means the ASF is identified without CF-CI. Under Assumption 3.1,

$$\begin{aligned}
\mu_2(x^o, (v_{2i}, \mathbf{z}_i)) &= \int_{\Re} 1\{\mathbf{x}^o\beta_o + u > 0\} f_{u|v,\mathbf{z}}(u; v_{2i}, \mathbf{z}_i)\mathrm{d}u \\
&= E_u(1\{\mathbf{x}^o\beta_o + u > 0\} \mid v_{2i}, \mathbf{z}_i) \\
&= \Phi\left(\frac{\mathbf{x}^o\beta_o + \boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i)\delta_o)}\right)
\end{aligned}$$

and the ASF is

$$ASF(\mathbf{x}^o) = E_{v_2,z}(\mu_2(\mathbf{x}^o, (v_{2i}, \mathbf{z}_i))) = E_{v_2,\mathbf{z}}\left(\Phi\left(\frac{\mathbf{x}^o\beta_o + \boldsymbol{h}(v_{2i}, \mathbf{z}_i)\gamma_o}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i)\delta_o)}\right)\right) \tag{6.3}$$

where the expectation is taken with respect to the unconditional distribution of v_{2i} and $\mathbf{z}_i$. A consistent method of moments estimator would replace the unknown parameter values with their consistent estimates, $(\hat{\pi}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$, and in place of the expectation, take sample averages.

6.3 Comparing ASF under CF-CI and CF-LI

So CF-CI plays a role in both identifying the parameters and deriving the ASF. Therefore the consequence of presuming CF-CI when it does not hold can be two-fold: inconsistent parameter estimates and incorrect ASF derivation. To better understand the later effect, we need to derive the ASF under CF-CI. Let $G_{CF-CI}(\cdot; v_{2i})$ be the conditional CDF of $-u_{1i} \mid v_{2i}, \mathbf{z}_i$, an unknown function that BP and Rothe estimate non-parametrically. The conditioning arguments $\mathbf{z}_i$ is excluded from the $G_{CF-CI}()$ function because CF-CI imposes this exclusion restriction. Then ASF derived under the presumption of CF-CI is,

$$ASF_{CF-CI}(\mathbf{x}^o) = E_{v_2}(G_{-u_1|v_2}(\mathbf{x}^o\beta_o; v_{2i})) \tag{6.4}$$

where the expectation is taken with respect to v_{2i}. Comparing equations (6.3) and (6.4), highlights the impact of incorrectly presuming CF-CI on interpretation. Since there is endogeneity, the effect of $\mathbf{x}^o$ on the predicted probability of success can be broken down between the structural direct effect and an endogenous indirect effect. The allure of the CF-CI assumption is it immediately distinguishes between the two effects in the conditional distribution function $G_{u_1|v_2}(\mathbf{x}^o\beta_o; v_{2i})$ where the first argument captures the structural direct effect and the second argument should entirely control for endogenous indirect effect. But when CF-CI fails, and this structure of the conditional CDF is still presumed, the lines between the structural direct effect and an endogenous indirect effect become blurred. Consequently, the ASF calculated when incorrectly imposing CF-CI does not correctly average out the endogenous indirect effect. This means that in addition to inconsistent parameter estimates, presuming CF-CI results in incorrect derivation of the ASF.

As a more flexible alternative to assuming CF-CI, Rothe proposes CF-LI which allows the conditional distribution to be a function of the instruments through the linear index $\mathbf{x}_i\beta_o$,

$$u_{1i} \mid v_{2i}, \mathbf{z}_i \sim u_{1i} \mid v_{2i}, \mathbf{x}_i\beta_o.$$

Utilizing results from Manski (1988), β_o and

$$G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i}) = F_{u_1|v_2,\mathbf{x}\beta_o}(\mathbf{x}_i\beta_o; v_{2i}, \mathbf{x}_i\beta_o),$$

the CDF of u_{1i} conditional on and evaluated at $\mathbf{x}_i\beta_o$, are both identified. As discussed previously, the CF-LI assumption is still a fairly strong restriction on the conditional distribution of $u_i \mid v_{2i}, \mathbf{z}_i$. Compared to the specification in Assumption 3.1, CF-LI require the control function and the heteroskedastic function to be constructed using only the linear index and not as more flexible functions of the instruments. But for now, consider the most optimistic case where the CF-LI assumption holds such that the Rothe SML estimator for the parameters is consistent. Can we also correctly identify the ASF under CF-LI?

Applying Lemma 2.1 from Wooldridge (2005), the true ASF when CF-LI holds is

$$ASF_{CF-LI}(\mathbf{x}^o) = E_{v_2,\mathbf{x}\beta_o}(F_{-u_1|v_2,\mathbf{x}\beta_o}(\mathbf{x}^o\beta_o; v_{2i}, \mathbf{x}_i\beta_o)) \tag{6.5}$$

where the expectation is taken with respect to the joint distribution of the conditioning arguments $(v_{2i}, \mathbf{x}_i\beta)$. The immediate issue is that the ASF cannot be written in terms of the identified function $G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i})$ that is estimated using the proposed SML estimator in Rothe. The identified function is the conditional CDF evaluated at and condition on the same linear index. Therefore one cannot distinguish between the direct structural effect and indirect endogenous effect of the linear index so the true ASF is not identified.

Rothe suggests using $E_{v_2}(G_{CF-LI}(\mathbf{x}^o\beta_o, v_{2i}))$ as the ASF but this only averages out the part of the endogenous indirect component due to v_{2i}, and does not average out any the effect due to the linear index. Therefore, the ASF proposed by Rothe is equal to the true ASF only when the more restrictive CF-CI assumption of BP holds. So although it may be tempting to consider the CF-LI assumption as a compromise to allow for flexibility in terms of the relationship between the latent error and the instruments, the true ASF is not identified under the CF-LI assumption.

7 Simulation

7.1 Set Up

This simulation study examines the finite-sample performance of the proposed gcfprobit estimator in several settings. The variety of settings explore when the proposed approach is best suited and when other approaches may outperform the proposed approach. In each design, there is one included and one excluded instrument that are both independent and normally distributed with mean 0 and variances 3 and 1 respectively. The common data generating process is

$$
\begin{aligned}
y_{1i} &= \begin{cases} 1 & y_{1i}^* \geq 0 \\ \\ 0 & y_{1i}^* < 0 \end{cases} \\
y_{1i}^* &= y_{2i}\beta_o + z_{1i} + u_{1i} \\
y_{2i} &= \pi_{1o} + \pi_{2o} z_{1i} + \pi_{3o} z_{2i} + v_{2i};
\end{aligned}
$$

where $\beta_o = 1$ and $\pi_o = \left(-1/\sqrt{2}, -1/\sqrt{6}, 1/\sqrt{2}\right)'$. The control variate v_{2i} is drawn from a $N(0,1)$. This means that there is a strong first stage with an R^2 of approximately 0.50.

