Correcting for Sample Selection Bias in Dyadic Regressions

An Application to Gravity Models

written by

Gabriela Mara Miyazato Szini

Tinbergen Institute

University of Amsterdam

supervised by

Prof. Dr. Frank Kleibergen

University of Amsterdam

Thesis submitted in partial fulfilment of the requirements for the Degree of

Master of Philosophy in Economics (research)

Amsterdam, August 2020

Acknowledgements

To the memory of István Szini

Abstract

We present methods to correct for sample selection bias in the estimation of dyadic regressions. Dyadic datasets can be seen as a pseudo panel data, where both dimensions tend to infinity as the number of individuals grows. We show that including fixed effects for both individuals forming a dyad leads to asymptotically biased estimates of the structural parameters in the first stage of the Heckman (1979) two step method. This is a consequence of the incidental parameter problem. We reconcile and modify existing approaches to similar problems in standard panel data models to this framework. Our Monte Carlo simulation exercise corroborates to the theoretical predictions that the standard Heckman approach yields biased estimates, while the proposed methods reduce such biases. We apply the proposed estimators to the gravity model for international trade flows. The suggested methods deliver different estimates for the coefficients of trade barriers.

Contents

1 Introduction

Tinbergen (1962) established gravity equations, which have been widely used for estimating models of international trade, migration, equity and FDI flows. For instance, in the literature of international trade flows, it is used to infer the effects of institutions such as customs unions, exchange rate mechanisms, ethnic ties, linguistic identity and international boarders on trade flows.

Even decades after its first appearance in trade models, there is still a substantial lack of understanding on features of its estimation, and econometric methods that considers some potential issues only emerged recently in the literature. The aim of this paper is to provide an ample and deep discussion about different estimation issues in gravity models focusing on its dyadic structure, highlighting and describing in detail proposed estimators that take it into account.

A great deal of applications rely on the specification of the gravity model proposed by Helpman et al. (2008), due to its tractability and the fact that it demonstrates explicitly how the sample selection problem plays a role in the macroeconomic model. Up to this seminal paper, most studies only considered observations with positive trade flows when estimating such model without imposing any special treatment on it. However, the authors point out that one should account for the self-selection of firms into export markets and their impact on trade volumes. In other words, one should account for sample selection in the estimation. To do so, the authors first relax some assumptions in the theoretical foundation of the gravity model proposed by Anderson and Van Wincoop (2003), then they propose a parametric estimation of a linear model controlling for the sample selection through the inverse Mills-Ratio approach (Heckman (1979)).

However, we show that this methodology suffers from the incidental parameter problem, Neyman and Scott (1948). This well-known problem relates to the fact that, as demonstrated by Fernández-Val and Weidner (2016), the fixed effects estimates in nonlinear models are asymptotically biased, even when both dimensions of a panel dataset tend to infinity. The model proposed by Helpman et al. (2008) undergoes such issue, once it accounts for fixed effects for both importer and exporter, which are known in the literature as the multilateral resistance terms.

In order to analyze such problems, and discuss possible solutions to it, we further simplify the estimated equations by Helpman et al. (2008). This simplification highlights the fact that such a framework is applicable to any dyadic data. As defined by Graham (2020), dyadic data reflects situations where the outcome of interest is determined by pairwise interactions between units. More specifically, the first stage of the estimation - that relies on estimating the probabilities that two units (nodes) interact with each other (that countries trade in the trade application) - delineates a specifi-

3

cation for network formation. One particularity of dyadic settings is that as the number of individuals increase, both dimensions of the pseudo panel increase.

The introduction of two-way fixed effects requires that in the first stage of the approach, a more involved estimator needs to be employed to difference out such fixed effects, delivering asymptotically unbiased and consistent estimates for the structural parameters. A clever approach is proposed by Charbonneau (2017) and employed in this study. This approach is based on a conditional likelihood estimator.

Once such estimates are obtained, the standard Heckman approach for correcting for sample selection in the observation equation requires that (i) estimates of the fixed effects itself in the first stage (selection equation) are provided such that the predicted probabilities are obtained; and (ii) a functional form for the inverse Mills-ratio is provided, which generally follows from both errors being normally distributed. The Charbonneau (2017) estimation results in additional problems related to both items: it does not deliver estimates for the fixed effects itself, and it relies on the assumption that the errors of the selection equation are logistically distributed.

We propose two different approaches to tackle such obstacles. The first is to retrive the estimates of the fixed effects through the traditional unconstrained MLE, by setting the structural parameters to be equal to the estimates provided by Charbonneau (2017). This method is denoted by the hybrid approach by Wooldridge et al. (2018). Once the fixed effects estimates are obtained, one can then calculate the predicted probabilities. Moreover, a transformation of the variables proposed by Lee (1983) guarantees that the traditional two-step approach given by Heckman (1979) can be employed, even if the error term of the selection equation is logistically distributed.

A second approach is developed by Kyriazidou (1997). It relies on a weighted least squares estimator based on the idea of differencing out the sample selection effects in the observation equation (the second stage equation of Heckman (1979)), and, in the trade application, in the estimates for the volumes traded amongst countries. The advantage of this method is that, as we will demonstrate later, it does not require estimates of the fixed effects in the first stage equation. While the original framework in Kyriazidou (1997) is a standard panel data model, where equations are differenced over the time dimension, we propose modifications of this approach to accomodate dyadic interactions, where equations are differenced over combinations of dyads. Such modification delivers an approach that can be employed even when the exogenous variables of the observation equation are invariant over time, varying only over dyads.

Even though the proposed approaches already exist in the literature separately (apart from our modification to the estimator proposed by Kyriazidou (1997)), this study adds to the literature in that it combines those

4

separate methods to account for sample selection in dyadic structures. Up to our knowledge, although the literature for sample selection corrections is abundant for standard panel data models, it is insufficient for dyadic datasets, and consequently for cases related to network formation.

The structure of this paper is as follows: Section 2 presents the baseline gravity model by Helpman et al. (2008), Section 3 presents the estimation strategy employed by the authors, Section 4 delineates our simplification of the model, highlighting its dyadic structure and its aspects related to network formation, Section 5 provides the standard two-stage Heckman approach for this model, Section 6 refers to a more formalized discussion of the incidental parameters problem, Section 7 discusses the approach by Charbonneau (2017), Section 8 shows our first proposed approach to correct for sample selectivity in the observation equation, Section 9 shows the second proposed approach, Section 10 provides our Monte Carlo simulations results for the standard Heckman approach and the two new methods proposed for a set of different designs, Section 11 provides an application to the estimation of the gravity model for international trade flows, and Section 12 concludes.

2   The gravity model by Helpman et al. (2008)

One of the baseline models for the gravity equation is the model derived by Anderson and Van Wincoop (2003), as this was one of the first studies to apply the theory of the gravity equations seriously to international trade flows. Specifically, the way the constant elasticity of substitution (CES) expenditure system is manipulated in the model allows for considering multilateral resistance terms for both countries involved in the trade flow in a tractable way. As Anderson and Van Wincoop (2003) explains, these terms are defined as a country's average trade barrier with all its trading partners, being invariant to the country. It is expected that in a pairwise trading relationship the multilateral resistance terms of both involved countries affect the outcome. This follows from, after controlling for the size of the economy of the countries, the more resistant a country is to trade with all other trading partners, the more the country is likely to trade with a given bilateral partner. Thus, bilateral trade is related to size, bilateral trade barriers and the multilateral resistance terms of both countries.

However, in this study, we use as a baseline the model specified by Helpman et al. (2008). Both models take into account the equation for trade flows for a country i that exports a positive quantity for a given country j. Therefore, observations related to pairs of countries that do not trade amongst themselves are not taken into account in the estimation of the equation for trade flows, which generates a sample selection problem. One advantage of the paper by Helpman et al. (2008) is that it provides a sound theoretical

framework that takes into account sample selection and also accommodates asymmetric trade flows whereas the model by Anderson and Van Wincoop (2003) does not.

The sample selection is accounted for by considering that it is likely that only a fraction of firms in a country i decides to export to country j (such fraction is allowed to be zero), and that those firms have individual heterogeneous productivity and face both variable and fixed costs of exportation. Therefore, the model is able to predict one of the stylized features of the data, which is that a number of pair of countries displays zero trade flows. Also, it allows for the decomposition of the impact of trade frictions into intensive and extensive margins (the first refers to the trade volume per exporter and the latter to the number of exporters).

We consider that a country i has a measure N_i of heterogeneous firms that differ in terms of productivity, which is measured by $\frac{1}{a}$, where a follows a cumulative distribution function, $G(a)$, with support $[a_L, a_H]$. In this model only the more productive firms decide to export, and the fraction of firms exporting is given by a zero-profit condition - this profitability varies by destination, according to the demand levels of the importing countries, the variable and the fixed costs.

It is further assumed that a firm in country i produces one unit of output with inputs that cost $c_i a$, where a is defined as above and measures the number of bundles of inputs used for production of a unit of output, and c_i measures the cost of the bundle (which is country-specific). If a producer sells its product in country j, then it faces additional costs. Those costs are split into a fixed cost, $c_i f_{ij}$, and a variable cost that takes the form of a "melting iceberg" specification [1], assuming that τ_{ij} units of a product needs to be shipped from country i to j for a unit to arrive. [2]

By also assuming CES preferences, that products are differentiated according to their country of origin, and that there is monopolistic competition in the final products, the model delivers the following system of equations:

$$(1 - \alpha) \left(\frac{\tau_{ij} c_i a_{ij}}{\alpha P_j} \right)^{1-\sigma} Y_j = c_i f_{ij} \tag{1}$$

This Equation (1) comes from the above mentioned zero-profit condition, where a_{ij} is defined such that at this point profits are exactly zero. We denote by P_j the price index of the economy of country j, and Y_j the size of its economy. This equation determines the fraction of country i's N_i firms that export to country j, given by $G(a_{ij})$. The fraction of firms exporting can be zero once $a_{ij} \leq a_L$. Note that here we denote that $\sigma = 1/(1 - \alpha)$ is

---

[1]Which, as defined by Krugman (1991), is modeled according to the fact that for each unit of goods shipped from one region to the other, only a fraction of it arrives.

[2]Note that those additional costs are country-specific but not firm-specific, nor depending on the firm productivity level.

the elasticity of substitution across products, that remains the same across countries (as in Anderson (1979)).

$$V_{ij} = \begin{cases} \int_{a_L}^{a_{ij}} a^{1-\sigma} dG(a) & \text{for } a_{ij} \geq a_L \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Where V_{ij} determines the bilateral trade volume.

$$Y_{1,ij} = \left(\frac{c_i \tau_{ij}}{\alpha P_j} \right)^{1-\sigma} Y_j N_i V_{ij} \tag{3}$$

Equation (3) defines the value of country j's imports from i ($Y_{1,ij}$). It can thus be defined as the gravity equation obtained from this model. One can note that when $a_{ij} \leq a_L$, V_{ij} will be equal to zero, and so will $Y_{1,ij}$. Moreover, from Equation (3) it is clear how the fraction of firms exporting plays a role in the trade flows defined by $Y_{1,ij}$.

$$P_j^{1-\sigma} = \sum_{i=1}^{I} \left(\frac{c_i \tau_{ij}}{\alpha} \right)^{1-\sigma} N_i V_{ij} \tag{4}$$

Finally, this Equation (4) defines the price indices of country j. One can note from these equations that not only trade flows can be explained, but also asymmetries can be explained, as $Y_{1,ij}$ can be different from $Y_{1,ji}$.

The derivations of this model can be found in the Appendix A.1. It is also shown that this model can be used to derive the model given by Anderson and Van Wincoop (2003) (which does not account for sample selection and asymmetric trade flows) under the assumptions of symmetric variable costs ($\tau_{ij} = \tau_{ji}$) and that V_{ij} can be decomposed into a deterministic function of terms that depend only on the importer, the exporter, and country-pair characteristics. For more details, I provide the equivalence between both models also in Appendix A.1.

3   Estimation method of Helpman et al. (2008)

Given the theoretical gravity model defined in the previous section, we now focus on the assumptions imposed by Helpman et al. (2008) on some variables in order to further pin down the model to be estimated.

The aim of the paper by Helpman et al. (2008) is to estimate Equation (3), the gravity equation, subject to countries that self-select into trading (i.e., the trading decision that introduces a sample selection). The estimated model is then used to understand the magnitude and to make inference on the coefficients of trade barriers.

First, the authors assume, as mentioned before:

ASSUMPTION 1. The firm productivity $1/a$ is Pareto distributed, with support $[a_L, a_H]$. We assume then that the distribution function of a follows $G(a) = (a^k - a_L^k)/(a_H^k - a_L^k)$, with $k > (\sigma - 1)$.

This framework allows for asymmetric trade flows $Y_{1,ij} \neq Y_{1,ji}$, once we can have that $a_{ij} < a_L$ for some pairs ij, while $a_{ji} > a_L$, leading to zero exports from i to j, but not the other way round. Therefore, if the productivity is drawn from a truncated Pareto distribution, asymmetric trade frictions are not necessary to generate asymmetric trade flows.

Assumption 1 implies that V_{ij} can be expressed as a function of a new variable W_{ij} defined below. Moreover, both variables are monotonic functions of the proportion of exporters from i to j, $G(a_{ij})$, once all parameters in those equations are fixed, with only a_{ij} varying:

$$V_{ij} = \frac{k a_L^{k-\sigma+1}}{(k - \sigma + 1)(a_H^k - a_L^k)} W_{ij}$$

$$W_{ij} = \max\left(\left(\frac{a_{ij}}{a_L}\right)^{k-\sigma+1} - 1, 0\right)$$

From now on, we will allow the model to be extended to several periods $t = 1, ...T$, since trade datasets are available for several periods. Two additional assumptions are imposed on the structure of the variable and fixed trade costs, where the first affects the volume of firm-level exports and the second the decision to trade:

ASSUMPTION 2. There are i.i.d. unmeasured country-pair specific trade frictions $u_{ij,t} \sim N(0, \sigma_u^2)$. These affect the variable trade costs $\tau_{ij,t}$, since this variable takes the form:
$$\tau_{ij,t}^{\sigma-1} \equiv D_{ij,t}^{\gamma} e^{-u_{ij,t}}$$
where $D_{ij,t}$ is the symmetric distance between i and j.

ASSUMPTION 3. There are i.i.d. unmeasured country-pair specific trade fictions $\nu_{ij,t} \sim N(0, \sigma_\nu^2)$ that may be correlated with $u_{ij,t}$. These affect only the fixed export costs $f_{ij,t}$, given by:

$$f_{ij,t} \equiv \exp\left(\phi_{EX,i} + \phi_{IM,j} + \kappa \phi_{ij,t} - \nu_{ij,t}\right)$$

where $\phi_{IM,j}$ is a fixed trade barrier imposed by the importing country on all exporters, $\phi_{EX,i}$ is a measure of fixed export costs common across all export destinations, and $\phi_{ij,t}$ is an observed measure of any additional country-pair specific fixed trade costs.

8

Given Assumption 2, and by log-linearizing Equation (3), which defines the export volume from country i to j, the authors arrive at the following equation to be estimated:

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + w_{ij,t} + u_{ij,t}, \tag{5}$$

where: (i) $y_{1,ij,t} = \ln Y_{1,ij,t}$; (ii) $w_{ij,t} = \ln W_{ij,t}$; (iii) β_1 is a vector that collects the coefficients of the remaining structural parameters, $\beta_1 = (\beta_0, \gamma'_1)'$; (iv) $x_{1,ij,t}$ is a vector that collects 1 and the vector $d_{ij,t} = \ln D_{ij,t}$; (v) $\chi_j = (\sigma - 1)\ln P_j + \ln Y_j$, which is an importer fixed effects; and finally, (vi) $\vartheta_i = -(\sigma - 1)\ln c_i + \ln N_i$, which is an exporter fixed effects.