The conditional distribution of the latent error u_{1i} will vary across the following four designs:

- Design I: Heteroskedastic Normal

$$u_{1i} \mid v_{2i}, \mathbf{z}_i \sim \mathrm{N}\left(-\frac{1}{3}v_{2i} + \frac{2}{5}(z_{1i} \times v_{2i}) + \frac{2}{3}(z_{2i} \times v_{2i}), \exp(\frac{1}{2}z_{1i} - \frac{1}{3}y_{2i})\right),$$

- Design II: Heteroskedastic Normal that satisfies CF-LI

$$u_{1i} \mid v_{2i}, \mathbf{z}_i \sim \mathrm{N}\left(\frac{2}{3}v_{2i} - \frac{4}{8}(v_{2i} \times (y_{2i}\beta_o + z_{1i}) - \sigma_v^2\beta_o), \exp(\frac{2}{3}(y_{2i}\beta_o + z_{1i}))\right),$$

- Design III: Symmetric Mixture of Normals

$$\begin{aligned} u_{1i} \mid v_{2i}, \mathbf{z}_i \sim & \left(-\frac{1}{3}v_{2i} + \frac{2}{5}(z_{1i} \times v_{2i}) + \frac{2}{3}(z_{2i} \times v_{2i})\right) \\ & + [0.5\,\mathrm{N}(-1,1) + 0.5\,\mathrm{N}(1,1)], \end{aligned}$$

- Design IV: Skewed and Bimodel Mixture of Normals

$$\begin{aligned} u_{1i} \mid v_{2i}, \mathbf{z}_i \sim & \left(-\frac{1}{3}v_{2i} + \frac{2}{5}(z_{1i} \times v_{2i}) + \frac{2}{3}(z_{2i} \times v_{2i})\right) \\ & + \sqrt{0.4} \times [0.8\,\mathrm{N}(-1,0.6) + 0.2\,\mathrm{N}(4,2)], \end{aligned}$$

The first distribution fits Assumption 3.1 in which CF-CI is violated due to the presence of heteroskedasticity and a general control function. In this setting, the proposed gcfprobit estimator should outperform any other approach that presumes CF-CI for identification.

The next distribution considers the case in which CF-CI is violated but CF-LI holds. This means the heteroskedastic and general control functions are functions of the instruments but only through the linear index $(y_{2i}\beta_o + z_{1i})$. In this setting, both the proposed gcfprobit and alternative SML estimator from Rothe (2009) should produce accurate parameter estimates. However, as discussed in the last section, the SML estimator cannot correctly identify the ASF and therefore the gcfprobit estimator will outperform in terms of interpretation.

The third distribution is a bimodal mixture of normals that includes a general control function as a violation of CF-CI. This design is introduced to investigate the performance of the proposed approach under distributional miss-specification. As discussed earlier and shown in the appendix, the inclusion of a heteroskedastic specification provides some flexibility in terms of distributional miss-specification. In this setting, the proposed approach should still produce consistent parameter and ASF estimates. The last design is a mixture of normals that is bimodal and skewed. This is a scenario in which the proposed approach makes no claims on identification or consistency. Because symmetry does not hold, the result from Khan (2013) does not apply and distributional miss-specification could strongly bias the gcfprobit estimates. For all three distributions, results are gathered for 1000 replications of sample sizes equal to 500, 1000, and 2000.

7.2 Results

In addition to the proposed gcfprobit estimator, the simulation experiment will employ the two-step control function probit estimator of RV (ivprobit) and the semi-parametric maximum likelihood (SML) estimator[13] of Rothe (2009) for comparison. All estimator used the same (correctly specified) first stage: OLS estimates from regressing y_{2i} on an intercept, z_{1i}, and z_{2i}.

[13] The SML estimator is implemented with a Gaussian kernel of order 2. Although asymptotically the SML estimator requires higher order kernels, Rothe finds that lower-order kernels perform better in small samples. As suggested in Rothe (2009), bandwidths for the SML estimator were treated as additional parameters to be optimized over. To avoid local maximum, 1/2 and 2 times the starting values (ivprobit estimates) were also used.

Table 1: Simulation Results for Design I

n		Bias	SD	RMSE	25%	50%	75%
500	ivprobit	−0.06989	0.101	0.123	0.868	0.931	0.997
	SML	−0.13603	0.098	0.168	0.794	0.854	0.930
	gcfprobit	−0.00529	0.121	0.121	0.916	0.995	1.072
1000	ivprobit	−0.06902	0.070	0.098	0.883	0.935	0.978
	SML	−0.13513	0.072	0.153	0.814	0.861	0.911
	gcfprobit	0.00015	0.077	0.077	0.946	1.003	1.057
2000	ivprobit	−0.07146	0.050	0.087	0.896	0.927	0.962
	SML	−0.13352	0.057	0.145	0.829	0.864	0.905
	gcfprobit	−0.00245	0.054	0.054	0.961	1.001	1.035

Table 2: Simulation Results for Design II

n		Bias	SD	RMSE	25%	50%	75%
500	ivprobit	−0.00454	0.151	0.151	0.904	1.001	1.092
	SML	0.00311	0.105	0.105	0.933	1.008	1.070
	gcfprobit	−0.00690	0.109	0.109	0.927	0.996	1.061
1000	ivprobit	0.00068	0.102	0.102	0.934	1.006	1.069
	SML	0.01867	0.071	0.074	0.971	1.021	1.070
	gcfprobit	−0.00002	0.072	0.072	0.950	1.000	1.050
2000	ivprobit	0.00006	0.072	0.072	0.953	1.004	1.051
	SML	0.02084	0.053	0.057	0.986	1.022	1.056
	gcfprobit	0.00003	0.053	0.053	0.965	1.001	1.034

During implementation, the proposed estimator was found to be fairly sensitive to different starting values. Using 16 randomized starting values (centered around the ivprobit estimates) helped to avoid local maxima.

Tables 1-4 reports the simulation results for simulation Designs I-IV respectively. In Table 1, the proposed gcfprobit estimator has a much smaller bias compared to the ivprobit or SML estimator and the smallest RMSE. This is unsurprising since this is the design in which only the proposed approach uses a valid identification strategy (and is therefore consistent).

For Design II, it is expected for both the proposed gcfprobit estimator and the SML estimator to perform well for parameter estimates because CF-LI holds (the minimal identification assumption needed for SML to be consistent). In this case all three estimators

Table 3: Simulation Results for Design III

n		Bias	SD	RMSE	25%	50%	75%
500	ivprobit	−0.02462	0.121	0.123	0.901	0.976	1.058
	SML	−0.09157	0.107	0.141	0.837	0.907	0.982
	gcfprobit	−0.00559	0.141	0.141	0.903	0.996	1.085
1000	ivprobit	−0.02118	0.086	0.088	0.923	0.976	1.036
	SML	−0.07551	0.076	0.107	0.877	0.926	0.977
	gcfprobit	−0.00778	0.095	0.095	0.929	0.993	1.057
2000	ivprobit	−0.02157	0.058	0.062	0.940	0.979	1.016
	SML	−0.07040	0.053	0.088	0.892	0.930	0.964
	gcfprobit	−0.00468	0.062	0.062	0.951	0.995	1.038

Table 4: Simulation Results for Design IV

n		Bias	SD	RMSE	25%	50%	75%
500	ivprobit	0.01329	0.118	0.119	0.939	1.015	1.096
	SML	−0.03431	0.086	0.093	0.908	0.968	1.020
	gcfprobit	0.01648	0.111	0.112	0.946	1.018	1.090
1000	ivprobit	0.01452	0.087	0.088	0.957	1.015	1.072
	SML	−0.02783	0.060	0.066	0.933	0.972	1.013
	gcfprobit	0.01754	0.077	0.079	0.967	1.016	1.070
2000	ivprobit	0.01302	0.061	0.062	0.973	1.014	1.054
	SML	−0.02598	0.044	0.051	0.945	0.975	1.001
	gcfprobit	0.01923	0.052	0.055	0.984	1.020	1.054

perform very well. The ivprobit estimator performs surprisingly well in this design but recall that the parameter estimate in this case is a ratio of coefficient estimates which means the impact of heteroskedasticity appears to be smaller than it actually is.

Table 3 reports the simulation results when introducing (symmetric) distributional misspecification. The SML estimator is a semi-parametric distribution-free estimator and therefore could perform better than the proposed approach which makes a distributional assumption (although CF-CI still does not hold). But because the distribution is symmetric, utilizing the result from Khan (2013), the proposed approach which incorporates heteroskedasticity will be "robust" to distributional misspecification[14] Consequently, the gcfprobit has a much

[14] This paper is not proposing a semi-parametric approach (nonparametric heteroskedasticity) that Khan (2013) uses but rather uses his result to explain why the parametric estimator does not produce inconsistent parameter estimates even under distributional misspecification.

smaller bias compared to the other two approaches. But when the sample size is small ($n = 500$), the RMSE for the gcfprobit estimator is as large as the RMSE for the SML estimator. This is because the proposed approach is fairly inefficient at small sample sizes. With the addition of the numerous parameters in the general control function and the heteroskedastic function, the gcfprobit estimator can be fairly imprecise.

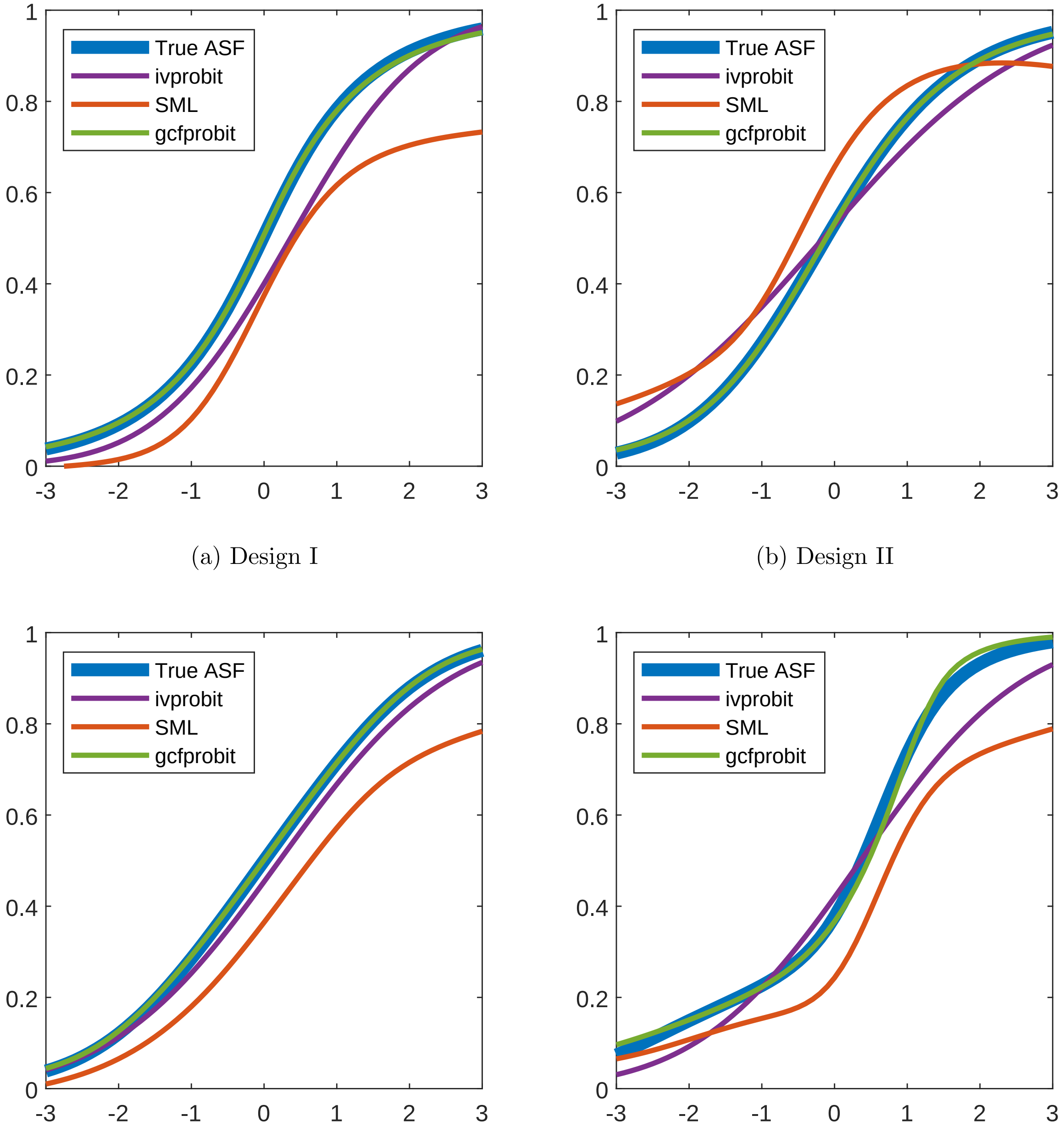
In the last design, the proposed estimator makes no claims on a strong performance. In this specification, there is non-symmetric distributional misspecification so the result from Khan (2013) does not apply. However, even in this setting, the proposed approach does not perform any worse than the SML estimator in terms of bias, or any worse than the ivprobit estimator in terms of RMSE. So even though the proposed approach does not perform as well as it does with the other designs, it does not appear to be performing any worse than the existing approaches.

Figure 1 looks at the ASF estimates for each of the four designs, derived as discussed in Section 6.2. Because CF-CI does not hold in Design I and III and the true distribution of the latent error is symmetric, the proposed gcfprobit estimator will produce accurate ASF estimates while the other approaches will perform poorly. Figures 1a and 1c confirm our expectation that only the proposed gcfprobit estimator produces ASF estimates that closely follow the true ASF.

For design II, CF-LI holds which means the SML estimator produced consistent parameter estimates. But recall the discussion in Section 6.3, the SML estimator cannot separately identify the different roles the linear index plays in calculating the ASF. As seen in Figure 1b, this means the SML estimator does not produce the correct ASF. This reiterates the conclusions in Section 6: imposing CF-CI plays a role in both identifying the parameters and deriving the ASF, if it does not hold, then there can be a doubly detrimental effect.

Finally, the proposed gcfprobit estimator produces the most accurate ASF estimates even in Design IV where there are no claims made on the accuracy of the proposed approach. This result is further encouragement that distributional miss-specification does not play as strong

Figure 1: Average Structural Function Estimates.



(a) Design I

(b) Design II

(c) Design III

(d) Design IV

Point-wise average ASF estimates over 1,000 simulations for sample size $n = 2000$.