One of the most important differences to the equation derived in previous studies, for instance, in the model of Anderson and Van Wincoop (2003), is the new variable $w_{ij,t}$, which controls for the fraction of firms exporting. If such term is not included in the RHS, the coefficients on any potential trade barrier can no longer be interpreted as the elasticity of a firm's trade with respect to distance, once trade barriers affect both $y_{1,ij,t}$ and $w_{ij,t}$ (through the proportion of exporters from i to j), leading to an omitted variable bias, as the latter is also correlated with $y_{1,ij,t}$.

Another bias in the estimation of (5) arises when country pairs of zero trade flows are excluded, yielding a sample selection bias. From a macroeconomic modelling perspective, such bias is due to the fact that country pairs with large observed trade barriers (high $d_{ij,t}$) that trade with each other are likely to have low unobserved trade barriers (high $u_{ij,t}$), otherwise it would not be profitable to trade.

Therefore, another important difference in comparison to previous studies, is that such selection bias is taken into account. To do so, a latent variable $Y^*_{2,ij,t}$ is defined, which is given by the ratio of variable export profits for the most productive firm (with productivity $1/a_L$) to the fixed export costs for exports from i to j (expressions that are given by the zero profit condition). Then, in this case, positive exports are observed if $Y^*_{2,ij,t} > 1$, and one can verify in the Appendix A.1 the expressions for $Y^*_{2,ij,t}$ and that $W_{ij,t}$ is a monotonic function of $Y^*_{2,ij,t}$: in the case of positive exports, $W_{ij,t} = Y^{*(k-\sigma+1)/(\sigma-1)}_{2,ij,t} - 1$.

By log-linearizing the equation for the latent variable $Y^*_{2,ij,t}$, and given Assumption 3, we have that:

$$y^*_{2,ij,t} = x'_{2,ij,t}\beta_2 + \xi_i + \zeta_j + \eta_{ij,t}, \tag{6}$$

where: (i) $\eta_{ij,t} \equiv u_{ij,t} + \nu_{ij,t} \sim N(0, \sigma_u^2 + \sigma_\nu^2)$ is i.i.d., but correlated with the error term $u_{ij,t}$ in Equation (5); (ii) $\xi_i = -\sigma \ln c_i + \phi_{EX,i}$ is an exporter fixed effect; (iii) $\zeta_j = (\sigma - 1)\ln P_j + \ln Y_j - \phi_{IM,j}$ is an importer fixed effect; (iv) β_2 is a vector that collects the coefficients of the remaining structural parameters, $\beta_2 = (\gamma_0, \gamma'_2)'$; (iv) $x_{2,ij,t}$ is a vector that collects 1, and the vectors $d_{ij,t} = \ln D_{ij,t}$, and $\phi_{ij,t}$. Note that $x_{2,ij,t}$ includes both the regressors

in $x_{1,ij,t}$ and additional regressors that affect only the decision to export, but not the volume of exports (satisfying then an exclusion restriction).

From an econometric perspective, the fact that $u_{ij,t}$ and $\eta_{ij,t}$ are correlated indicates that the data on $x_{1,ij,t}$ is not randomly missing, which points out to a sample selection problem. The Appendix A.2 demonstrates how the bias in the estimators of Equation (5) arises in the form of an omitted variable bias when this sample selection is not accounted for. The intuition for the existence of this bias is that the estimates for the effects of the trade barriers based on the sample of countries that do trade amongst themselves do not deliver a reliable estimate of these effects for countries that do not trade, had they traded. More specifically, the fitted regressions confound the behavior of such parameters for the volume of trade with the parameters of the equation determining the probability of two countries trading.

In order to estimate the coefficients on trade barriers in Equation (5) in an unbiased manner, Helpman et al. (2008) follow the procedure proposed by Heckman (1979). The first step of such procedure is to obtain estimates of the selection equation, given in this case by Equation (6).

Although $y_{2,ij,t}^*$ is unobserved, we observe whether countries decide or not to trade, and we know that $Y_{2,ij,t}^* > 1$ and, therefore, $ln(Y_{2,ij,t}^*) = y_{2,ij,t}^* > 0$ when i exports to j (setting an indicator variable $y_{2,ij,t} = 1$), and $y_{2,ij,t}^* \leq 0$ when it does not (setting an indicator variable $y_{2,ij,t} = 0$). Thus, Equation (6) can be estimated as a discrete choice model, such as a probit model, given that the errors are normally distributed.

Finally, as we do not want to impose that $\sigma_\eta^2 = \sigma_u^2 + \sigma_\nu^2 = 1$, we can divide the equation above by the standard deviation σ_η, yielding:

$$y_{2,ij,t}^{**} = x_{2,ij,t}'\beta_2^* + \xi_i^* + \zeta_j^* + \eta_{ij,t}^* \tag{7}$$

Therefore, given Assumption 3 and Equation (7), the authors start by specifying the Probit equation:

$$\mu_{ij,t} = \Pr\left(y_{2,ij,t} = 1 | \text{ observed variables }\right) = \Phi\left(x_{2,ij,t}'\beta_2^* + \xi_i^* + \zeta_j^*\right), \tag{8}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. This probit equation not only allows for an estimate of the inverse Mills-ratio, used for taking into account the sample selection in the Equation (5), but also allows for obtaining a consistent estimate of $\mathbb{E}[w_{ij}|\cdot, y_{2,ij,t} = 1]$ [3], which is then incorporated in the estimation of Equation (5). Note that, we use an estimate for the variable conditional on the population of countries that do trade amongst themselves, as this is the sample considered in the estimation of Equation (5).

---

[3]It is noteworthy that the distributional assumptions on joint normality of the unobserved trade costs and the Pareto distribution of firm-level productivity affect the functional form of the trade flow equation via the functional form of the controls for firm heterogeneity ($w_{ij,t}$) and the sample selection.

In the following sections a more simplified model will be defined, together with a more detailed explanation of the standard Heckman procedure for correcting for sample selection. We will then motivate why this procedure may not yield unbiased coefficients for the observation equation (the gravity equation) and the correct inference due to the incidental parameters problem (Neyman and Scott (1948)).

4 A simplified model of dyadic interactions

Taking the model proposed by Helpman et al. (2008) as a motivation, and simplifying it for now, focusing on a specification considering two-way fixed effects and a possible selection bias, we study the following model throughout the next sections:

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + u_{ij,t} \tag{9}$$

$$y_{2,ij,t} = \mathbb{1}(y^{**}_{2,ij,t} > 0) \tag{10}$$

$$y^{**}_{2,ij,t} = x'_{2,ij,t}\beta^*_2 + \xi^*_i + \zeta^*_j + \eta^*_{ij,t}, \tag{11}$$

$$(i = 1, ...N; j = 1, ...N, t = 1, ...T, i \neq j)$$

where $y_{1,ij,t}$ is observed only when $y_{2,ij,t} = 1$. However, all explanatory variables are assumed to be observed in all periods for all i and j, being a typical case of a censored panel. Moreover, $\beta_1 \in \mathbb{R}^k$ is a vector that collects the coefficients of the structural parameters, including a coefficient of an intercept (β_0), and of the explanatory variables; $x_{1,ij,t}$ is a vector that collects 1 and the vector of the explanatory variables; and χ_j and ϑ_i are fixed effects. Similarly, $\beta^*_2 \in \mathbb{R}^q$ is a vector that collects the coefficients of the structural parameters, including the coefficient of an intercept (β^*_0), and of the explanatory variables; $x_{2,ij,t}$ is a vector that collects 1 and the column vectors of the explanatory variables; and ξ^*_i and ζ^*_j are fixed effects. In both equations, the unobservable individual-specific effects may depend on the observable explanatory variables in an arbitrary way, and are, therefore, considered as nuisance parameters to be estimated. Moreover, the errors of the equations might be correlated.

Note that also in this generalized framework we can have directed links and directed outcomes, meaning that $y_{2,ij,t}$ does not necessarily have the same value of $y_{2,ji,t}$ and vice versa. The same holds for $y_{1,ij,t}$ and $y_{1,ji,t}$.

This model is simplified compared to the estimated equations of Helpman et al. (2008) because it does not take into account the variable w_{ij}. This variable, in the setting of international trade, refers to the fraction of exporting firms in a country. It is obtained from the estimation of the single

index in the inverse Mills-ratio, and it introduces a further non-linearity in the model, which requires the observation equation to be estimated through NLS.

Note that this is a so-called type 2 tobit model, in which the sample selectivity (in the case of correlated errors) introduces non-linearity in the equation of interest with respect to the unobserved characteristics. Moreover, such fixed effects in Equation (11), unlike in linear models, cannot be easily differenced away.

In order to estimate such a model by MLE, one would need not only to specify the distribution of underlying time-varing errors, but also to specify a functional form for how the fixed effects depend on observed variables. Besides being non-robust to distributional misspecification, this fully parametric "random effects" approach involves multiple numerical integrations, becoming computationally intensive.

Even though the motivation for this model comes from the trade literature, this specification is general enough to cover any possible case of dyadic data. As defined by Graham (2020), dyadic data reflects situations where the outcome of interest is determined by pairwise interactions between units. Other fields where such data is analysed includes international financial flows, development economics and anthropology. Moreover, the selection model given by Equations (10) and (11) is a typical model of network formation, where links between units are formed or not. In this model, we consider a directed network - where, as expected, the direction of the links and the resulting outcome matters, while an unit i may form a link with an j, the opposite may not be true. In this setting, the link formation given by the variable $y_{2,ij,t}$ determines the observation of an outcome, given by the variable $y_{1,ij,t}$. More specifically, our framework is very similar to that of a network formation under dyadic interaction with degree heterogeneity (captured by the fixed effects) [4], a framework embedded in the so called β-models in the network literature.

5 The Heckman 2-stage approach for sample selection

In this section we will present the standard approach presented by Heckman (1979) to correct for the sample selection bias in this analysed model. It is important to mention beforehand that such bias arises from a possible correlation between the errors $u_{ij,t}$ and $\eta^*_{ij,t}$. We start with some standard assumptions:

---

[4]Degree heterogeneity is defined, in the networks literature, as the fact that the number of links (degree) across nodes (units) varies.

ASSUMPTION 4. The errors $u_{ij,t}$ and $\eta^*_{ij,t}$ satisfy:

$$\mathbb{E}[u_{ij,t}] = 0$$
$$\mathbb{E}[\eta^*_{ij,t}] = 0$$
$$\mathbb{E}[u_{ij,t}u_{i'j',t'}] = \begin{cases} \sigma^2_u \text{ if } i = i', j = j', t = t' \\ 0 \text{ otherwise} \end{cases}$$
$$\mathbb{E}[\eta^*_{ij,t}\eta^*_{ij,t}] = \begin{cases} \sigma^2_{\eta^*} = 1 \text{ if } i = i', j = j', t = t' \\ 0 \text{ otherwise} \end{cases}$$
$$\mathbb{E}[u_{ij,t}\eta^*_{ij,t}] = \begin{cases} \sigma_{u\eta^*} \text{ if } i = i', j = j', t = t' \\ 0 \text{ otherwise} \end{cases}$$

If we seek to estimate Equation (9), we need to take into account that the estimation will only consider a selected sample, which is given by Equations (10) and (11). We also assume, without loss of generality, that we observe $y_{1,ij,t}$ for the first N_i observations related to the index i and the first N_j observations related to the index j. Moreover, all the links between the observations N_i of i are formed with all observations N_j of j and vice-versa, and therefore, ultimately $N_i = N_j$. Thus, conditioning on observing the outcomes of $y_{1,ij,t}$ only for the pairs that form links, we can rewrite Equation (9) as:

$$\mathbb{E}[y_{1,ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j, y_{2,ij,t} = 1] \tag{12}$$
$$= x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + \mathbb{E}[u_{ij,t}|\eta^*_{ij,t} \geq x'_{2,ij,t} - \beta^*_2 - \xi^*_i - \zeta^*_j]$$

If we had that $u_{ij,t}$ is independent of $\eta^*_{ij,t}$, and therefore, the conditional expectation of $u_{ij,t}$ is zero, the regression function for the selected subsample would be the same as the populational regression function. However, in the general case given by Assumption 4, we have that there is correlation between these errors. Then, the conditional mean of the error $\eta^*_{ij,t}$ in the incomplete sample is a function of the explanatory variables $x_{2,ij,t}$ and the fixed effects ξ^*_i and ζ^*_j. Therefore, the sample selection bias arises because this last term is not taken into account.

More generally, we can denote the single index:

$$z_{ij,t} = \frac{-x'_{2,ij,t}\beta^*_2 - \xi^*_i - \zeta^*_j}{\sigma^2_{\eta^*}} = -x'_{2,ij,t}\beta^*_2 - \xi^*_i - \zeta^*_j.$$

We can also assume that the conditional expectation in the last term of Equation (12) is an unknown, smooth function $\varphi_{ij,t}$ of the scalar index $z_{ij,t}$ and other distributional parameters in the parametric case, or a function only of the scalar index in a nonparametric case. Therefore, we can write:

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + \varphi_{ij,t}(z_{ij,t}) + \nu_{ij,t}, \tag{13}$$

13

where, by construction, $\mathbb{E}[\nu_{ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j, \varphi_{ij,t}(z_{ij,t}), y_{2,ij,t}^{**} > 0] = 0$.

For the identification of the regression coefficients of (9), we need that either the function $\varphi_{ij,t}$, which ultimately is a function of the regressors in (11), is specified, otherwise, we need strong exclusion restrictions. As pointed out by Ahn and Powell (1993), for example, taking any nontrivial linear combination $x_{1,ij,t}'\alpha$ of the regressors $x_{1,ij,t}$, then Equation (13) can be rewritten:

$$
\begin{aligned}
y_{1,ij,t} &= x_{1,ij,t}'(\beta_1 + \alpha) + \vartheta_i + \chi_j + [\varphi_{ij,t}(z_{ij,t}) - x_{1,ij,t}'\alpha] + \nu_{ij,t} \qquad (14) \\
&= x_{1,ij,t}'\beta_1^* + \vartheta_i + \chi_j + \varphi_{ij,t}^*(z_{ij,t}) + \nu_{ij,t}^*
\end{aligned}
$$

where: $\varphi_{ij,t}^*(z_{ij,t}) = \varphi_{ij,t}(z_{ij,t}) - \mathbb{E}[x_{1,ij,t}'\alpha|z_{ij,t}, y_{2,ij,t}^{**} > 0]$; and $\nu_{ij,t}^* = \nu_{ij,t} - x_{1,ij,t}'\alpha + \mathbb{E}[\alpha'x_{1,ij,t}|z_{ij,t}, y_{2,ij,t}^{**} > 0]$.

Therefore, we have that this model still satisfies that $\mathbb{E}[\nu_{ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j, \varphi_{ij,t}(z_{ij,t}), y_{2,ij,t}^{**} > 0] = 0$, but β_1^* is not distinguishable from β_1 without further restrictions on the functional form of $\varphi_{ij,t}(\cdot)$. This can be done either through additional variables that satisfies exclusion restrictions from the equation of interest (meaning that they affect the link formation, and thus, are included in the regressors $x_{2,ij,t}$, but they are not included in the regressors $x_{1,ij,t}$ and are uncorrelated with the errors of equation 9), or through specifying a functional form of $\varphi_{ij,t}(\cdot)$.