Table 5: Summary Statistics

Variable		Mean	SD	Mean (If $y_1 = 0$)	Mean (If $y_1 = 1$)
Employed	(y_1)	0.601	0.490		
Non-Wife Inc ($1000)	(y_2)	30.534	26.627	35.835	27.012
Experience	(z_{11})	15.924	5.664	16.087	15.816
Has kids (age<6)	(z_{12})	0.485	0.500	0.566	0.431
Education	(z_{13})	13.185	2.565	12.694	13.511
Husband's Education	(z_2)	13.429	2.846	13.202	13.580
Observations		2,808		1,121	1,687

1991 CPS data on Married Women Labor force participation.

of a role as incorrectly presuming identification assumptions in terms of the accuracy of ASF estimates.

8 Empirical Example

To showcase the estimator in an empirical example, I examine married women's labor force participation using 1991 CPS data.[15] Table 5 provides some summary statistics for the data set. The dependent variable is Employed (=1 when the individual is in the labor force) where approximately 58% of married women in the sample participate in the labor force. The last two columns divide the sample over the binary outcome and reports the summary statistics for the other observable characteristics.

The model for labor force participation is

$$\begin{aligned} emp_i = 1\{&\beta_1 + nwifinc_i\beta_2 + nwifinc_i \times kidslt6_i\beta_3 \\ &+ exper_i\beta_4 + kidslt6_i\beta_5 + educ_i\beta_6 - u_{1i} \geq 0\} \end{aligned}$$

where the economic interest is in estimating the effect of annual non-wife income on the probability of being in the labor force. Since there is a trade-off between work and leisure,

[15] Data is part of the supplementary material provided with Wooldridge (2010). Data can be downloaded at https://mitpress.mit.edu/books/econometric-analysis-cross-section-and-panel-data

by relaxing the budget constraint such that an individual has other sources of income, one would expect the individual to be less likely to work. From the summary statistics, those not working tend to have higher non-wife income. But this cannot be interpreted as a causal effect since there is concern that other sources of income are endogenously determined with the wife's labor force participation. In particular, husband's employment, which partly determines the non-wife income, would be decided simultaneously with wife's employment.

Moreover, the employment decision-making process differs over the number and ages of children in the household. For instance, if there are very young children in the household then the trade-off is not just between work and leisure but must also consider the cost of childcare if both parents enter the workforce. This sort of interaction would violate CF-CI.

Utilizing husband's education level as an instrument, the causal effect of non-wife income on wife's labor force participation can be parsed out. Since education and the probability of working are generally correlated, the instrument is relevant with a fairly high F-statistic of 182.1. Excludability of the instrument follows from the argument that husband's education level should not directly affect the wife's choice of labor force participation except through the channels of how it affects the non-wife income. The other controls considered in this example are the wife's education level, experience, and dummy variables for whether or not they have kids younger than 6 and kids 6 and older.

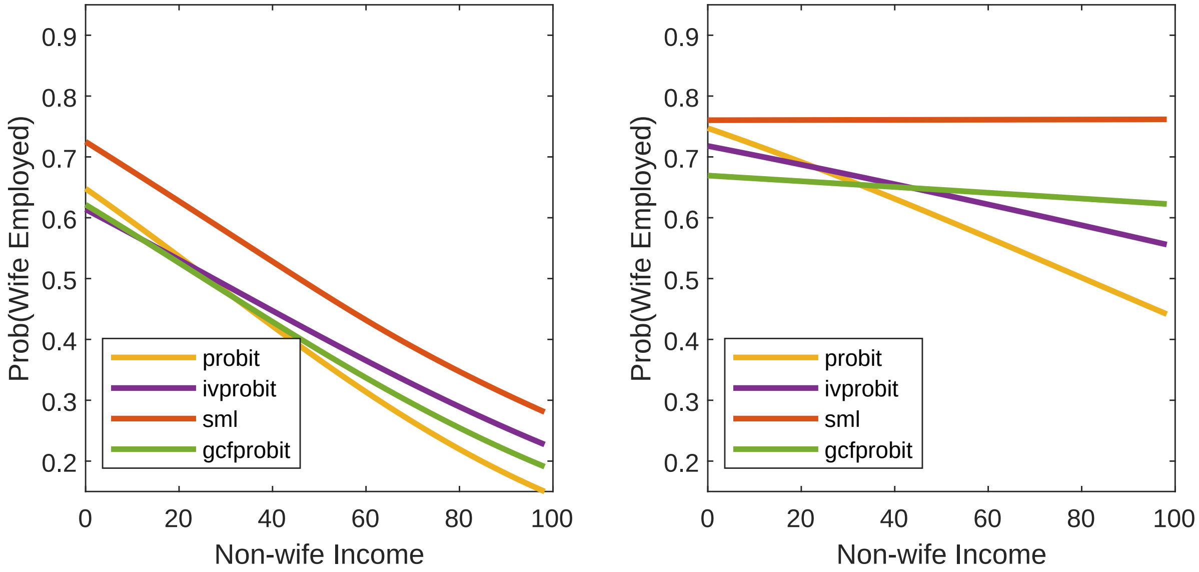
Table 6 reports the second stage parameter estimates. The coefficient on education is normalized to one to allow for comparisons across the different estimators. The first column specifies a standard probit model that assumes no endogeneity and homoskedasticity in the latent error. The next column is the ivprobit estimator from RV, which is only valid when the latent error is normally distributed and CF-CI holds. Columns (3) and (4) are the SML estimator and the proposed gcfprobit estimator respectively. They both produce much smaller (in absolute value) coefficient estimates for non-wife income, and relatively larger (in absolute value) coefficient estimates for the interaction of non-wife income with the presence of children in the household. This setting also highlights a major concern with the proposed

Table 6: Coefficient Estimates for Married Women's LFP

Variable	probit (1)	ivprobit (2)	SML (3)	gcfprobit (4)
Non-wife Income	−0.073	−0.044	0.000	−0.011
	(0.012)	(0.034)	(0.037)	(0.047)
Non-wife Income	−0.054	−0.061	−0.156	−0.098
× Has Kids (Age<6)	(0.018)	(0.022)	(0.052)	(0.266)
Experience	−0.077	−0.111	−0.269	−0.034
	(0.052)	(0.078)	(0.114)	(0.599)
Has Kids (Age<6)	−2.536	−2.855	−1.146	−1.147
	(0.799)	(1.022)	(1.560)	(4.385)
Education	1	1	1	1
Control Function Parameters				
$\hat{v}_{2i}$		−0.040		−0.087
		(0.044)		(0.076)
$\hat{v}_{2i}\times$ Has Kids (Age<6)				0.084
				(0.210)
Heteroskedastic Parameters				
Non-wife Income				−0.004
				(0.003)
Education				0.025
				(0.102)
Experience				0.024
				(0.120)

1991 CPS data on Married Women Labor force participation. Standard errors given in parenthesis and calculated using the formula in Section 5. The standard errors for the SML estimates are obtained using a boostrap with 500 replications. All estimators use the same first stage estimates from regressing non-wife income on an intercept, experience, dummy variable for young kids, education, and husband's education. F-statistic from the first stage is 182.116 (for statistical significance of the coefficient on the excluded instrument).