To introduce a functional form for $\varphi_{ij,t}(\cdot)$, additional distributional assumptions are made:

ASSUMPTION 5. The error term $u_{ij,t}$ is identically and independently distributed over ij and t, such that $u_{ij,t} \sim N(0, \sigma_u^2)$.

ASSUMPTION 6. The error term $\eta_{ij,t}^*$ is identically and independently distributed over ij and t, such that $\eta_{ij,t}^* \sim N(0, 1)$.

ASSUMPTION 7. The errors $u_{ij,t}$ and $\eta_{ij,t}^*$ are correlated such that:

$$
\mathbb{E}[u_{ij,t}\eta_{i'j',t'}^*] = \begin{cases} \sigma_{u\eta^*} & \text{if } i = i', \quad j = j', \quad t = t' \\ 0 & \text{otherwise} \end{cases}
$$

Under these assumptions, and supposing that $B(u_{ij,t}, \eta_{ij,t}^*; \rho)$ is the joint bivariate normal density of $u_{ij,t}$ and $\eta_{ij,t}^*$, with correlation coefficient $\rho = \frac{\sigma_{u\eta^*}}{\sigma_u \sigma_{\eta^*}}$, we then have from standard results of the bivariate normal distribution that:

$$
\mathbb{E}[u_{ij,t}|\eta_{ij,t}^* \geq -x_{2,ij,t}'\beta_2^* - \xi_i^* - \zeta_j^*] = \frac{\sigma_{u\eta^*}}{\sigma_{\eta^*}}\lambda_{ij,t}(z_{ij,t}) = \sigma_{u\eta^*}\lambda_{ij,t}(z_{ij,t}) \quad (15)
$$

where:

$$
\lambda_{ij,t}(z_{ij,t}) = \frac{\phi(z_{ij,t})}{1 - \Phi(z_{ij,t})} = \frac{\phi(z_{ij,t})}{\Phi(-z_{ij,t})} \qquad (16)
$$

14

where ϕ is the density function for a standard normal variable and Φ is its distribution function. Note that $\lambda_{ij,t}(z_{ij,t})$ is the inverse Mills-ratio, which is a monotone decreasing function of the probability of observing $y_{1,i,j,t}$. Some of its properties are: $\lim_{\Phi(-z_{ij,t})\to 1} \lambda_{ij,t} = 0$, $\lim_{\Phi(-z_{ij,t})\to 0} \lambda_{ij,t} = \infty$, and $\partial\lambda_{ij,t}/\partial\Phi(-z_{ij,t}) < 0$.

At this point, it is interesting to note that under the distributional Assumptions 5 - 7, the previously given function $\varphi_{ij,t}(z_{ij,t})$ takes the form $\varphi_{ij,t}(z_{ij,t}) = \sigma_{u\eta^*}\lambda_{ij,t}(z_{ij,t})$. Thus, in general, we will obtain inconsistent coefficient estimates for the observation equation if there is misspecification of the parametric form of the index function and of errors.

We can rewrite Equation (13) as:

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + \sigma_{u\eta^*}\lambda_{ij,t}(z_{ij,t}) + \nu_{ij,t} \tag{17}$$

Therefore, if one knew $z_{ij,t}$, one could enter $\lambda_{ij,t}(z_{ij,t})$ as a regressor in Equation (13) and estimate it by OLS (note that, in our specification, the two-way fixed effects could be estimated by the inclusion of dummy variables). In that case, the least squares estimators of β_1 and $\sigma_{u\eta^*}$ are unbiased but inefficient, due to the heteroskedasticity in $\mathbb{E}[\nu_{ij,t}^2]$, as shown in Heckman (1979):

$$\mathbb{E}[\nu_{ij,t}^2] = \sigma_u^2\Big((1-\rho^2) + \rho^2(1 + z_{ij,t}\lambda_{ij,t} - \lambda_{ij,t}^2)\Big)$$

As $0 \leq 1 + z_{ij,t}\lambda_{ij,t} - \lambda_{ij,t}^2 \leq 1$, the usual estimator of the covariances is downward biased. The standard GLS procedure could be employed to obtain appropriate standard errors for estimated coefficients of the first equation.

In practice, as we do not know $z_{ij,t}$, Heckman (1979) suggests the following procedure:

- Step 1: Estimate the probability that $y_{2,ij,t}^{**} \geq 0$ using probit analysis on the sample, given by Equations (10) and (11).

- Step 2: From this estimator (provided that it is consistent), one can obtain $\hat{z}_{ij,t}$ consistently.

- Step 3: The estimated value of $\lambda_{ij,t}(z_{ij,t})$ is used as a regressor in Equation (17) fit on the subsample. The regression estimators are then consistent for β_1 and $\sigma_{u\eta^*}$.

- Step 4: One can then consistently estimate σ_u^2 from the following. From step (3), we consistently estimate $\sigma_{u\eta^*}$, through the estimator $\hat{\sigma}_{u\eta^*}$. Denote the residuals for each observation from step 3 as $\hat{\nu}_{ij,t}$. Then, using $\hat{z}_{ij,t}$ and $\hat{\lambda}_{ij,t}$ the estimated values from step (2), an estimator of σ_u^2 is:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N_i}\sum_{j\neq i}\sum_{t=1}^{T}\hat{\nu}_{ij,t}^2}{N_i(N_i-1)T} - \frac{\hat{\sigma}_{u\eta^*}}{N_i(N_i-1)T}\sum_{i=1}^{N_i}\sum_{j\neq i}\sum_{t=1}^{T}(\hat{\lambda}_{ij,t}\hat{z}_{ij,t} - \hat{\lambda}_{ij,t}^2)$$

Finally, to obtain the limiting distribution of the estimates of β_1 to conduct inference, Heckman (1979) proposes to first look at Equation (17) with an estimated value of $\lambda_{ij,t}$ used in place of its true value.

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + \sigma_{u\eta^*}\hat{\lambda}_{ij,t} + \sigma_{u\eta^*}(\lambda_{ij,t} - \hat{\lambda}_{ij,t}) + \nu_{ij,t} \qquad (18)$$

We can interpretate $\sigma_{u\eta^*}(\lambda_{ij,t} - \hat{\lambda}_{ij,t}) + \nu_{ij,t}$ as the error term of this equation. From this, the following theorem arises:

THEOREM 1. *The estimator of β_1 will be an asymptotically unbiased estimator, i.e., will converge to a limiting distribution centered around its true value only if $\mathbb{E}[(\lambda_{ij,t} - \hat{\lambda}_{ij,t})] = 0$.*

Proof. From standard properties of OLS/FGLS estimators, we have that one of the conditions for $\hat{\beta}_1$ to be asymptotically unbiased [5] is that:

$$\mathbb{E}[\sigma_{u\eta^*}(\lambda_{ij,t} - \hat{\lambda}_{ij,t}) + \nu_{ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j] = 0$$

By linearity of expectations, this translates to:

$$\sigma_{u\eta^*}\mathbb{E}[(\lambda_{ij,t} - \hat{\lambda}_{ij,t})|x_{1,ij,t}, \vartheta_i, \chi_j] + \sigma_{u\eta^*}\mathbb{E}[\nu_{ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j] = 0$$

We know that by construction, the last term of this expression equals to zero. Moreover, in the case where sample selection is present, we have that $\sigma_{u\eta^*} > 0$. Therefore, the condition simplies to:

$$\mathbb{E}[(\lambda_{ij,t} - \hat{\lambda}_{ij,t})|x_{1,ij,t}, \vartheta_i, \chi_j] = 0$$

If this condition holds, we have by the Law of Iterated Expectations that:

$$\mathbb{E}[(\lambda_{ij,t} - \hat{\lambda}_{ij,t})] = \mathbb{E}[\mathbb{E}[(\lambda_{ij,t} - \hat{\lambda}_{ij,t})|x_{1,ij,t}, \vartheta_i, \chi_j]] = 0$$

$$\square$$

The limiting distribution of the standard Heckman estimator for a simple linear regression without fixed effects can be found in Helpman et al. (2008).

In the next section, I discuss that in our framework, the incidental parameters problem will imply a the condition of Theorem 1 is not satisfied.

---

[5] Here, we use the formal definition of an asymptotically unbiased estimator $\hat{\theta}_n$ for a parameter θ to be:

$$r_n(\hat{\theta}_n - \theta) \xrightarrow{d} H \qquad (19)$$

Where r_n is some sequence and where the expected value of H is 0

6   The incidental parameter problem

Even though Helpman et al. (2008) directly estimated Equation (8) by maximum likelihood, it is well known in the literature that the fixed effects estimators for nonlinear panel data models suffers from the incidental parameter problem (Neyman and Scott (1948)), yielding asymptotically biased estimates of the parameters.

As highlighted by Arellano and Hahn (2007) in a standard panel data regression with one way fixed effects and dimensions $i = 1, ...N$ and $t = 1, ...T$, if T is fixed and $N \to \infty$, there will be an estimation error in the estimates of the fixed effects, as only a finite number T of observations are available to estimate each fixed effect. As we allow for the fixed effects to be correlated with the exogenous regressors (and its distribution is left unspecified), this estimation error contaminates the estimates of the other parameters as well, as they are not informationally orthogonal. For large enough T, this bias should be small. However, even under $T \to \infty$ and $N \to \infty$, the fixed effects estimator will be asymptotically biased, leading to incorrect inference over the parameters and the average partial effects.

The same argument holds for our framework of a dyadic regression with two-way fixed effects. In our panel data model, we have 3 dimensions: $i = 1, ..N$, $j = 1, ...N$ and $t = 1, ...T$. Thus, the first two dimensions grow at rate N, and the latter at rate T. As most of the available datasets in the trade literature have the dimension T fixed, we will consider asymptotic results such that $N \to \infty$ and T is fixed.

Note as well that for each new country in the dataset, the number of observations is increased by $2(N-1)T$. Moreover, for each fixed effect in Equation (11) there are $(N-1)T$ observations available for their estimation.

We will now use results shown by Fernández-Val and Weidner (2016) to demonstrate how the incidental parameter problem arises in our framework, delivering consistent but asymptotic biased estimators, keeping in mind that as $N \to \infty$ both dimensions i and j go to infinity and also the number of observations $N(N-1)T$ go to infinity.

Given the dataset of $N(N-1)T$ observations $\{(y_{2,ij,t}, x'_{2,ij,t})' : 1 \leq i \leq N, 1 \leq j \leq N, 1 \leq t \leq T, i \neq j\}$ with $y_{2,ij,t} = \mathbb{1}(y^{**}_{2,ij,t} > 0)$, and $y^{**}_{2,ij,t}$ specified by Equation (11) where $\eta^*_{ij,t}$ is i.i.d., we have that $y_{2,ij,t}$ is generated by the process:

$$y_{2,ij,t}|x_{2,ij,t}, \xi^*, \zeta^*, \beta_2^* \sim f_{Y_2}(\cdot|x_{2,ij,t}, \xi^*, \zeta^*, \beta_2^*)$$

where: $\xi^* = (\xi_1^*, ...\xi_N^*)$, $\zeta^* = (\zeta_1^*, ...\zeta_N^*)$, f_{Y_2} is a known probability function and ξ_i^*, ζ_j^* are the unobserved fixed effects. Note here that this approach is semi-parametric in the sense that is does not specify the distribution of the fixed effects or their relationship with the explanatory variables.

We can further model the conditional distribution of $y_{2,ij,t}$ using a single-

index specification with fixed effects, since it is a binary response model:

$$f_{Y_2}(y_{2,ij,t}|x_{2,ij,t}, \xi^*, \zeta^*, \beta_2^*) = F(x'_{2,ij,t}\beta_2^* + \xi_i^* + \zeta_j^*)^{y_{2,ij,t}}$$
$$\times [1 - F(x'_{2,ij,t}\beta_2^* + \xi_i^* + \zeta_j^*)]^{1-y_{2,ij,t}}, \qquad (20)$$

where, clearly $y_{2,ij,t} \in \{0,1\}$ and F is a cumulative distribution function, defined to be a standard normal according to Assumption 6.

We can then collect all the fixed effects to be estimated in the vector $\omega_{NN}^* = (\xi_1^*, ...\xi_N^*, \zeta_1^*, ...\zeta_N^*)'$, which can be seen as a nuisance parameter vector. Then, the true values of the parameters, denoted by $\beta_{2,0}^*$ and $\omega_{NN,0}^* = (\xi_{1,0}^*, ...\xi_{N,0}^*, \zeta_{1,0}^*, ...\zeta_{N,0}^*)'$ are the solution to the population conditional maximum likelihood maximization:

$$\max_{(\beta_2^*, \omega_{NN}^*) \in \mathbb{R}^{\dim\beta^* + \dim\omega_{NN}^*}} \mathbb{E}_\omega[\mathcal{L}(\beta_2^*, \omega_{NN}^*)] \qquad (21)$$

with

$$\mathcal{L}(\beta_2^*, \omega_{NN}^*) \qquad (22)$$

$$= (N(N-1)T)^{-1}\left\{ \sum_{i=1}^{N}\sum_{j \neq i}\sum_{t=1}^{T} \log f_{Y_2}(y_{2,ij,t}|x_{2,ij,t}, \xi^*, \zeta^*, \beta_2^*) - b(\iota'_{NN}\omega_{NN}^*)^2/2 \right\}$$

where $\mathbb{E}_\omega$ denotes the expectation with respect to the distribution of the data conditional on the unobserved effects and strictly exogenous variables, $b > 0$ is an arbitrary constant, $\iota_{NN} = (1'_N, -1'_N)'$ and 1_N denotes a vector of ones of dimension N.

The second term of $\mathcal{L}$ relates to a penalty that imposes a normalization to identify the fixed effects in models with two-way fixed effects that enter in the log-likelihood function as $\xi_i^* + \zeta_j^*$. To be more specific, in this case, adding a constant to all ξ_i^* and subtracting the same constant from all ζ_j^* would not change $\xi_i^* + \zeta_j^*$. Thus, without this normalization, the parameters ξ_i^* and ζ_j^* are not identifiable.