Figure 2: ASF Estimates for Households with and without Young Children



(a) Has Young Kids (age< 6)

(b) Does not have Young Kids

1991 CPS data on Married Women Labor force participation. Average Structural Function is with respect to non-wife income (between 5th and 95th percentile) for a married woman with high school education and 10 years of experience.

approach, that it can be drastically inefficient relative to the other approaches

Figure 2 displays the ASF with respect to non-wife income for a married woman with high school education, 10 years of experience. The first thing to note is that the ASF using probit estimates is generally more negatively sloped in both the figures. This is because without addressing the issue of endogeneity, (i.e., if the husband works then the wife is less likely to work and vice versa, but these decisions are made simultaneously) one would expect to see this sort of substitution effect. Once endogeneity is controlled for in ivprobit, the slope lessens. The proposed gcfprobit and SML estimators start to deviate more from the ivprobit ASF estimates for households that do not have young children. This intuitively makes sense since without children at home, there is a much higher opportunity cost to not working and so one would expect a fairly flat ASF. Both the gcfprobit and SML estimators flexibly allow differences in the presence of children to change the effect endogeneity has on the estimates. For the SML estimator, this is restricted through the linear index (CF-LI), whereas the proposed approach is much more flexible. Moreover, because the ASF are fairly linear and smooth across all estimators, including SML, one would suspect that distributional misspecification is not making a large impact and differences in estimates are mostly due to differences in identification strategies.

9 Conclusion

This paper presents a new control function approach to endogeneity in a binary response model that does not impose the Conditional Independence control function assumption utilized previously in the literature. Applying a similar framework as Kim and Petrin (2011), this paper uses the conditional mean restriction as an alternative approach to identification. Consequently proposing a new estimation procedure, gcfprobit, that is shown to be consistent and asymptotically normal even when traditional conditional independence assumptions do not hold. A major concern with the new approach is that is can be relatively inefficient compared to existing approaches. A possible avenue for further research is to develop a more efficient semi-parametric estimator possibly through regularization.

A Observational Equivalence with Endogeneity

Theorem 2.1 of Khan (2013) states that a heteroskedastic probit model with a non-parametric heteroskedasticity function is observationally equivalent to a "distribution free" model with only a conditional median restriction. The equivalence between the two models is in terms of the choice probabilities: $\mathrm{P}(y_{1i} = 1 \mid \mathbf{x}_i)$. This means both models will generate the same choice probability functions and therefore cannot be distinguished from one another on that basis. Khan (2013) only considers the case of no endogeneity – zero conditional mean. This extension shows observational equivalence for the case of a known non-zero conditional mean. This means the observational equivalence result holds when there is arbitrary endogeneity by incorporates the general control function. The following assumptions are variations on those presented in Khan (2013),

Assumption A.1 (General Conditional Median Restriction)**.** *In the set up described by equation (2.1),*

(i) $(v_{2i}, \boldsymbol{z}_i) \in \Re^{1+k_1+k_2}$ *has a density with respect to Lebesgue measure that is positive on the set* $(\mathcal{V} \times \mathcal{Z}) \subseteq \Re^{1+k_1+k_2}$.

(ii) *Let* $p_o(t, v_{2i}, \boldsymbol{z}_i)$ *denote* $\mathrm{P}(u_{1i} < t \mid v_{2i}, \boldsymbol{z}_i)$*, and assume*

(a) $p_o(\cdot, \cdot, \cdot)$ *is continuous on* $\Re \times (\mathcal{V} \times \mathcal{Z})$*,*

(b) $p_o'(t, v_{2i}, \boldsymbol{z}_i) = \partial p_o(t, v_{2i}, \boldsymbol{z}_i)/\partial t$ *exists and is continuous and positive on all* $\Re$ *for all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$*,*

(c) $p_o(h_o(v_{2i}, \boldsymbol{z}_i), v_{2i}, \boldsymbol{z}_i) = 1/2$ *where* $h_o(v_{2i}, \boldsymbol{z}_i)$ *is continuous on all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$*,*

(d) $\lim_{t \to -\infty} p_o(t, v_{2i}, \boldsymbol{z}_i) = 0$*,* $\lim_{t \to \infty} p_o(t, v_{2i}, \boldsymbol{z}_i) = 1$ *for all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$*.*

Assumption A.2 (Endogenous Heteroskedastic Probit)**.** *In the set up described by equation (2.1),*

(i) $(v_{2i}, \boldsymbol{z}_i) \in \Re^{1+k_1+k_2}$ *has a density with respect to Lebesgue measure that is positive on the set,*

(ii) $u_{1i} = \sigma_o(v_{2i}, \boldsymbol{z}_i) e_{1i} + h_o(v_{2i}, \boldsymbol{z}_i)$ *where* $\sigma_o(v_{2i}, \boldsymbol{z}_i)$ *is continuous and positive on* $(\mathcal{V} \times \mathcal{Z})$*,* $h_o(v_{2i}, \boldsymbol{z}_i)$ *that is continuous on all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$*, and* e_i *is independent of* $(v_{2i}, \boldsymbol{z}_i)$ *with any known (e.g. logistic, normal) distribution with median 0 and has a density function that is positive and continuous on the real line.*

Modifying the observational equivalence result to this setting is straightforward. Instead of focusing the model on a zero median restriction, it is acknowledged that the median is non-zero but a general control function, $h_o(v_{2i}, z_i)$, is specified.

Theorem A.1 (Observational Equivalence)**.** *The models described by Assumptions A.1 and A.2 are observationally equivalent in that,*

$$\mathrm{P}(y_{1i} = 1 \mid \boldsymbol{x}_i) = p_o(\boldsymbol{x}_i \beta_o, v_{2i}, \boldsymbol{z}_i) = \Phi\left(\frac{\boldsymbol{x}_i \beta_o + \boldsymbol{h}(v_{2i}, \boldsymbol{z}_i)\gamma_o}{\exp(\boldsymbol{g}(y_{2i}, \boldsymbol{z}_i))\delta_o}\right)$$

.

Proof. Write $u_{1i} = h_o(\mathbf{z}_i, v_{2i}) + \epsilon_i$ where $\mathrm{Med}(\epsilon_i \mid \mathbf{z}_i, v_{2i}) = 0$. Plugging into equation (2.1) and redefining: $\tilde{x}_i = (\mathbf{x}_i, h_o(\mathbf{z}_i, v_{2i}))$ and $\tilde{\beta}_o' = (\beta_o', 1)$, one can apply Theorem 2.1 of Khan (2013) to

$$y_i = 1\{\tilde{\mathbf{x}}_i \tilde{\beta}_o + \epsilon \geq 0\}$$

and obtain the observational equivalence result. □

B Proofs for Identification

The following lemma is an extension of corollary 1 given in Carlson (2019) to allow for multivariate **X** and **Z**.

Lemma B.1. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Z}$ *be vectors of random variables with continuous support. Suppose the following conditions hold*

(i) *$\boldsymbol{Z}$ does not contain a constant,*

(ii) $\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})$ *is non-singular,*

(iii) $\mathrm{E}(\boldsymbol{Z}'\boldsymbol{Z})$ *is non-singular,*

(iv) β_o *is non-zero,*

(v) *Each element of $\boldsymbol{Z}$ is a polynomial function of an element in $\boldsymbol{X}$, such that*

$$Z_j = X_k^{p_j^k}$$

where K is the dimension of $\boldsymbol{X}$ and $p_j^k \in \{1, 2, 3, ...\}$ is the order of the polynomial on the k^{th} term in $\boldsymbol{X}$ that composes the j^{th} term in $\boldsymbol{Z}$,

then for all parameters $(\beta, \delta) \in \Theta$ (the parameter space) if

$$\boldsymbol{X}(\beta_o - \exp(Z(\delta - \delta_o))\beta) = 0$$

with probability 1, then $(\beta, \delta) = (\beta_o, \delta_o)$.