To estimate the parameters, we solve the sample analogue of the following equation:

$$\max_{(\beta_2^*, \omega_{NN}^*) \in \mathbb{R}^{\dim\beta_2^* + \dim\omega_{NN}^*}} \mathcal{L}(\beta_2^*, \omega_{NN}^*) \qquad (23)$$

In order to analyze the statistical properties of β_2^*, we first concentrate out the nuisance parameters ω_{NN}^*, such that for given β_2^*, the optimal $\hat{\omega}_{NN}^*(\beta_2^*)$ is:

$$\hat{\omega}_{NN}^*(\beta_2^*) = \text{argmax}_{\omega_{NN}^* \in \mathbb{R}^{\dim\omega_{NN}^*}} \mathcal{L}(\beta_2^*, \omega_{NN}^*) \qquad (24)$$

Thus, the fixed effects estimator of β_2^* and ω_{NN}^* are, by plugging in the previous expression for $\hat{\omega}_{NN}^*(\beta_2^*)$:

$$\hat{\beta}_2^* = \text{argmax}_{\beta_2^* \in \mathbb{R}^{\dim \beta_2^*}} \mathcal{L}(\beta_2^*, \hat{\omega}_{NN}^*(\beta_2^*)) \tag{25}$$

$$\hat{\omega}_{NN}^*(\beta_2^*) = \hat{\omega}_{NN}^*(\hat{\beta}_2^*) \tag{26}$$

The source of the problem is that the dimension of the nuisance parameters ω_{NN} increases with the sample size under asymptotic approximations where $N \to \infty$. To further describe the incidental parameter problem, denote:

$$\bar{\beta}_2^* = \text{argmax}_{\beta_2^* \in \mathbb{R}^{\dim \beta_2^*}} \mathbb{E}_\omega \Big[\mathcal{L}(\beta_2^*, \hat{\omega}_{NN}^*(\beta_2^*)) \Big]$$

Using an asymptotic expansion for smooth likelihoods under appropriate regularity conditions, provided by Fernández-Val and Weidner (2016), we have that:

$$\bar{\beta}_2^* = \beta_{2,0}^* + \frac{\bar{B}_\infty}{(N-1)T} + \frac{\bar{D}_\infty}{(N-1)T} + \text{o}_P(((N-1)T)^{-1}) \tag{27}$$

For some constants $\bar{B}_\infty$ and $\bar{D}_\infty$. The derivation for this expression can be found in the Appendix of Fernández-Val and Weidner (2016). As explained by the authors, the expansion is obtained by first taking a first-order Taylor expansion of the Equation (25) around the true value $\beta_{2,0}^*$, as it is usually done to obtain the asymptotic properties of such estimator. Then, one should additionally take a second-order Taylor expansion of the obtained term $\frac{\partial \mathcal{L}(\beta_{2,0}^*, \hat{\omega}_{NN}^*)}{\partial_{\beta_2^*}}$ around the true values of the nuisance terms. Intuitively, this second step demonstrates how the estimates of the fixed effects affect the estimates of the structural parameter β_2^*. We provide a more formalized form of this argument in Appendix A.3. To obtain the exact form of the expressions $\bar{B}_\infty$ and $\bar{D}_\infty$ a quite involved derivation is needed. However, this is not the focus of our study, since we show later that there are other possibilities to correct for the asymptotic bias generated by these terms other than deriving the biases themselves.

Moreover, by the properties of the maximum likelihood estimator we have that, under regularity conditions:

$$\sqrt{N(N-1)T}(\hat{\beta}_2^* - \bar{\beta}_2^*) \xrightarrow{d} N(0, \bar{V}_{B\infty}) \tag{28}$$

For some $\bar{V}_{B\infty}$. By substituting the expression for $\beta_{2,0}^*$ obtained in Equa-

tion (27), we obtain that, by Slutsky's theorem:

$$\sqrt{N(N-1)T}(\hat{\beta}_2^* - \beta_{2,0}^*) \tag{29}$$
$$= \sqrt{N(N-1)T}(\hat{\beta}_2^* - \bar{\beta}_2^*)$$
$$+ \sqrt{N(N-1)T}\Big(\frac{\bar{B}_\infty}{(N-1)T} + \frac{\bar{D}_\infty}{(N-1)T} + o_P(((N-1)T)^{-1})\Big)$$
$$\xrightarrow{d} N\Big(\frac{\bar{B}_\infty}{\sqrt{T}} + \frac{\bar{D}_\infty}{\sqrt{T}}, \bar{V}_{B\infty}\Big)$$

We can see from Equation (27) that, as $N \to \infty$, $\hat{\beta}_2^* \xrightarrow{p} \beta_{2,0}^*$ ($\beta_{2,0}^*$ being the true value of the parameter), thus, the estimates of $\beta_{2,0}^*$ are consistent. However, from Equation (29) we see that when T is fixed, the estimates converge to a distribution that is not centered at zero, which leads to incorrect asymptotic confidence intervals. This demonstrates the incidental parameters problem, that boils down to an asymptotic bias in the estimates of $\beta_{2,0}^*$. This asymptotic bias arises as the order of the bias is higher than the inverse of the sample size because of the small rate of convergence of the fixed effects.

CONJECTURE 1. *The asymptotic bias in the estimates of $\beta_{2,0}^*$ carries over both to the estimates of the fixed effects, $\hat{\omega}_{NN}^*$ and to the estimates of the inverse Mills-ratio, $\hat{\lambda}_{ij,t}$.*

This conjecture follows from Theorem B.1. in the appendix of Fernández-Val and Weidner (2016), where it is shown that the asymptotic expansion of $\hat{\omega}_{NN}^*$ around $\omega_{NN,0}$ depends on $(\beta_2^* - \beta_{2,0}^*)$. Moreover, as we see from Equation (26), the estimates of the fixed effects are a function of the estimates of β_2^*. Therefore, it is expected that given an asymptotically biased estimator $\hat{\beta}_2^*$, the estimators of the fixed effects will also converge to a limiting distribution not centered around its true values. Also, as mentioned by Fernández-Val and Weidner (2016), its rate of convergence is slower than that of the structural parameter, $\sqrt{(N-1)T}$. Thus, we expect that, under fixed T:

$$\sqrt{(N-1)T}(\hat{\omega}_{NN}^* - \omega_{NN,0}) \xrightarrow{d} N\Big(\bar{A}_\infty, \bar{V}_{A\infty}\Big). \tag{30}$$

For some constant $\bar{A}_\infty$. Remembering that $\hat{\lambda}_{ij,t}$ is a function of the estimated fixed effects and of the estimated structural parameters, if the rates of convergence of both random variables would be the same, one could straightforwardly apply the Delta method to determine the distribution of $\hat{\lambda}_{ij,t}$.

In our case, as the fixed effects converge at a slower rate, we expect that the joint distribution will only converge at the slower rate. On this note, if

we scale the random variable $(\hat{\beta}_2^* - \beta_{2,0}^*)$ by this rate it is expected that it converges to a constant denoted by $\bar{C}_\infty$. Thus:

$$\sqrt{(N-1)T}\begin{pmatrix} \hat{\beta}_2^* - \beta_{2,0}^* \\ \hat{\omega}_{NN}^* - \omega_{NN,0} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \bar{C}_\infty \\ \bar{A}_\infty \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & \bar{V}_{A\infty} \end{pmatrix}\right)$$

Having defined this joint distribution, one can apply the Delta method easily, verifying that the limiting distribution of $\hat{\lambda}_{ij,t}$ is asymptotically biased even if the constant $\bar{C}_\infty$ is zero. In case the constant is not zero, the only possibility for it being asymptotically unbiased would be if the biases from the fixed effects and from the structural parameters cancel out. However, there is no reason to believe that this should be the case.

Under this conjecture, it is noteworthy that the conditions of Theorem 1 do not hold. Therefore, if the selection equation is estimated by probit in this case, it is likely that the estimators of β_1 in Equation (9) will also be asymptotically biased. A more formalized version of this conjecture is left for further research. For now, we aim to explore the finite sample properties of such estimator through Monte Carlo simulations.

To start addressing this issue, one can start looking at a possible asymptotically unbiased estimator for the structural parameter β_2^*. Fernández-Val and Weidner (2016) tackles this problem by deriving asymptotic corrections, based on defining analytically the values of $\bar{B}_\infty$, $\bar{D}_\infty$ and $\bar{V}_{B\infty}$. Other studies propose different solutions for the incidental parameter problem in nonlinear panel data models. To name a few, Jochmans (2019) proposes modifications of the profile likelihood in models for dyadic interactions, covering the so called β-models in the framework of network formation, and delivering asymptotic unbiased estimators; Graham (2017) provides an estimator for undirected dyadic link formation in network models, which takes into account the degree heterogeneity of agents (the fixed effects in our approach) through a tetrad logit (TL) estimator that conditions on sufficient statistics for the degree heterogeneity; and Dzemski (2019) also provides analytical asymptotic corrections for estimating a model of directed dyadic link formation.

We propose in this study, instead of relying on approximate asymptotic corrections for such biases, to estimate the parameters through a conditional logit approach proposed by Charbonneau (2017). Such approach differences away both fixed effects. The advantage of the estimator proposed by Charbonneau (2017) is that the analytical form of the biases does not need to be specified, while still delivering consistent and unbiased estimates of the parameters.

7 The conditional logit estimation of Charbonneau (2017)

The method displayed in this section relies on eliminating the two-way fixed effects of the model in order to obtain estimates of the parameters β_2^* that are not contaminated by the incidental parameter problem. This approach is proposed by Charbonneau (2017) and is inspired on a well-known method to eliminate fixed effects in single fixed effects model. The author shows that it is possible to generalize the conditional maximum likelihood approach to include two-way fixed effects only for logit models. Therefore, in order to estimate the model given by Equations (10) and (11), we modify the previous Assumption 6 to:

ASSUMPTION 8. The idiosyncratic terms η_{ij}^* are i.i.d. across ij and follow a standard logistic distribution conditional on the regressors and fixed effects.

To adopt this method, we consider only one period t for now. In the following we will follow closely the exposition of Arellano and Honoré (2001) for the approach to eliminate a single fixed effects from the model. Supposing that the observations are given by:

$$y_{2,ij} = \mathbb{1}\{x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^* + \eta_{ij}^* \geq 0\} \text{ for } i = 1,...N \quad j = 1,...N \quad i \neq j$$

which, given Assumption 8 implies that:

$$\Pr(y_{2,ij} = 1|x_2, \xi^*, \zeta^*, \beta_2^*) = \frac{\exp(x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^*))}{1 + \exp(x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^*))}$$

where x_2 denotes the vectors of observations $x_{2,ij}$ for all possible pairs ij. Fixing distinct indices i, j, k and l from the set form by individuals in $i' = 1,...,N$, it is straightforward to show that the conditional likelihood can be written such that it eliminates the fixed effects ξ:

$$\Pr(y_{2,lj} = 1|x_2, \xi^*, \zeta^*, \beta_2^*, y_{2,lk} + y_{2,lj} = 1)$$
$$= \frac{\Pr(y_{2,lj} = 1|x_2, \xi^*, \zeta^*, \beta_2^*) \quad \Pr(y_{2,lk} = 0|x_2, \xi^*, \zeta^*, \beta_2^*)}{\Pr(y_{2,lj} = 1, y_{2,lk} = 0|x_2, \xi^*, \zeta^*, \beta_2^*) + \Pr(y_{2,lj} = 0, y_{2,lk} = 1|x_2, \xi^*, \zeta^*, \beta_2^*)}$$
$$= \frac{\exp((x_{2,lj} - x_{2,lk})'\beta_2^* + \zeta_j^* - \zeta_k^*)}{1 + \exp((x_{2,lj} - x_{2,lk})'\beta_2^* + \zeta_j^* - \zeta_k^*)}$$

which is again the specification of a logit model, with $x_{2,lj,t} - x_{2,lk,t}$ as explanatory variables and $\zeta_j^* - \zeta_k^*$ as a fixed effect. We can also apply the same procedure to another pair of observations ij, ik:

$$\Pr(y_{2,ij} = 1|x_2, \xi^*, \zeta^*, \beta_2^*, y_{2,ij} + y_{2,ik} = 1)$$
$$= \frac{\Pr(y_{2,ij} = 1|x_2, \xi^*, \zeta^*, \beta_2^*) \quad \Pr(y_{2,ik} = 0|x_2, \xi^*, \zeta^*, \beta_2^*)}{\Pr(y_{2,ij} = 1, y_{2,ik} = 0|x_2, \xi^*, \zeta^*, \beta_2^*) + \Pr(y_{2,ij} = 0, y_{2,ik} = 1|x_2, \xi^*, \zeta^*, \beta_2^*)}$$
$$= \frac{\exp((x_{2,ij} - x_{2,ik})'\beta_2^* + \zeta_j^* - \zeta_k^*)}{1 + \exp((x_{2,ij} - x_{2,ik})'\beta_2^* + \zeta_j^* - \zeta_k^*)}$$

These conditional likelihoods are free of the fixed effects ξ^*, but are still dependent on the fixed effects ζ^*. The idea now is to combine the above two expressions, by adding a condition on $y_{2,ij,t} + y_{2,lj,t} = 1$.

Defining : $C \equiv \{y_{2,lk} + y_{2,lj} = 1, y_{2,ij} + y_{2,ik} = 1\}$, we can now write:

$$\Pr(y_{2,lj} = 1 | x_2, \xi^*, \zeta^*, \beta_2^*, y_{2,lk} + y_{2,lj} = 1, y_{2,ij} + y_{2,ik} = 1, y_{2,ij} + y_{2,lj} = 1) \tag{31}$$

$$
\begin{aligned}
&= \frac{\Pr(y_{2,lj} = 1, y_{2,ij} + y_{2,lj} = 1 | x_2, \xi^*, \zeta^*, \beta_2^*, C)}{\Pr(y_{2,ij} + y_{2,lj} = 1 | x_2, \xi^*, \zeta^*, \beta_2^*, C)} \\
&= \frac{\Pr(y_{2,lj} = 1 | x_2, \xi^*, \zeta^*, \beta_2^*, C) \quad \Pr(y_{2,ij} = 0 | x_2, \xi^*, \zeta^*, \beta_2^*, C)}{\Pr(y_{2,lj} = 1, y_{2,ij} = 0 | x_2, \xi^*, \zeta^*, \beta_2^*, C) + \Pr(y_{2,lj} = 0, y_{2,ij} = 1 | x_2, \xi^*, \zeta^*, \beta_2^*, C)} \\
&= \frac{\exp(((x_{2,lj} - x_{2,lk}) - (x_{2,ij} - x_{2,ik}))'\beta_2^*)}{1 + \exp(((x_{2,lj} - x_{2,lk}) - (x_{2,ij} - x_{2,ik}))'\beta_2^*)}
\end{aligned}
$$

This final probability no longer depends on fixed effects, allowing to solve the incidental parameters problem in the presence of two-way fixed effects. It is now possible to either write a conditional maximum likelihood function or to apply the last equation to all quadruples of observations. Charbonneau (2017) argues that the latter is easier to implement, with a function to maximize:

$$\sum_{i=1}^{N} \sum_{j \neq i} \sum_{l,k \in Z_{ij}} \log \left(\frac{\exp(((x_{2,lj} - x_{2,lk}) - (x_{2,ij} - x_{2,ik}))'\beta_2^*)}{1 + \exp(((x_{2,lj} - x_{2,lk}) - (x_{2,ij} - x_{2,ik}))'\beta_2^*)} \right) \tag{32}$$

where Z_{ij} is the set of all potential k and l that satisfy: $y_{2,lk} + y_{2,lj} = 1, y_{2,ij} + y_{2,ik} = 1, y_{2,ij} + y_{2,lj} = 1$ for a given pair ij. If we would instead use the conditional maximum likelihood function, we would have that the sufficient statistics for the fixed effects are given by $\sum_{j=1}^{N} y_{2,ij}$, $\sum_{i=1}^{N} y_{2,ij}$.

Even though Charbonneau (2017) does not provide the asymptotic properties of this estimator, the later study by Jochmans (2018) provides such results. To demonstrate them, we first define some variables for an easier understanding:

$$z = \frac{(y_{2,lj} - y_{2,lk}) - (y_{2,ij} - y_{2,ik})}{2}$$

$$r = (x_{2,lj} - x_{2,lk}) - (x_{2,ij} - x_{2,ik})$$

Given that $y_{2,ij}$ for any ij is a binary variables, z can take values from the set $\{-1, -1/2, 0, 1/2, 1\}$. The event $z \in \{-1, 1\}$ corresponds to the condition that for any ij, l and k satisfies $y_{2,lk} + y_{2,lj} = 1, y_{2,ij} + y_{2,ik} = 1, y_{2,ij} + y_{2,lj} = 1$. From Equation (31) it follows that conditional on x_2 and $z \in \{-1, 1\}$, the distribution of z is logistic and does not depend on the

fixed effects. Note that when $z = 1$, we have necessarily that $y_{2,lk} = 1$, and when $z = -1$, it is necessarily zero.