Proof. Suppose there is a $(\beta, \delta) \in \Theta$, such that

$$\mathbf{X}(\beta_o - \exp(\mathbf{Z}(\delta - \delta_o))\beta) = 0$$

with probability 1. Then, because of condition (iv), I can rearrange

$$\mathbf{Z}(\delta - \delta_o) = \ln\left(\frac{\mathbf{X}\beta_o}{\mathbf{X}\beta}\right). \tag{B.1}$$

Let A denote the set of k such that the set $\{p_j^l : l = k\}$ is non-empty, then there exists a maximum polynomial order, $\tilde{p}_k = \max_j \{p_j^l : l = k\}$ for each $k \in A$. Then for each $k \in A$, take the partial derivative with respect to X_k, $\tilde{p}_k + 1$ times,

$$0 = (-1)^{\tilde{p}_k+1}\left[\left(\frac{\beta_{ko}}{\mathbf{X}\beta_o}\right)^{\tilde{p}_k+1} - \left(\frac{\beta_k}{\mathbf{X}\beta}\right)^{\tilde{p}_k+1}\right] \tag{B.2}$$

which implies $\frac{\beta_{ko}}{\mathbf{X}\beta_o} = \frac{\beta_k}{\mathbf{X}\beta}$. There are two cases: either for all k such that $\{p_j^l : l = k\}$ is not an empty set, $\beta_k = \beta_{ko} = 0$ or there exists at least one $\hat{k}$ such that $\beta_{\hat{k}} \neq 0$ and $\beta_{\hat{k}o} \neq 0$. In the first case, this reduces to the scenario that $\mathbf{X}$ and $\mathbf{Z}$ are not functionally related in which Theorem 1 of Carlson (2019) can be applied to obtain identification. In the second case, equation (B.2) implies $\frac{\beta_{ko}}{\beta_k} = \frac{\mathbf{X}\beta_o}{\mathbf{X}\beta}$ and plugging into equation (B.1),

$$\mathbf{Z}(\delta - \delta_o) = \ln\left(\frac{\beta_{ko}}{\beta_k}\right) \tag{B.3}$$

the right hand side is a constant. By conditions (i) and (iii) equation (B.3) can only hold if $\delta_o - \delta = 0$ and by condition (ii) this implies $\beta_o = \beta$. □

Proof of Theorem 4.1. Parts (i)-(iii) of Assumption 4.1 insure that

$$\mathrm{E}((\mathbf{x}_i, \boldsymbol{h}(v_{2i}, \mathbf{z}_i))'(\mathbf{x}_i, \boldsymbol{h}(v_{2i}, \mathbf{z}_i)))$$

is non-singular. To show, let $(a', b')'$ be a non-random vector such that

$$\begin{pmatrix} \mathbf{x}_i & \boldsymbol{h}(v_{2i}, \mathbf{z}_i) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \mathbf{x}_i a + \boldsymbol{h}(v_{2i}, \mathbf{z}_i) b = 0.$$

Taking the conditional expectation with respect to $\mathbf{z}_i$,

$$\mathrm{E}(\mathbf{x}_i \mid \mathbf{z}_i)a + \mathrm{E}(\boldsymbol{h}(v_{2i}, \mathbf{z}_i) \mid \mathbf{z}_i)b = \mathrm{E}(\mathbf{x}_i \mid \mathbf{z}_i)a = 0$$

where the first equality holds because of 4.1(ii)-(iii). Part (i) of Assumption 4.1 implies a is a zero vector and it follows that b is also a zero vector. Part (v) of Assumption 4.1 restricts how the heteroskedastic function may be specified to avoid issues non-identification due to the non-linear setting. Applying Lemma B.1 (which requires part 4.1(iv)), identification follows. □

C Proofs for Large Sample Properties

Let $||\cdot||$ denotes the euclidean norm, let $\bigtriangledown_x f(x)$ denote the $k \times 1$ gradient with respect to the $k \times 1$ vector x, and let $\bigtriangledown_x^2 f(x)$ denote the $k \times k$ Hessian with respect to the $k \times 1$ vector x.

Proof of Theorem 5.1. First, the following sample and population objective functions are defined respectively for the second stage estimator,

$$\begin{aligned} Q_n(\pi, \theta) &= n^{-1} L_n(\pi, \beta, \gamma, \delta), \\ Q_o(\pi, \theta) &= \mathrm{E}\left(y_{1i} \log\left[\Phi\left(\frac{\mathbf{x}_i\beta + \boldsymbol{h}(y_{2i} - m(\mathbf{z}_i, \pi), \mathbf{z}_i)\gamma}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i)\delta)}\right)\right]\right. \\ &\qquad \left. + (1 - y_{1i}) \log\left[1 - \Phi\left(\frac{\mathbf{x}_i\beta + \boldsymbol{h}(y_{2i} - m(\mathbf{z}_i, \pi), \mathbf{z}_i)\gamma}{\exp(\boldsymbol{g}(y_{2i}, \mathbf{z}_i)\delta)}\right)\right]\right). \end{aligned}$$

Using Theorem 2.1 of Newey and McFadden (1994) I need to show the following,

(i) Identification: $Q_o(\pi_o, \theta)$ is uniquely maximized at θ_o,

(ii) Compactness: Θ is compact,

(iii) Continuity: $Q_o(\pi, \theta)$ is continuous in π and θ,

(iv) Uniform Convergence: $Q_n(\hat{\pi}, \theta)$ converges uniformly in probability to $Q_o(\pi_o, \theta)$.

Identification, part (i), holds under Assumption 4.1. Part (ii) is assumed. Part (iii) is evident given the heteroskedastic probit specification in Assumption 3.1 and the continuity of the control function and first stage conditional mean. Part (iv) follows from Lemma 2.4 of Newey and McFadden (1994) and the triangle inequality:

$$\begin{aligned}
&\sup_{\theta\in\Theta}||Q_n(\hat{m},\theta)-Q_o(m_o,\theta)|| \\
&\qquad=\sup_{\theta\in\Theta}||Q_n(\hat{m},\theta)-Q_n(m_o,\theta)+Q_n(m_o,\theta)-Q_o(m_o,\theta)|| \\
&\qquad\leq\sup_{\theta\in\Theta}||Q_n(\hat{m},\theta)-Q_n(m_o,\theta)||+\sup_{\theta\in\Theta}||Q_n(m_o,\theta)-Q_o(m_o,\theta)||
\end{aligned}$$

where the first term is $o_p(1)$ by consistency of the first stage estimator and the Continuous Mapping Theorem (again requiring continuity of the control function and first stage conditional mean). Given the finite second moment conditions given in Assumption 4.1, the dominance condition in Lemma 2.4 easily holds because both y_{1i} and $\Phi(\cdot)$ are bounded. Therefore Lemma 2.4 establishes $\sup_{\theta\in\Theta}||Q_n(m_o,\theta)-Q_o(m_o,\theta)||=o_p(1)$ so uniform convergence hold and the conditions of Theorem 2.1 are satisfied. □