Thus, the conditional log-likelihood of a given quadruple is:

$$\mathbb{1}\{z = 1\} \log(F(r'\beta_2^*)) + \mathbb{1}\{z = -1\} \log(1 - F(r'\beta_2^*))$$

where F is the standard logistic distribution. This indicates that the approach by Charbonneau (2017) boils down to a standard logistic regression with the transformed variables z and r, and considering only a subset of the sample. This can also be seen in the objective function $\mathcal{L}_{CL}$ defined later in this section. Conditioning on the event $z \in \{-1, 1\}$ means that not all quadruples will be used to estimate β_2^*, which is a possible drawback of this approach. Intuitively, conditioning on this event means that for an unit i, if a link is formed with another unit j, such that $y_{2,ij} = 1$, then a link must not be formed with the unit k, such that $y_{2,ik} = 0$. The same holds for the other units in the quadruples considered. In the terminology of Rasch (1960) those units are refered to as movers.

In the entire dataset, there are m_n distinct quadruples:

$$m_n = \binom{N}{2} \binom{N-2}{2} = \frac{N(N-1)(N-2)(N-3)}{4}$$

Jochmans (2018) introduces further a function σ that maps the possible quadruples in the dataset to an index set $N_{m_n} = \{1, 2, ... m_n\}$, such that it is possible to write the objective function in a comprehensible manner. Based on this, we extend the notation by defining the following random variables:

$$z(\sigma\{l, i; j, k\}) = \frac{(y_{2,lj} - y_{2,lk}) - (y_{2,ij} - y_{2,ik})}{2}$$

$$r(\sigma\{l, i; j, k\}) = (x_{2,lj} - x_{2,lk}) - (x_{2,ij} - x_{2,ik}))$$

With this extended notation, the estimator can be written as:

$$\hat{\beta}_2^* = \operatorname{argmax}_{\beta_2^* \in \Theta} \mathcal{L}_{CL}(\beta_2^*)$$

where Θ refers to the parameter space searched over, and $\mathcal{L}_{CL}$ is the objective function given by:

$$\mathcal{L}_{CL}(\beta_2^*) = \sum_{\sigma \in N_{m_n}} \mathbb{1}\{z(\sigma\{l, i; j, k\}) = 1\} \log(F(r(\sigma\{l, i; j, k\})'\beta_2^*))$$
$$+ \mathbb{1}\{z(\sigma\{l, i; j, k\}) = -1\} \log(1 - F(r(\sigma\{l, i; j, k\})'\beta_2^*))$$

Even though the objective function sums over all possible quadruples, only the quadruples that are in the set given by the event $z \in \{-1, 1\}$ are

incorporated into the sum. Therefore, we can denote that the number of such quadruples is given by:

$$m_n^* = \sum_{\sigma \in N_{mn}} \mathbb{1}\{z \in \{-1, 1\}\}$$

We can further write the expected fraction of quadruples in the dataset that contributes to the log-likelihood as a function of the random variable m_n^*:

$$p_n = \frac{\mathbb{E}(m_n^*)}{m_n} = \frac{\sum_{\sigma \in N_{mn}} \Pr\{z \in \{-1, 1\}\}}{m_n}$$

Now we turn to the assumptions needed for both the consistency and asymptotic normality and unbiasedness of the estimator, provided by Jochmans (2018).

ASSUMPTION 9. $\beta_{2,0}^*$ is interior to Θ, which is a compact subset of $\mathbb{R}^{\dim \beta_2^*}$.

ASSUMPTION 10. For all $(i, j) \in N \times N, \mathbb{E}(||x_{2,ij}||^2) < c$, where c is a finite constant.

ASSUMPTION 11. $Np_n \to \infty$ as $N \to \infty$ and the matrix

$$\lim_{N \to \infty} (m_n p_n)^{-1} \sum_{\sigma \in N_{mn}} \mathbb{E}(r(\sigma\{l, i; j, k\}) r(\sigma\{l, i; j, k\})' f(r(\sigma\{l, i; j, k\})' \beta_{2,0}^*) \mathbb{1}\{z \in \{-1, 1\}\})$$

has maximal rank, where f is the density of the logistic function.

Assumptions 9 and 10 are standard to establish consistency in nonlinear models. Assumption 11 is made to guarantee identifiability of the parameter. Note that it allows the expected fraction of quadruples to enter the log-likelihood, and be informative in this context, to go to zero as N grows. The requirement that p_n does not shrink faster than N^{-1} is needed for the uniform convergence of $\mathcal{L}_{CL}(\beta_2^*)$. As m_n is of order $O(N^4)$, this condition implies that $\mathbb{E}(m_n^*) = m_n p_n \to \infty$. Thus, even if the expected fraction of quadruples to enter the log-likelihood is allowed to go to zero as N grows, it is still needed that the accumulation of informative quadruples does not cease as N grows. In practice, this translates to the fact that even if a lower percentage of the quadruples enter the log-likelihood when new individuals are added to the dataset, the total number of observations considered in the estimation should continue to grow. The second part of Assumption 11 is a standard identification condition. Then, the following theorem arises:

THEOREM 2. *Let assumptions 8 - 11 hold. Then,* $\hat{\beta}_2^* \xrightarrow{p} \beta_{2,0}^*$ *as $N \to \infty$.*

The proof of this theorem is provided in the online Appendix of Jochmans (2018). In order to establish the limiting distribution of the estimator, a stronger form of Assumption 10 is made:

ASSUMPTION 12. For all $(i,j) \in N \times N$, $\mathbb{E}(||x_{2,ij}||^6) < c$, where c is a finite constant.

Defining:

$$s(\sigma; \beta_2^*) = r_\sigma \{\mathbb{1}\{z_\sigma = 1\}(1 - F(r_\sigma' \beta_2^*)) - \mathbb{1}\{z_\sigma = -1\}F(r_\sigma' \beta_2^*)\},$$

where $r_\sigma = r(\sigma\{l,i;j,k\})$ and $z_\sigma = z(\sigma\{l,i;j,k\})$. Then, further defining:

$$\upsilon_{ij}(\beta_2^*) = \sum_{i \neq l,j} \sum_{k \neq i,l,j} s(\sigma\{l,i;j,k\}; \beta_2^*)$$

$$\Upsilon(\beta_2^*) = \sum_{l=1}^{N} \sum_{j \neq l} \upsilon_{ij} \upsilon_{ij}'$$

Denoting S the score vector, we have that $\Upsilon(\beta_{2,0}^*)^{-1/2} S(\beta_{2,0}^*) \xrightarrow{d} N(0,I)$, where I is the identity matrix of dimension of β_2^*. The Hessian matrix is given by:

$$H(\beta_2^*) = \frac{\partial^2 \mathcal{L}_{CL}(\beta_2^*)}{\partial \beta_2^* \partial \beta_2^{*\prime}} = \sum_{\sigma \in N_{m_n}} r_\sigma r_\sigma' f(r_\sigma' \beta_2^*) \mathbb{1}\{z \in \{-1,1\}\}$$

Finally, defining:

$$\Omega = H(\hat{\beta}_2^*)^{-1} \Upsilon(\hat{\beta}_2^*) H(\hat{\beta}_2^*)^{-1}$$

Jochmans (2018) arrives at the following result:

THEOREM 3. *Let Assumptions 8 - 11 hold. Then $||\hat{\beta}_2^* - \beta_{2,0}^*|| = O_p(1/\sqrt{N(N-1)p_n})$ and*

$$\Omega^{-1/2}(\hat{\beta}_2^* - \beta_{2,0}^*) \xrightarrow{d} N(0,I)$$

as $N \to \infty$

The proof of this theorem is also provided in the online Appendix of Jochmans (2018). The intuition behind this result is relatively simple. Even though the objective function has the functional form of a standard logistic regression, the asymptotic variance of the estimator is not the usual textbook formula, where the information matrix equality holds, but a sandwich form of it. This is due to the fact that the transformed variables are functions of quadruples of units, with the same units appearing in several combinations of quadruples. Therefore, this introduces cross-sectional dependences that must be corrected.

There are also several aspects of network theory embedded in this model and respective estimators. When studying network formation, this framework allows to control for degree heterogeneity, which is given by the fixed-effects (that even though are not estimated, are controlled for). While previous studies that relied on asymptotic bias corrections (for instance, Dzemski (2019)) allowed it only for dense networks, this framework allows it also for sparse networks. This difference comes from the fact that in asymptotic bias corrections approaches, the fixed-effects need to be estimated. In the case of sparse networks, those estimates may not be consistent or converge only at a very slow rate. On the other hand, in this method, the fixed-effects need not be estimated, and the probability of link formation is allowed to shrink towards zero.

This estimation method also permits to test for homophily in network formation, once, given the limiting distribution, it is possible to conduct inference on variables that reflect how similar the characteristics of units i and j in a dyad are.

The drawback of this approach is that the independence of the error terms in Assumption 8 results in links being conditionally independently formed. Therefore, it is a not well-suited estimator in cases where the link decisions made by a given unit (node) is dependent on the decisions of other units. Evidently, it also rules out cases where a high degree of transitivity is present in the network [6].

In the next sections, we will provide two approaches to estimate the parameters of the observation Equation (9) given these estimates for the selection equation, denoted from now on $\beta^*_{2,CL}$. The first approach relies on retrieving estimates of the fixed-effects in the selection equation, and applying the Heckman methodolody to control for the sample selection. The second approach relies on differencing out both the fixed-effects and the regressor that corrects for the sample selection bias in the observation equation.

8   The first approach: Hybrid estimates and Lee's transformation

In this section we introduce a method to retrieve the estimates for the fixed-effects in Equation (11) once consistent and asymptotically unbiased estimates of the parameter β^*_2 are given. In this context, we use the estimates that the previously outlined approach delivers, given by $\hat{\beta}^*_{2,CL}$.

Once the fixed effects estimates are available, we then introduce the method described by Lee (1983). It essentially consists of the standard

---

[6]A high degree of transitivity arises in situations where two units are more likely to be linked if there is a higher overlap between the sets of units that they are already linked to Graham (2020).

Heckman approach, with the only difference being that some transformations are made to the selection equation, once estimators are now coming from a logistic regression instead of a probit.

8.1 The hybrid approach of Wooldridge et al. (2018)

The conditional logit approach defined above delivers $\sqrt{N(N-1)p_n}$-consistent and asymptotically unbiased estimators of the parameters β_2^* as long as the covariates are strictly exogenous. However, this approach alone does not yield average partial effects estimates or the predicted probabilities of the outcome $y_{2,ij,t}$ as the distribution of the unobserved heterogeneity is not specified such that it can be integrated out. In our setting, the predicted probabilities play a role in the correction for sample selection bias, and therefore, are needed. Such estimates can be obtained with the conditional logit approach by inserting as well an estimate of the unobserved fixed effects. This hybrid approach consists of obtaining such estimates.

According to Wooldridge et al. (2018), the fixed effects ξ^* and ζ^* can be estimated from the unconditional maximum likelihood when the slope parameters are restricted to be equal to their conditional logit estimates. Then, the estimated slope parameters and the first stage estimated fixed effects are combined to form both APE estimates and predicted probabilities.

We can obtain the estimates of the fixed effects from:

$$\hat{\omega}_{NN}^*(\beta_2^*) = \mathrm{argmax}_{\omega_{NN}^* \in \mathbb{R}^{\dim \omega_{NN}^*}} \mathcal{L}(\hat{\beta}_{2,CL}^*, \omega_{NN}^*)$$

Where the unconditional log-likelihood $\mathcal{L}$ is given by Equations (22) and (20). The adjustment here is that we consider a single period t, and according to Assumption 8, F is now a standard logistic distribution function. Also, likewise in Section 6, we collect all the fixed effects to be estimated in the vector $\omega_{NN}^* = (\xi_1^*, ... \xi_N^*, \zeta_1^*, ... \zeta_N^*)'$.

This approach was also proposed in other papers such as Bartolucci and Pigini (2019). However, in both papers (Wooldridge et al. (2018) and Bartolucci and Pigini (2019)), additional bias corrections are proposed, given that in their setting one of the dimensions of the panel data is fixed (and therefore, fixed effects estimates are inconsistent). In our framework, as both dimensions go to infinity, it is expected that given an asymptotically unbiased and consistent estimate of the structural parameters, the fixed effects should also share those properties.

When retrieving the fixed effects, we also take into account only units for which the outcome presents variability. This means that, for a given fixed index i, there are $N-1$ observations available such that i refers to the first index in the dyad, i.e. y_{ij} for all $j \neq i$. Those $N-1$ observations are only taken into account if y_{ij} varies over the different indices j. The other way round is also taken into account, i.e., for fixing an index j and considering

the $N-1$ observations y_{ij} for all $i \neq j$. This is done because, as explained by Kunz, Staub, and Winkelmann (2017), the fixed effects related to the fixed index do not exist in cases where there is no variability in the outcome, a problem known in the literature as the perfect prediction problem (Maddala (1983)). As it is demonstrated in Kunz et al. (2017), the perfect prediction means that for such observations, the maximum-likelihood estimator for the fixed effects will not have a finite solution. This will lead to fixed effects tending to minus infinity or plus infinity. This problem is formally illustrated through a simple example in Appendix A.4.

Note that this hybrid framework is not well-suited for sparse networks, once we are proposing to estimate the fixed effects.

Finally, for now we will focus on analysing only the finite sample properties of the estimates given by this hybrid approach through a simulation exercise. While in frameworks where only one of the fixed effects in Equation (11) is present relies on partitioning the log-likelihood in the sum of the individual log-likelihoods that depend on the structural parameters and on the fixed effect for a given individual, in this case, this strategy is not feasible, since no partition of the dataset depends on only one fixed effect. Therefore, a more careful analysis of the asymptotic properties is left for further research.

8.2 Sample selection estimator based on Lee (1983)

Once the estimates for both β_2^* and the fixed effects ω_{NN}^* are obtained, it is possible to take into account in the estimation of Equation (9) the sample selection, even if we use a logistic model for the first stage of the Heckman (1979) approach. The method is delineated in Lee (1983).

The idea is, rewriting Equation (9) as:

$$y_{1,ij,t} = x'_{1,ij}\beta_1 + \vartheta_i + \chi_j + \sigma_u u_{ij}^* \tag{33}$$

where $u_{ij}^* \sim N(0,1)$ and $\sigma_u > 0$. Considering the selection Equation (11), under Assumptions 4, 5, 8, and assuming that $x_1, x_2, \vartheta_i, \chi_j, \xi_i^*, \zeta_j^*$ are exogenous variables, which essentially already follows from previously defined assumptions:

ASSUMPTION 13. The errors are such that, given the covariates x_1, x_2 and the fixed effects:

$$\mathbb{E}(u^*|x_1, x_2, \vartheta_i, \chi_j, \xi_i^*, \zeta_j^*) = 0$$
$$\mathbb{E}(\eta^*|x_1, x_2, \vartheta_i, \chi_j, \xi_i^*, \zeta_j^*) = 0$$
$$\mathrm{Var}(\eta^*|x_1, x_2, \vartheta_i, \chi_j, \xi_i^*, \zeta_j^*) = 1$$

We now denote that the disturbances u^* and η^* conditional on x_1, x_2, ϑ_i, $\chi_j, \xi_i^*, \zeta_j^*$ have continuous distribution functions $\Phi(u^*)$ and $F(\eta^*)$, that are

29

completely specified, according to the previous assumptions, as a standard normal distrubution and a standard logistic distribution, respectively.