Proof of Theorem 5.2. Let $\mathbf{w}_i=(y_{1i},y_{2i},\mathbf{z}_i)$ with support $\mathcal{W}$ and

$$\begin{aligned}
S(\mathbf{w}_i,\pi,\theta)&=C(\mathbf{w}_i,\pi,\theta)H(\mathbf{w}_i,\pi,\theta),\\
C(\mathbf{w}_i,\pi,\theta)&=\frac{(y_{1i}-P(\boldsymbol{w}_i,\pi,\theta))\,p(\boldsymbol{w}_i,\pi,\theta)}{P(\boldsymbol{w}_i,\pi,\theta)\,(1-P(\boldsymbol{w}_i,\pi,\theta))\exp(\boldsymbol{g}(y_{2i}),\mathbf{z}_i)\delta)},\\
P(\boldsymbol{w}_i,\pi,\theta)&=\Phi\left(\frac{x_i\beta+\boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi),z_i)\gamma}{\exp(\boldsymbol{g}(y_{2i},\mathbf{z}_i)\delta)}\right),\\
p(\boldsymbol{w}_i,\pi,\theta)&=\phi\left(\frac{x_i\beta+\boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi),z_i)\gamma}{\exp(\boldsymbol{g}(y_{2i},\mathbf{z}_i)\delta)}\right),\\
H(\mathbf{w}_i,\pi,\theta)&=\begin{pmatrix}\mathbf{x}_i'\\ \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi),\mathbf{z}_i)'\\ -(\mathbf{x}_i\beta+\boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi),\mathbf{z}_i))\boldsymbol{g}(y_{2i},\mathbf{z}_i)'\end{pmatrix}.
\end{aligned}$$

Proving asymptotic normality for the two-step estimator will be done by showing asymptotic normality for the GMM estimator in which the first and second stage are stacked. Let $M(\mathbf{w}_i,\pi,\theta)$ denote the following stacked moment equations:

$$M(\mathbf{w}_i,\pi,\theta)=\begin{pmatrix}(y_{2i}-m(\mathbf{z}_i,\pi))\bigtriangledown_\pi m(\mathbf{z}_i,\pi)\\ S(\mathbf{w}_i,\pi,\theta)\end{pmatrix}.$$

Then using Theorem 6.1 of Newey and McFadden (1994), I need to show the following:

(i) $\pi_o\in\text{int}(\Pi)$ and $\theta_o\in\text{int}(\Theta)$, both of which are compact,

(ii) $M(\mathbf{w}_i,\pi,\theta)$ is continuously differentiable in a neighborhood of (π_o,θ_o) with probability approaching one,

(iii) $\text{E}(M(\mathbf{w}_i,\pi_o,\theta_o))=0$,

(iv) $\text{E}(||M(\mathbf{w}_i,\pi_o,\theta_o)||^2)$ is finite,

(v) $\mathrm{E}(\sup_{(\pi,\theta)\in\Pi\times\Theta} || \bigtriangledown_{(\pi',\theta')'} M(\mathbf{w}_i,\pi,\theta)||) < \infty$,

(vi) $G'G$ is non-singular,

where $G = \mathrm{E}(\bigtriangledown_{(\pi',\theta')'} M(\mathbf{w}_i,\pi_o,\theta_o))$. Part (i) is assumed and (ii) is evident given the nonlinear LS and probit specifications. Part (iii) holds by Assumption 3.1 (correct conditional mean specification in the first stage and Fischer consistency in the second stage). Part (iv) can be verified:

$$\begin{aligned}||M(\mathbf{w}_i,\pi_o,\theta_o)||^2 =& ||(y_{2i}-m(\mathbf{z}_i,\pi_o)) \bigtriangledown_\pi m(\mathbf{z}_i,\pi_o)||^2 + ||S(\mathbf{w}_i,\pi_o,\theta_o)||^2 \\ =& ||(y_{2i}-m(\mathbf{z}_i,\pi_o)) \bigtriangledown_\pi m(\mathbf{z}_i,\pi_o)||^2 \\ &+ C(\mathbf{w}_i,\pi_o,\theta_o)^2\Big(||\mathbf{x}_i||^2 + ||\boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)||^2 \\ &+ ||(\mathbf{x}_i\beta_o + \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi)_o,\mathbf{z}_i)\gamma_o)||^2||\boldsymbol{g}(y_{2i},\mathbf{z}_i)||^2\Big).\end{aligned}$$

Taking the conditional expectation,

$$\begin{aligned}&E\Big(||M(\mathbf{w}_i,\pi_o,\theta_o)||^2 \mid \mathbf{z}_i, y_{2i}\Big) \\ &=||(y_{2i}-m(\mathbf{z}_i,\pi_o)) \bigtriangledown_\pi m(\mathbf{z}_i,\pi_o)||^2 + \mathrm{E}(C(\mathbf{w}_i,\pi_o,\theta_o)^2 \mid \mathbf{z}_i, y_{2i}) \\ &\quad\times \Big(||\mathbf{x}_i||^2 + ||\boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)||^2 \\ &\quad+ ||(\mathbf{x}_i\beta_o + \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi)_o,\mathbf{z}_i)\gamma_o)||^2||\boldsymbol{g}(y_{2i},\mathbf{z}_i)||^2\Big) \\ &=||(y_{2i}-m(\mathbf{z}_i,\pi_o)) \bigtriangledown_\pi m(\mathbf{z}_i,\pi_o)||^2 \\ &\quad+ \left(\frac{\lambda(\mathbf{w}_i,\pi_o,\beta_o,\gamma_o,\delta_o)\lambda(\mathbf{w}_i,\pi_o,-\beta_o,-\gamma_o,\delta_o)}{\exp(2\boldsymbol{g}(y_{2i},\mathbf{z}_i)\delta_o)}\right) \\ &\quad\times \Big(||\mathbf{x}_i||^2 + ||\boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)||^2 \\ &\quad+ ||(\mathbf{x}_i\beta_o + \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)\gamma_o)||^2||\boldsymbol{g}(y_{2i},\mathbf{z}_i)||^2\Big)\end{aligned}$$

where $\lambda(\mathbf{w}_i,\pi_o,\beta_o,\gamma_o,\delta_o)$ is the inverse mills ratio. Since $\lambda_i(\pi_o,\beta_o,\gamma_o,\delta_o)\lambda_i(\pi_o,-\beta_o,-\gamma_o,\delta_o)$ is bounded (and bounded away from 0), $\mathrm{E}(||M(y_{1i},y_{2i},\mathbf{z}_i;\pi_o,\theta_o)||^2)$ is finite as long as v_{2i}, $\mathbf{x}_i$, $\mathbf{z}_i$, $\boldsymbol{h}(v_{2i}\mathbf{z}_i)$, $\boldsymbol{g}(y_{2i},\mathbf{z}_i)$, and $\bigtriangledown_\pi m(\mathbf{z}_i,\pi_o)$ have finite second moments (which is assumed). Part (v) follows from boundedness of the first derivative of the inverse mills ratio and the assumptions in the theorem. To show (vi), let $G=(G_\pi,G_\theta)$ where