As denoted before already, the dependent variable $y_{1,ij}$ conditional on $x_1, x_2, \vartheta_i, \chi_j, \xi_i^*, \zeta_j^*$ has a well defined marginal distribution but it is not observed unless $y_{2,ij}^{**} \geq 0$. The observed samples of $y_{1,ij}$ are thus censored and follow $y_{1,ij} = x_{1,ij}'\beta_1 + \vartheta_i + \chi_j + \sigma_u u_{ij}^*$ if and only if $x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^* \geq -\eta_{ij}^*$.

The distributions of u^* and η^* are allowed to be correlated (supposing a correlation of ρ, as denoted before in Section 5). Moreover, only the marginal distributions are specified but not the joint bivariate distribution of u^* and η^*. Therefore, the idea of Lee's transformation is to suggest a proper bivariate distribution that can be applied to the Heckman sample selection method with the specified marginal distributions.

We start by applying the following transformation to the error term:

$$\eta^{**} = J(\eta^*) = \Phi^{-1}(F(\eta^*))$$

This transformation guarantees that the transformed variables is distributed as a standard normal with zero mean and unit variance. A bivariate distribution having the required marginal distributions $F(\eta^*)$ of η^* and $\Phi(u^*)$ of u^* can be specified, from the following assumption.

ASSUMPTION 14. The transformed random variables η^{**} and u^* are jointly normally distributed with zero means, unit variances and correlation of ρ.

Denoting by $B(\cdot, \cdot, \rho)$ a bivariate normal distribution $N(0, 0, 1, 1, \rho)$, a proper bivariate distribution of (η^*, u^*), denoted by H, is derived such that:

$$H(\eta^*, u^*, \rho) = B(J(\eta^*), u^*; \rho)$$

From the model specification, recall that $y_{2,ij} = 1$ if and only if $x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^* \geq -\eta_{ij}^*$. Given the logistic distribution, which is absolutely continuous, $F(\cdot)$, the transformation $J(\cdot) = \Phi^{-1}(F(\cdot))$ is strictly increasing, such that $y_{2,ij} = 1$ if and only if $J(x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^*) \geq J(-\eta_{ij}^*)$, or, equivalently, $J(-x_{2,ij}'\beta_2^* - \xi_i^* - \zeta_j^*) < J(\eta_{ij}^*)$

Therefore, the previously censored regression model given by Equations (33) and (11), with given standard normal and logistic marginal distributions $\Phi(u^*)$ and $F(\eta^*)$ and the previously defined bivariate distribution, is statistically equivalent to the model:

$$y_{1,ij,t} = x_{1,ij}'\beta_1 + \vartheta_i + \chi_j + \sigma_u u_{ij}^* \tag{34}$$

$$y_{2,ij}^{***} = J(x_{2,ij}'\beta_2^* + \xi_i^* + \zeta_j^*) + \eta_{ij}^{**} \tag{35}$$

where $y_{2,ij}^{***} = J(y_{2,ij}^{**})$.

We are now in a setting where the standard Heckman approach from Section 5 can be applied. It follows from this specification that the inverse Mills-ratio is now of the form:

$$y_{2,ij}^{***} = J(x_{2,ij}'\beta_2^{*'} + \xi_i^* + \zeta_j^*) + \eta_{ij}^{**} \tag{36}$$

Defining as before, $z_{ij} = -x_{2,ij}'\beta_2^* - \xi_i^* - \zeta_j^*$, we have that the inverse Mills-ratio can be written now as:

$$\lambda_{ij}(z_{ij}) = \frac{\phi(J(z_{ij}))}{1-\Phi(J(z_{ij}))} = \frac{\phi(J(z_{ij}))}{1-\Phi(\Phi^{-1}(F(z_{ij})))} = \frac{\phi(J(z_{ij}))}{1-F(z_{ij})} = \frac{\phi(J(z_{ij}))}{F(-z_{ij})}$$

where for the last equality we used the symmetry of the logistic distribution. Given the form of the inverse Mills-ratio one can correct Equation (33) for the sample selection bias by including the term:

$$y_{1,ij} = x_{1,ij}'\beta_1 + \vartheta_i + \chi_j + (\sigma_u\rho)\lambda_{ij}(z_{ij}) + \nu_{ij} \tag{37}$$

Again, by construction, $\mathbb{E}[\nu_{ij}|x_{1,ij}, \vartheta_i, \chi_j, \lambda_{ij}(z_{ij}), y_{2,ij}^{***} > 0] = 0$. Thus, as before, we can include the inverse Mills-ratio as a regressor in the equation, and treat $(\sigma_u\rho)$ as a parameter to be estimated.

If we had the limiting distribution of the fixed effects and thus, the joint distribution of the fixed effects and the structural parameters of (36), the asymptotic properties of the parameters of this regression would follow easily from the proof of the limiting distribution of the Heckman approach provided in Heckman (1979). The only adjustment to be made is to account for the fixed effects ϑ_i and χ_j in the derivation. We expect that, once the estimates $\hat{\beta}_{2,CL}^*$ are not asymptotically biased, the estimates for the fixed effects of the selection equation provided by the hybrid approach not to be asymptotically biased as well. Moreover, they should also be consistent once we have that $(N-1) \to \infty$. Therefore, the estimated inverse-Mills ratio should satisfy Theorem 1, even if they will likely converge only at a slower rate than $\sqrt{N(N-1)}$, as the fixed effects converge at a slower rate.

9 Second approach: a modification to Kyriazidou (1997)

9.1 An outline of the approach proposed in Kyriazidou (1997)

To outline the approach proposed by Kyriazidou (1997) to estimate Equation (9) correcting for a possible sample selection bias, for now we assume that we are back in a setting where $T \geq 2$. As the estimation proposed here relies on a consistent estimator[7] for β_2^*, we will take into account the estimator

---

[7]In her paper, Kyriazidou (1997) imposes that only a consistent estimator for the first stage is needed. Therefore, in theory one could even estimate it by the unconditional MLE probit in this setting.

defined in section 5, denoted by $\hat{\beta}^*_{2,CL}$. We will then impose the strong assumption for now:

ASSUMPTION 15. The true value $\beta^*_{2,0}$ does not vary over time.

Even if this is a strong assumption, our specified model given by Equations (9) - (11) accomodates it. In a general framework, we usually have that the exogenous regressors $x_{2,ij,t}$ in Equation (11) are time-varying. However, suppose that they were constant in time, or that a linear combination of those variables, up to a finite number of estimable parameters β^*_2 is constant over time ($\beta^{*\prime}_2 x_{2,ij,1} = \beta^{*\prime}_2 x_{2,ij,2}$). In this scenario, for a dyad ij observed in two consecutive periods, it is expected that the magnitude of the selection effect in the observation equation to be the same in both periods. Therefore time differencing on the Equation (9) would eliminate both the fixed effects ϑ_i and χ_j and the sample selection effect. While this is unfeasible in most settings, as the sample selection effect is most likely to vary over time, this remark is the basis for the estimator proposed in this section.

We will follow a similar idea as the standard Heckman approach for sample selection. In the first step, the parameters β^*_2 are estimated by the approach defined by Charbonneau (2017), yielding $\hat{\beta}^*_{2,CL}$. In the second step, those estimates are used to estimate the time differenced observation equation by a weighted least squares regression. The weights given to each differenced dyad is such that, given the estimates $\hat{\beta}^*_{2,CL}$, observations for which the difference in the sample selection effects approaches zero, a higher weight is assigned, and vice-versa. Therefore, by considering that both the observation equation is differenced and that the weights depend on the difference of the sample selection effects, all fixed effects in the model given by Equations (9) - (11) are differenced out, and their estimates are no longer needed. [8]

Besides the fact that estimators for the fixed effects are no longer needed, one key advantage of this approach compared to the standard Heckman approach is that while the latter requires a full specification of the distribution of the error terms, in general, this approach allows for distributions of both error terms to be unspecified, provided that a semiparametric estimator for the first stage is available. However, in our specified framework, we maintain the distributional assumptions of the error terms given by Assumptions 4, 5 and 8, and that such errors are independent of the fixed effects and the explanatory variables, given by Assumption 13. We will provide a more rigorous treatment of the arguments presented up to now and the proposed estimator.

In the following we will denote the vector of explanatory variables for

---

[8]Therefore, in a standard panel data regression with fixed T, the fact that a possible estimator for the fixed effects in the selection equation is inconsistent does not harm this approach.

both time periods $t = 1, 2$ and fixed effects by $\varsigma_{ij} \equiv (x_{1,ij,1}, x_{1,ij,2}, x_{2,ij,1}, x_{2,ij,2},$
$\xi_i^*, \zeta_j^*, \vartheta_i, \chi_j)$.

The fixed-effects in the observation Equation (9) could easily be differenced out (timewise) for equations that are observed in both periods (i.e., $y_{2,ij,1} = y_{2,ij,2} = 1$). However, the sample selection problem remains, as we can see from the populational regression:

$$\mathbb{E}(y_{1,ij,1} - y_{1,ij,2}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij}) \tag{38}$$
$$= (x_{1,ij,1} - x_{1,ij,2})'\beta_1 + \mathbb{E}(u_{ij,1} - u_{ij,2}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij})$$

In a general framework, and also in our framework, there is no reason to expect that the last term is zero, which would be given by either $\mathbb{E}(u_{ij,1}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij}) = \mathbb{E}(u_{ij,2}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij}) = 0$, or simply $\mathbb{E}(u_{ij,1}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij}) = \mathbb{E}(u_{ij,2}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij})$.

For ease of exposition note that for each time period, given that an observation is observed in both $t = 1, 2$:

$$\mathbb{E}(y_{1,ij,t}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij}) \tag{39}$$
$$= x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + \mathbb{E}(u_{ij,t}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij})$$

Using the same definition as in Section 5, we can assume that the conditional expectation in the last term of Equation (39) is a smooth function $\varphi_{ij,t}$ that depends on ς_{ij} and the joint conditional distribution of the error terms, denoted by $H_{ij,t}(u_{ij,t}, \eta^*_{ij,1}, \eta^*_{ij,2}|\varsigma_{ij})$.

$$\varphi_{ij,t} \equiv \mathbb{E}(u_{ij,t}|y_{2,ij,1} = y_{2,ij,2} = 1, \varsigma_{ij}) \tag{40}$$
$$= \mathbb{E}(u_{ij,t}|\eta^*_{ij,1} \le x'_{2,ij,1}\beta_2^* + \xi_i^* + \zeta_j^*, \eta^*_{ij,2} \le x'_{2,ij,2}\beta_2^* + \xi_i^* + \zeta_j^*, \varsigma_{ij})$$
$$= \varphi(x'_{2,ij,1}\beta_2^* + \xi_i^* + \zeta_j^*, '_{2,ij,2}\beta_2^* + \xi_i^* + \zeta_j^*; H_{ij,t}(u_{ij,t}, \eta^*_{ij,1}, \eta^*_{ij,2}|\varsigma_{ij}))$$
$$= \boldsymbol{\varphi_{ij,t}}(x'_{2,ij,1}\beta_2^* + \xi_i^* + \zeta_j^*, '_{2,ij,2}\beta_2^* + \xi_i^* + \zeta_j^*; \varsigma_{ij})$$

As we are considering that the errors are independent and identically distributed over dyads ij and over time, and that they are independent of ς_{ij}, this further reduces to:

$$\varphi_{ij,t} = \mathbb{E}(u_{ij,t}|y_{2,ij,t} = 1) \tag{41}$$
$$= \mathbb{E}(u_{ij,t}|\eta^*_{ij,t} \le x'_{2,ij,t}\beta_2^* + \xi_i^* + \zeta_j^*)$$
$$= \varphi(x'_{2,ij,t}\beta_2^* + \xi_i^* + \zeta_j^*)$$

This translates to the smooth function given by $\boldsymbol{\varphi_{ij,t}}$, which is used to construct the inverse Mills-ratio as a function of the single index $x'_{2,ij,t}\beta_2^* + \xi_i^* + \zeta_j^*$, becoming invariant over dyads and time. Moreover, this function depends on the joint distribution of the errors. If Assumption 14 is imposed and the according transformations of Section 8.2 are made in the errors η^*, the function takes the form $\varphi(x) = \frac{\phi(J(x))}{F(-x)}$ with J defined as in Section 8.2,

and F the standard logistic distribution. However, this assumption is not needed and neither are the transformations since the functional form of φ is not needed in this approach as we will see later.

We can then rewrite the observation Equation (9) with the sample selection correction as:

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + \varphi_{ij,t} + \nu_{ij,t} \tag{42}$$

where, again, $\nu_{ij,t} = u_{ij,t} - \varphi_{ij,t}$. Therefore, by construction it satisfies that $\mathbb{E}(\nu_{ij,t}|y_{2,ij,1} = y_{2,ij,1} = 1, \varsigma) = 0$.

As mentioned before, the main idea would be to difference out in Equation (42) both the fixed effects ϑ_i and χ_j and the sample selection effect $\varphi_{ij,t}$. However, as pointed out before, in general, $\varphi_{ij,1} \neq \varphi_{ij,2}$ in the case that there is sample selection in the model. The equality would only hold if $x'_{2,ij,1}\beta_2^* = x'_{2,ij,2}\beta_2^*$. If that was the case for all the individuals observed in $t = 1, 2$ one could simply employ a first difference in the equation. In this case the possibility of differencing out the equation for consistent estimation would hold under weaker distributional assumptions: since the first differences over time is taken in the dyad level, the errors are not required to be i.i.d. across dyads and nor to be independent of ς_{ij}. Then, the functional form of $\varphi_{ij,t}$ would be allowed to vary across individuals.

In this scenario of a more flexible form of ς_{ij}, it would be sufficient that the following assumption holds:

ASSUMPTION 16. Consider that $(\eta^*_{ij,1}, \eta^*_{ij,2}, u_{ij,1}, u_{ij,2})$ and $(\eta^*_{ij,2}, \eta^*_{ij,1}, u_{ij,2}, u_{ij,1})$ are identically distributed conditional on ς_{ij}, i.e.,

$$H(\eta^*_{ij,1}, \eta^*_{ij,2}, u_{ij,1}, u_{ij,2}) = H(\eta^*_{ij,2}, \eta^*_{ij,1}, u_{ij,2}, u_{ij,1})$$

Then, for an individual that has $x'_{2,ij,1}\beta_2^* = x'_{2,ij,2}\beta_2^*$:

$$\begin{aligned}
\varphi_{ij,1} &= \mathbb{E}(u_{ij,1}|\eta^*_{ij,1} \leq x'_{2,ij,1}\beta_2^* + \xi_i^* + \zeta_j^*, \eta^*_{ij,2} \leq x'_{2,ij,2}\beta_2^* + \xi_i^* + \zeta_j^*, \varsigma_{ij}) \\
&= \mathbb{E}(u_{ij,2}|\eta^*_{ij,2} \leq x'_{2,ij,1}\beta_2^* + \xi_i^* + \zeta_j^*, \eta^*_{ij,1} \leq x'_{2,ij,2}\beta_2^* + \xi_i^* + \zeta_j^*, \varsigma_{ij}) \\
&= \varphi_{ij,2}
\end{aligned}$$

Thus, differencing would still be feasible under these weaker distributional assumptions. Defining:

$$\Psi_{ij} = \mathbb{1}(x'_{2,ij,1}\beta_2^* = x'_{2,ij,2}\beta_2^*)$$

$$\Phi_{ij} = \mathbb{1}(y_{2,ij,1} = y_{2,ij,2} = 1)$$

Denoting Δ as the first differences operator, an OLS estimator for this case would be:

$$\breve{\beta}_1 = \Big[\sum_{i=1}^N \sum_{j \neq i} \Delta x'_{1,ij} \Delta x_{1,ij} \Psi_{ij} \Phi_{ij}\Big]^{-1} \Big[\sum_{i=1}^N \sum_{j \neq i} \Delta x'_{1,ij} \Delta y_{1,ij} \Psi_{ij} \Phi_{ij}\Big]$$

Under appropriate regularity conditions, this estimator would be consistent and asymptotically unbiased. However, in most cases we will likely have that $\Psi_{ij} = 0$ for all dyads ij. The approach then is to assign weights for observations according to how close to zero $x'_{2,ij,1}\beta_2^* - x'_{2,ij,2}\beta_2^*$ is. If φ is a smooth function and provided that our estimator $\hat{\beta}_{2,CL}^*$ is consistent, observations that have this difference close to zero will also have that $\Delta\varphi_{ij}$ is close to zero, and therefore, the sample selection effects will be close to be differenced out.