$$\begin{aligned}G_\pi &= \begin{pmatrix} G_{1\pi} \\ G_{2\pi}\end{pmatrix} = \begin{pmatrix}\mathrm{E}(\bigtriangledown_\pi m(\mathbf{z}_i,\pi_o)(\bigtriangledown_\pi m(\mathbf{z}_i,\pi_o))') \\ \mathrm{E}(\bigtriangledown_\pi S(\mathbf{w}_i,\pi_o,\theta_o))\end{pmatrix}, \\ G_\theta &= \begin{pmatrix} G_{1\theta} \\ G_{2\theta}\end{pmatrix} = \begin{pmatrix}\mathbf{0} \\ \mathrm{E}(\bigtriangledown_\theta S(\mathbf{w}_i,\pi_o,\theta_o))\end{pmatrix},\end{aligned}$$

and

$$
\begin{aligned}
\mathrm{E}(\bigtriangledown_{\pi} S(\mathbf{w}_i,\pi_o,\theta_o)) &= \mathrm{E}(\Lambda_i \bigtriangledown_{v_{2i}} \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)\gamma_o(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o))'),\\
\mathrm{E}(\bigtriangledown_{\theta} S(\mathbf{w}_i,\pi_o,\theta_o)) &= \mathrm{E}(\Lambda_i H(\mathbf{w}_i,\pi,\theta)(H(\mathbf{w}_i,\pi,\theta))'),\\
\Lambda_i &= \frac{\lambda(\mathbf{w}_i,\pi_o,\beta_o,\gamma_o,\delta_o)\lambda(\mathbf{w}_i,\pi_o,-\beta_o,-\gamma_o,\delta_o)}{\exp(2\boldsymbol{g}(y_{2i},\mathbf{z}_i)\delta_o)}.
\end{aligned}
$$

Then

$$
G'G = \begin{pmatrix} G'_{\pi}G_{\pi} & G'_{\pi}G_{\theta} \\ G'_{\theta}G_{\pi} & G'_{\theta}G_{\theta} \end{pmatrix},
$$

where

$$
\begin{aligned}
G'_{\pi}G_{\pi} =& \,\mathrm{E}(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o)(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o))')\,\mathrm{E}(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o)(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o))')\\
&+\mathrm{E}(\Lambda_i \bigtriangledown_{v_{2i}} \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)\gamma_o(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o)))\\
&\times \mathrm{E}(\Lambda_i \bigtriangledown_{v_{2i}} \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)\gamma_o(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o))'),\\
G'_{\pi}G_{\theta} =& \,\mathrm{E}(\Lambda_i \bigtriangledown_{v_{2i}} \boldsymbol{h}(y_{2i}-m(\mathbf{z}_i,\pi_o),\mathbf{z}_i)\gamma_o(\bigtriangledown_{\pi} m(\mathbf{z}_i,\pi_o)))\\
&\times \mathrm{E}(\Lambda_i H(\mathbf{w}_i,\pi,\theta)(H(\mathbf{w}_i,\pi,\theta))'),\\
G_{\theta}G_{\theta} =& \,\mathrm{E}(\Lambda_i H(\mathbf{w}_i,\pi,\theta)(H(\mathbf{w}_i,\pi,\theta))')\,\mathrm{E}(\Lambda_i H(\mathbf{w}_i,\pi,\theta)(H(\mathbf{w}_i,\pi,\theta))'),
\end{aligned}
$$

which is non-singular by Theorem 5.2(iii) and Assumption 4.1(ii). □

Acknowledgements

This paper is based on a chapter of my PhD thesis at Michigan State University. I am grateful to Jeffrey Wooldridge, Kyoo il Kim, Joe Herriges, David Kaplan, Ben Zou, Tim Vogelsang, Mike Conlin, and Nicole Mason for all their direction and advice. I would also like to thank Xiaoxia Shi, participants of the Women's Mentoring Workshop at MEG 2018, conference attendees at MEG 2018 and IAAEC 2018.

References

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

Blundell, R. and R. L. Matzkin (2014): "Control functions in nonseparable simultaneous equations models," *Quantitative Economics*, 5, 271–295.

Blundell, R. and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *Econometric society monographs*, 36, 312–357.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71, 655–679.

Carlson, A. (2019): "Parametric identification of multiplicative exponential heteroscedasticity," *Oxford Bulletin of Economics and Statistics*, 81, 686–696.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014): "Local identification of nonparametric and semiparametric models," *Econometrica*, 82, 785–809.

D'Haultfœuille, X. and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," *Econometrica*, 83, 1199–1210.

Dong, Y. and A. Lewbel (2015): "A simple estimator for binary choice models with endogenous regressors," *Econometric Reviews*, 34, 82–105.

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016): "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7, 561–589.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191–1206.

Gandhi, A., K. i. Kim, and A. Petrin (2011): "Identification and estimation in discrete choice demand models when endogenous variables interact with the error," NBER Working Papers 16894, National Bureau of Economic Research, Inc.

Hahn, J. and G. Ridder (2011): "Conditional moment restrictions and triangular simultaneous equations," *Review of Economics and Statistics*, 93, 683–689.

Hall, P. and J. L. Horowitz (2005): "Nonparametric methods for inference in the presence of instrumental variables," *The Annals of Statistics*, 33, 2904–2929.

Hong, H. and E. Tamer (2003): "Endogenous binary choice model with median restrictions," *Economics Letters*, 80, 219–225.

Horowitz, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 505–531.

Imbens, G. W. and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

Kasy, M. (2011): "Identification in triangular systems using control functions," *Econometric Theory*, 27, 663–671.

Khan, S. (2013): "Distribution free estimation of heteroskedastic binary response models using probit/logit criterion functions," *Journal of Econometrics*, 172, 168–182.

Kim, K. i. and A. Petrin (2011): "A new control function approach for non-parametric regressions with endogenous variables," NBER Working Papers 16679, National Bureau of Economic Research, Inc.

Krief, J. M. (2014): "An integrated kernel-weighted smoothed maximum score estimator for the partially linear binary response model," *Econometric Theory*, 30, 647–675.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97, 145–177.

Lin, W. and J. M. Wooldridge (2015): "On different approaches to obtaining partial effects in binary response models with endogenous regressors," *Economics Letters*, 134, 58 – 61.

Manski, C. F. (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 313–333.

——— (1988): "Identification of binary response models," *Journal of the American Statistical Association*, 83, 729–738.

Newey, W. K. (2013): "Nonparametric instrumental variables estimation," *American Economic Review*, 103, 550–56.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Newey, W. K. and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67, 565–603.

Petrin, A. and K. Train (2010): "A control function approach to endogeneity in consumer choice models," *Journal of Marketing Research*, 47, 3–13.

Pinkse, J. (2000): "Nonparametric two-step regression estimation when regressors and error are dependent," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28, 289–300.

Rivers, D. and Q. H. Vuong (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39, 347 – 366.

Rothe, C. (2009): "Semiparametric estimation of binary response models with endogenous regressors," *Journal of Econometrics*, 153, 51–64.

Rothenberg, T. J. (1971): "Identification in parametric models," *Econometrica*, 577–591.

Smith, R. J. and R. W. Blundell (1986): "An exogeneity test for a simultaneous equation tobit model with an application to labor supply," *Econometrica*, 679–685.

Su, L. and A. Ullah (2008): "Local polynomial estimation of nonparametric simultaneous equations models," *Journal of Econometrics*, 144, 193–218.

Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," *Econometrica*, 83, 1185–1197.

Wooldridge, J. M. (2005): "Unobserved heterogeneity and estimation of average partial effects," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 27–55.

——— (2010): *Econometric analysis of cross section and panel data*, MIT press.