Therefore, given this argument, we can propose the estimator:

$$\hat{\beta}_1 = \Big[\sum_{i=1}^{N} \sum_{j \neq i} \hat{\Psi}_{ij} \Delta x'_{1,ij} \Delta x_{1,ij} \Phi_{ij} \Big]^{-1} \Big[\sum_{i=1}^{N} \sum_{j \neq i} \hat{\Psi}_{ij} \Delta x'_{1,ij} \Delta y_{1,ij} \Phi_{ij} \Big] \quad (43)$$

where $\hat{\Psi}_{ij}$ is an estimated weight, based on $\hat{\beta}_{2,CL}^*$, that declines to zero as the magnitude of $|x'_{2,ij,1}\hat{\beta}_{2,CL}^* - x'_{2,ij,2}\hat{\beta}_{2,CL}^*|$ increases. Thus, we guarantee that a higher weight is attributed to observations where the sample selection effects almost cancel out. More precisely, these weights are given by:

$$\hat{\Psi}_{ij} = \frac{1}{h_n} K\Big(\frac{\Delta x'_{2,ij}\hat{\beta}_{2,CL}^*}{h_n}\Big) \quad (44)$$

where K is a kernel density function, and h_n is a sequence of bandwidths which tends to zero as $N(N-1) \to \infty$. For a fixed magnitude of the difference $|x'_{2,ij,1}\hat{\beta}_{2,CL}^* - x'_{2,ij,2}\hat{\beta}_{2,CL}^*|$, the weight $\hat{\Psi}_{ij}$ shrinks as the sample size increases.

Finally, note that this approach requires an exclusion restriction on the set of regressors, i.e., that at least one of the variables in the vector x_2 is not contained in the vector x_1. This is due to that, in cases where the differences $|x'_{2,ij,1}\hat{\beta}_{2,CL}^* - x'_{2,ij,2}\hat{\beta}_{2,CL}^*|$ shrinks to zero, if $x_1 = x_2$, the regressors in the observation equation will also shrink and the usual rank conditions for the weighted least squares will not hold.

9.2 The choice of the bandwidth h_n

To compute the estimator, a kernel function K and a bandwidth parameter h_n need to be chosen. Kyriazidou (1997) states that the asymptotic performance of the estimator is more influenced by the choice of the bandwidth than by the choice of the kernel. Therefore, she focuses on this choice, setting in most of her simulations the kernel to be a standard normal density function. The choice of h_n relies on the asymptotic results presented in the paper, and intuitively consists on choosing it such that a measure of distance between the estimator and the parameters true values is minimized. For brevity, we now present only the main results and intuitive explanations. For a more formalized view, refer to Kyriazidou (1997).

Using arguments provided by Bierens (1987), Kyriazidou (1997) shows that for a given order of differentiability r of the expression $\mathbb{E}(\Delta x_1' \boldsymbol{\varphi} \Phi | \Delta x_2' \beta_2^*)$ and a given sample size $N(N-1)$, $h_n = h(N(N-1))^{-\mu}$ should be chosen such that $\mu = 1/(2(r+1)+1)$. This result follows from the rate of convergence of the distribution of the estimator $\hat{\beta}_1$, which is maximized by setting μ as small as possible. However, μ should be contained in the range $1 - 2p < \mu < p/2$, where p is the rate of convergence of the estimator for β_2^* for the asymptotic results derived by the author to hold. The estimator for β_2^* should also converge fast enough for the results to hold, namely, at least at a rate $p > (r+1)/(2(r+1)+1)$. Therefore, combining the inequalities for p and μ, we arrive at the result that the minimum μ is achieved at $\mu = 1/(2(r+1)+1)$. For this chosen parameter, the estimator $\hat{\beta}_1$ will converge at a rate $(N(N-1))^{-(r+1)/(2(r+1)-1)}$.

Thus, the problem of choosing a bandwith boils down to choosing a constant h. Kyriazidou (1997) shows that for any positive initially chosen h, the distance between the estimator and the parameters' true values is minimized through a correction given by the following corollary:

COROLLARY 1. *Let $\hat{\beta}_1$ be the estimator with window width $h_n = h(N(N-1))^{-1/(2(r+1)+1)}$, and $\hat{\beta}_{1,\delta}$ the estimator with window width $h_{n,\delta} = h(N(N-1))^{-\delta/(2(r+1)+1)}$ where $\delta \in (0,1)$.*
Define:
$$\hat{\hat{\beta}}_1 = \frac{\hat{\beta}_1 - N(N-1)^{-(1-\delta)(r+1)/(2(r+1)+1)}\hat{\beta}_{1,\delta}}{1 - N(N-1)^{-(1-\delta)(r+1)/(2(r+1)+1)}}$$

Then, $(\hat{\hat{\beta}}_1 - \beta_1)$ will converge to a normal distribution centered around 0 at rate $(N(N-1))^{-(r+1)/(2(r+1)-1)}$.

She shows that this Corollary results in obtaining an optimal bandwidth h^*, given an initial chosen value of h. While h^* can be pinned down and it is necessary for the estimates of the asymptotic variance, one does not need to pin down its value for obtaining the asymptotically corrected estimates of β_1 (however the estimates of β_1 and the value of h^* are tied down).

The steps to implement this estimator are:

- Step 1: For a given r and $N(N-1)$, choose any $h_n = h(N(N-1))^{-1/(2(r+1)+1)}$ and any $h_{n,\delta} = h(N(N-1))^{-\delta/(2(r+1)+1)}$ with an arbitrary positive h and $0 < \delta < 1$.

- Step 2: Compute both $\hat{\beta}_1$ and $\hat{\beta}_{1,\delta}$.

- Step 3: From the Corollary above, obtain the asymptotically unbiased estimator $\hat{\hat{\beta}}_1$.

Another interesting result obtained in Kyriazidou (1997) is that given an appropriately chosen h_n as mentioned above, using an estimated value for β_2^*

does not affect the limiting distribution of the estimator for β_1. This follows from the lower bound imposed on μ, that guarantees that β_1 is estimated at a lower rate than β_2^*.

More formally, defining the scalar index $X_{2,ij} = \Delta x_{2,ij}\beta_2^*$, its estimated counterpart $\hat{X}_{2,ij} = \Delta x_{2,ij}\hat{\beta}_{2,CL}^*$ and the following quantities:

$$S_{xx} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{h_n}K\Big(\frac{X_{2,ij}}{h_n}\Big)\Delta x'_{1,ij}\Delta x_{1,ij}\Phi_{ij}$$

$$\hat{S}_{xx} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{h_n}K\Big(\frac{\hat{X}_{2,ij}}{h_n}\Big)\Delta x'_{1,ij}\Delta x_{1,ij}\Phi_{ij}$$

$$S_{x\nu} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{h_n}K\Big(\frac{X_{2,ij}}{h_n}\Big)\Delta x'_{1,ij}\Delta \nu_{ij}\Phi_{ij}$$

$$\hat{S}_{x\nu} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{h_n}K\Big(\frac{\hat{X}_{2,ij}}{h_n}\Big)\Delta x'_{1,ij}\Delta \nu_{ij}\Phi_{ij}$$

$$S_{x\varphi} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{h_n}K\Big(\frac{X_{2,ij}}{h_n}\Big)\Delta x'_{1,ij}\Delta \varphi_{ij}\Phi_{ij}$$

$$\hat{S}_{x\varphi} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{h_n}K\Big(\frac{\hat{X}_{2,ij}}{h_n}\Big)\Delta x'_{1,ij}\Delta \varphi_{ij}\Phi_{ij}$$

By choosing an appropriate h_n, Kyriazidou (1997) shows that the quantities $\hat{S}_{xx}$, $\hat{S}_{x\nu}$ and $\hat{S}_{x\varphi}$ have the same probability limits as their infeasible counterparts S_{xx}, $S_{x\nu}$ and $S_{x\varphi}$ (which consider β_2^* instead of its estimator).

Then, (i) by taking the first differences of Equation (42) and premultiplying it by $\hat{\Phi}_{ij}\Delta x'_{1,ij}$, (ii) considering the estimator $\hat{\beta}_1$, and (iii) defining $\tilde{\beta}_1$ to be the unfeasible estimator (i.e., evaluated at the true β_2^*), we can easily write:

$$\tilde{\beta}_1 - \beta_1 = S_{xx}^{-1}(S_{x\nu} + S_{x\varphi})$$
$$\hat{\beta}_1 - \beta_1 = \hat{S}_{xx}^{-1}(\hat{S}_{x\nu} + \hat{S}_{x\varphi})$$

Therefore, it follows that if the quantities $\hat{S}_{xx}$, $\hat{S}_{x\nu}$ and $\hat{S}_{x\varphi}$ have the same probability limits as their infeasible counterparts, the estimators $\tilde{\beta}_1$ and $\hat{\beta}_1$ will also have the same limiting distributions.

9.3 A modification to the approach

Even though in most settings we have that the regressors $x_{1,ij,t}$ vary over time, in particular cases such as in trade applications, most of the explanatory variables in the observation equation do not have any variability over time. For instance, traditional variables used in trade applications are the distance between countries, whether they share the same currency, or have the same official language, etc. This implies that, in the delineated approach by Kyriazidou (1997), the coefficients of such regressors are not estimable, since the method relies on time differencing.

We propose in this section a modification to the previous method, that is feasible due to the dyadic structure of our data. Instead of differencing over time, we propose that the variables are differenced over dyads, such that, fixing a single time period t, and for a quadruple given by the indices i, j, k and l:

$$\Delta_{ijlk} y_{1,ij} = (y_{1,ij} - y_{1,ik}) - (y_{1,lj} - y_{1,lk})$$

Given this definition for the operator Δ_{ijlk}, it is easy to see from Equation (9) that such method would difference out the fixed effects ϑ_i and χ_j. By redefining the vector ς_{ij} to collect the explanatory variables for the quadruple ij, ik, lj, lk, and their respective fixed effects, yielding:

$$\varsigma_{ijlk} \equiv (x_{1,ij}, x_{2,ij}, x_{1,ik}, x_{2,ik}, x_{1,lj}, x_{2,lj}, x_{1,lk}, x_{2,lk}, \xi_i^*, \xi_l^*, \zeta_j^*, \zeta_k^*, \vartheta_i, \vartheta_l, \chi_j, \chi_k)$$

It follows that:

$$\mathbb{E}[\Delta_{ijlk} y_{1,ij} | y_{2,ij} = y_{2,ik} = y_{2,lj} = y_{2,lk} = 1, \varsigma_{ijlk}]$$
$$= \Delta_{ijlk} x_{1,ij}' \beta_1 + \mathbb{E}(\Delta_{ijlk} u_{1,ij} | y_{2,ij} = y_{2,ik} = y_{2,lj} = y_{2,lk} = 1, \varsigma_{ijlk})$$

Based on the previously outlined baseline approach of Kyriazidou (1997), if we then define:

$$\varphi_{ijlk} = \mathbb{E}(u_{ij} | y_{2,ij} = y_{2,ik} = y_{2,lj} = y_{2,lk} = 1, \varsigma_{ijlk})$$

We can write the equivalent of Equation (42) for this modification:

$$y_{1,ij} = \beta_1' x_{1,ij} + \vartheta_i + \chi_j + \varphi_{ijlk} + \nu_{ij}$$

For a given dyad ij there is more than one associated quadruple, as there are several combinations of l and k (the restriction being that all indices must be different). However, for all such quadruples, it follows that, as before, $\nu_{ij} = u_{ij} - \varphi_{ijlk}$. Therefore, by construction it is sufficient that $\mathbb{E}(\nu_{ij} | y_{2,ij} = y_{2,ik} = y_{2,lj} = y_{2,lk} = 1, \varsigma_{ijlk}) = 0$.

Moreover, under the Assumptions 4, 5 and 8, it follows that the equivalent of the conditional exchangeability Assumption 16 for this modification is satisfied.

Therefore, following the same logic as in the previous subsection, we can construct an estimator for β_1 such that:

$$\hat{\beta}_1 = \Big[\sum_{i=1}^{N} \sum_{j \neq i} \sum_{k \neq i,j} \sum_{l \neq i,j,k} \hat{\Psi}_{ijkl} \Delta_{ijkl} x'_{1,ij} \Delta_{ijkl} x_{1,ij} \Phi_{ijkl} \Big]^{-1} \qquad (45)$$

$$\Big[\sum_{i=1}^{N} \sum_{j \neq i} \sum_{k \neq i,j} \sum_{l \neq i,j,k} \hat{\Psi}_{ijkl} \Delta_{ijkl} x'_{1,ij} \Delta_{ijkl} y_{1,ijkl} \Phi_{ijkl} \Big]$$

Accordingly, Φ_{ijkl} is defined as: $\Phi_{ijkl} = \mathbb{1}(y_{2,ij} = y_{2,ik} = y_{2,lj} = y_{2,lk} = 1)$. Also in this modification, $\hat{\Psi}_{ijkl}$ is an estimated weight, based on $\hat{\beta}^*_{2,CL}$ that declines to zero as the magnitude of $|\Delta_{ijkl} x'_{2,ij} \hat{\beta}^*_{2,CL}|$ increases. Thus, we continue to guarantee that a higher weight is attributed to observations where the sample selection effects almost cancel out. More precisely, in this modification these weights are given by:

$$\hat{\Psi}_{ijkl} = \frac{1}{h_n} K \Big(\frac{\Delta_{ijkl} x'_{2,ij} \hat{\beta}^*_{2,CL}}{h_n} \Big) \qquad (46)$$

where K is a kernel density function, and h_n is a sequence of bandwidths. The remaining of the modification follows from the previous subsection.

The potential drawbacks of this method are that, as we need that all quadruples need to be observed for the differentiation to hold, the approach may not be well suited for sparse networks, where few links are formed given a number of units N.

Moreover, even though Kyriazidou (1997) has provided results for the asymptotic variance of the estimator defined in the previous subsection, we do not consider this here, since the same indices are present in several combinations of quadruples, and a correlation is introduced in the transformed error terms, such that the asymptotic variance should account for that. This was not an issue before, once the differentiation was over time for a given dyad and errors are i.i.d. over time and dyads.

10   Simulations

In this section we aim to access the finite sample behavior of the proposed estimators through a Monte Carlo simulation exercise.

We compare, for different designs, how the estimators for the first stage equation (selection equation) given by the standard Heckman approach and by the methodology proposed by Charbonneau (2017) perform in finite samples. For this first stage, the standard Heckman approach boils down to a probit estimation with dummies for the fixed effects (when the specified model contains fixed effects, otherwise a standard probit is employed). In cases where there are fixed effects in the selection equation, we also employ

a probit estimation that corrects for the perfect prediction problem, i.e., we drop (only for the first stage estimation) observations where for one of the units in the pairwise interaction (dyad) the outcome in the first stage is invariant. It is expected that, due to the incidental parameters problem, the estimator given by Charbonneau (2017) provides better results than both methods in terms of unbiasedness since they present an asymptotic bias, as explained in Section 6.

For the second stage (observation equation), we analyze the behavior of the estimators given by:

i. The standard Heckman approach, where the estimates of the probit in the first stage for both the structural parameters and the fixed effects of the selection equation are plugged into the functional form of the inverse Mills-ratio (given by Equation (16)). A standard OLS with dummies for the fixed effects is estimated for the observation equation, including the inverse Mills-ratio as a regressor.

ii. The standard Heckman approach, similar to the previous item with the only difference being that we take into account the estimates from the probit corrected for the perfect prediction problem in the first stage.

iii. The hybrid approach outlined in Section 8. In this approach the estimates given by the method of Charbonneau (2017) for the selection equation are taken as given. The fixed effects for the selection equation are estimated through the hybrid approach using the unconditional log-likelihood, restricting the structural parameters to be equal to their estimates. Finally, the inverse Mills-ratio is calculated for each observation according to the transformation proposed by Lee (1983), and included as a regressor in a standard OLS regression for the observation equation with dummies for the fixed effects. In designs where there are no fixed effects, the estimator takes into account only the Charbonneau estimates of the first stage and applies the transformation proposed by Lee (1983).

iv. The modified estimator based on Kyriazidou (1997). Again, we take as given the estimates given by the method outlined by Charbonneau (2017) for the selection equation. With such an estimator, we can calculate the weights for each observation (setting the Kernel to be a standard normal density function, as in Kyriazidou (1997)), and we transform the variables such as outlined in Section 9.3 to perform a weighted least squares regression for the observation equation.

As the possible biasedness of the first stage estimator for the standard Heckman approach might carry over to the estimates of the second stage, we expect the latter two approaches to provide better finite sample properties.

10.1 Data generating process and different designs

The data for the simulations are generated according to the following general DGP:

$$y_{1,ij} = x_{11,ij}\beta_{11} + x_{12,ij}\beta_{12} + \vartheta_i + \chi_j + u_{ij} \tag{47}$$

$$y_{2,ij} = \mathbb{1}(y_{2,ij}^{**} > 0) \tag{48}$$

$$y_{2,ij}^{**} = x_{21,ij}\beta_{21}^* + x_{22,ij}\beta_{22}^* + x_{23,ij}\beta_{23}^* + \xi_i^* + \zeta_j^* + \eta_{ij}^* \tag{49}$$

$$(i = 1, ...N; j = 1, ...N, i \neq j)$$

Throughout this simulation exercise we consider only one time period. Given the general DGP, we provide seven different designs. For all the designs we set $\beta_{11} = 1$, $\beta_{12} = 2.5$, $\beta_{21}^* = 0.8$, $\beta_{22}^* = 1$ and $\beta_{23}^* = 2$.

Design 1: We consider that there are no fixed effects, setting $\vartheta_i = \chi_j = \xi_i^* = \zeta_j^* = 0 \quad \forall i, j$. The remaining explanatory variables are generated according to:

- $x_1 = rndn$, where $rndn$ is a standard normal.

- $x_2 = \mathbb{1}\{rndn \leq 0.5\}$, being a bivariate variable.

- $x_3 = \mathbb{1}\{rndn \leq 0.5\}$, being a bivariate variable that satisfies an exclusion restriction with respect to the observation Equation (47).

Design 2: There are fixed effects in the selection equation, but not in the observation equation, setting $\vartheta_i = \chi_j = 0 \quad \forall i, j$. The variables x_1, x_2 and x_3 are generated as in Design 1. The fixed effects ξ_i^* and ζ_j^* are drawn from a standard normal distribution and are uncorrelated with the explanatory variables.

Design 3: There are fixed effects in the observation equation, but not in the selection equation, setting $\xi_i^* = \zeta_j^* = 0 \quad \forall i, j$. The variables x_1, x_2 and x_3 are generated as in Design 1. The fixed effects ϑ_i and χ_j are drawn from a standard normal distribution and are uncorrelated with the explanatory variables.

Design 4: There are fixed effects in both equations, and they are generated according to Designs 2 and 3, being uncorrelated with the explanatory variables. The variables x_1, x_2 and x_3 are generated as in Design 1.

Design 5: There are fixed effects in the selection equation, but not in the observation equation, setting $\vartheta_i = \chi_j = 0 \quad \forall i, j$. The fixed effects ξ_i^*

and ζ_j^* are generated according to Design 2, but are now correlated with the explanatory variable x_1. While x_2 and x_3 are obtained as in Design 1, we now set: $x_1 = rndn + \xi_i^* + \zeta_j^*$, where $rndn$ is a standard normal.

Design 6: There are fixed effects in the observation equation, but not in the selection equation, setting $\xi_i^* = \zeta_j^* = 0 \quad \forall i, j$. The fixed effects ϑ_i and χ_j are generated according to Design 3, but are now correlated with the explanatory variable x_1. While x_2 and x_3 are obtained as in Design 1, we now set: $x_1 = rndn + \vartheta_i + \chi_j$, where $rndn$ is a standard normal.

Design 7: There are fixed effects in both equations, and they are generated according to Designs 2 and 3, but are now correlated with the explanatory variables x_1. While x_2 and x_3 are obtained as in Design 1, we now set: $x_1 = rndn + \xi_i^* + \zeta_j^* + \vartheta_i + \chi_j$, where $rndn$ is a standard normal.

The error terms are generated by assuming that there is no misspecifications in the distributional assumptions made by each estimator. This translates to:

- For the Heckman approaches we consider that the error terms of Equations (47) and (49) follow a bivariate normal distribution, with both means set to zero, unit variances and a correlation of -0.7.

- For the Hybrid and Kyriazidou approaches, as both rely on the estimates obtained by Charbonneau estimators in the first stage, we consider that the error term of Equation (49) is logistically distributed with mean zero and variance one, and the error term of Equation (47) is normally distributed with mean zero and variance one. However, to ensure that Assumption 14 is satisfied, such that Lee's transformation can be applied, we start with generating two random variables that are bivariate normally distributed with with both means set to zero, unit variances and a correlation of -0.7. Then, we apply a transformation $J(\cdot) = F^{-1}(\Phi(\cdot))$ to one of them turning it into a logistically distributed variable, where F is the standard logistic distribution function, and Φ is the standard normal distribution function.

We also assume that there is no misspecification in the parametric form of models. For designs in which either all fixed effects or the fixed effects for a given equation are set to zero, we estimate only the remaining parameters according to the approaches highlighted in the beginning of this section. That is, for instance, in Design 1, we estimate the parameters β_{11}, β_{12}, β_{21}^*, β_{22}^* and β_{23}^*, but not the fixed effects.

For each design, we consider 500 Monte Carlo simulations. For all designs, we consider a sample size given by $N = 25$, i.e., giving a total of $N(N-1) = 600$ observations (where obviously some are set to have missing

values for y_1, given by the sample selection). Ideally, a set of simulations for a larger N should be provided. However, for now, we do not provide such results given the computational burden to estimate mainly the approach deliniated by Charbonneau (2017) and the modified version for the approach of Kyriazidou (1997). The computational costs arise as the estimation involves the combinations of all possible quadruples.

10.2 First stage (estimation of the selection equation)

In this section we present the results obtained from the Monte Carlo simulations for the first stage (selection equation). We denote Probit the estimators obtained with the standard Heckman procedure, which boils down to a MLE probit estimator with dummies for fixed effects when the design presents fixed effects. We denote Probit (PP) the estimators given by the same approach but corrected for the perfect prediction problem (notice that this estimator is only employed in designs which there are fixed effects in the selection equation). We denote by Charbonneau the estimators of the structural parameters using Charbonneau (2017)'s approach. Finally, we denote by Hybrid the estimators for the fixed effects obtained by the unconstrained MLE restricting the structural parameters to be the estimated by Charbonneau.

We report the mean bias, the mean standard deviation (used as a measure of precision), the mean standard errors and the test size given by the *t-statistic*, considering a 5% level of significance. These values can be seen in Tables 1 - 4 for the seven proposed designs.

We can see at first that for designs that do not contain fixed effects in the selection equation (1,3 and 6), both approaches deliver similar results. In some cases the Probit estimators even present less bias than Charbonneau. Given the reduced sample size, the size of the tests are close enough to 5%. However one should note that the Charbonneau estimator is more imprecise (higher standard deviations).

For designs that do contain fixed effects in the selection equation (Designs 2,4,5 and 7), the Probit and Probit (PP) estimates of the structural parameters are severely biased. The Charbonneau estimator, despite slightly biased, reduces dramatically the bias compared to the previous estimators. Moreover, in her simulation design Charbonneau (2017) shows that the bias is even further reduced for larger sample sizes. As we will mention later, the distribution of estimates of the fixed effects for the Probit (PP) are closer to the real distribution when comparing to the Probit estimates, since the perfect prediction problem is corrected. However, the biases in the structural paramaters remain of the same magnitude. This indicates that such biases in the structural parameters reflect to a larger extent the incidental parameter problem than the bias carried over from the fixed effects generated by the perfect prediction problem.

43

	Mean Bias	Mean Std. Deviation	Mean Std. Error	Test Size
Design 1				
Probit				
β_{21}^*	0.0182	0.1013	0.0955	0.062
β_{22}^*	0.0119	0.1342	0.1309	0.048
β_{23}^*	0.0367	0.2048	0.1964	0.054
z_{12}	0.0249	0.2153		
Logit Charbonneau				
β_{21}^*	0.0167	0.1467	0.1313	0.0880
β_{22}^*	0.0174	0.2642	0.2402	0.0640
β_{23}^*	0.0285	0.2989	0.2671	0.0700
z_{12}	0.0150	0.3258		
Design 2				
Probit				
β_{21}^*	0.2229	0.1638	0.1352	0.3620
β_{22}^*	0.2647	0.2592	0.2265	0.2040
β_{23}^*	0.5103	0.3669	0.2895	0.3900
ξ_1^*	0.3885	2.2845		
ζ_2^*	-0.1244	2.1568		
z_{12}	1.2188	4.2462		
Probit Perfect Prediction				
β_{21}^*	0.2226	0.1790	0.1342	0.3440
β_{22}^*	0.2633	0.2693	0.2258	0.2380
β_{23}^*	0.5299	0.3576	0.2887	0.4280
ξ_1^*	-0.2378	1.4069		
ζ_2^*	0.0988	1.3968		
z_{12}	-0.1289	2.6744		
Logit Charbonneau + Hybrid				
β_{21}^*	0.0282	0.1556	0.1424	0.0720
β_{22}^*	0.0254	0.2773	0.2609	0.0740
β_{23}^*	0.0624	0.3384	0.2902	0.0900
ξ_1^*	0.0678	0.7491		
ζ_2^*	0.0230	0.7340		
z_{12}	0.0986	1.0803		

Table 1: Simulation results for the estimated coefficients of the selection Equation 49 with $N = 25$ and 500 iterations. Test size refers to the size of the *t-test* at the 5% significance level.The estimator Probit Perfect Prediction is only obtained for designs with fixed effects in the selection equation. Empty cells refer to cases where the estimator is still not available.

	Mean Bias	Mean Std. Deviation	Mean Std. Error	Test Size
Design 3				
Probit				
β_{21}^*	0.0128	0.0968	0.0951	0.0400
β_{22}^*	0.0119	0.1298	0.1308	0.0400
β_{23}^*	0.0341	0.1956	0.1957	0.0460
z_{12}	0.0197	0.1899		
Logit Charbonneau				
β_{21}^*	0.0271	0.1406	0.1319	0.0660
β_{22}^*	0.0295	0.2525	0.2403	0.0520
β_{23}^*	0.0609	0.2990	0.2700	0.0640
z_{12}	0.0203	0.2993		
Design 4				
Probit				
β_{21}^*	0.2128	0.1577	0.1349	0.3166
β_{22}^*	0.2705	0.2516	0.2280	0.1964
β_{23}^*	0.5251	0.3674	0.2913	0.3888
ξ_1^*	0.3053	2.1540		
ζ_2^*	0.0048	2.3284		
z_{12}	1.0570	4.1692		
Probit Perfect Prediction				
β_{21}^*	0.2079	0.1638	0.1332	0.3020
β_{22}^*	0.2701	0.2689	0.2254	0.2160
β_{23}^*	0.5226	0.3773	0.2887	0.3920
ξ_1^*	-0.1273	1.3715		
ζ_2^*	-0.1223	1.3494		
z_{12}	0.2169	2.7302		
Logit Charbonneau				
β_{21}^*	0.0374	0.1495	0.1437	0.0640
β_{22}^*	0.0379	0.2885	0.2628	0.0640
β_{23}^*	0.0649	0.3130	0.2913	0.0580
ξ_1^*	-0.0366	0.7244		
ζ_2^*	0.0456	0.7549		
z_{12}	0.0458	1.0436		

Table 2: Simulation results for the estimated coefficients of the selection Equation 49 with $N = 25$ and 500 iterations. Test size refers to the size of the *t-test* at the 5% significance level.The estimator Probit Perfect Prediction is only obtained for designs with fixed effects in the selection equation. Empty cells refer to cases where the estimator is still not available.

	Mean Bias	Mean Std. Deviation	Mean Std. Error	Test Size
Design 5				
Probit				
β_{21}^*	0.2781	0.2143	0.1634	0.3640
β_{22}^*	0.3388	0.3435	0.2746	0.2340
β_{23}^*	0.7094	0.4625	0.3600	0.4680
ξ_1^*	0.3898	3.0340		
ζ_2^*	-0.0624	3.1310		
z_{12}	1.5217	5.9475		
Probit Perfect Prediction				
β_{21}^*	0.2842	0.2476	0.1634	0.3600
β_{22}^*	0.3496	0.3613	0.2718	0.2420
β_{23}^*	0.7083	0.5127	0.3578	0.4340
ξ_1^*	-0.1602	1.7806		
ζ_2^*	0.0107	1.7311		
z_{12}	0.1085	3.3272		
Logit Charbonneau + Hybrid				
β_{21}^*	0.0347	0.1809	0.1661	0.0760
β_{22}^*	0.0148	0.3191	0.3030	0.0580
β_{23}^*	0.0500	0.3730	0.3391	0.0680
ξ_1^*	-0.0173	0.8763		
ζ_2^*	-0.0368	0.8702		
z_{12}	-0.0069	1.2908		
Design 6				
Probit				
β_{21}^*	0.0125	0.0751	0.0751	0.0400
β_{22}^*	0.0082	0.1512	0.1403	0.0720
β_{23}^*	0.0351	0.1893	0.1922	0.0400
z_{12}	0.0147	0.2176		
Logit Charbonneau				
β_{21}^*	0.0203	0.1537	0.1394	0.0820
β_{22}^*	0.0138	0.2701	0.2530	0.0620
β_{23}^*	0.0424	0.3025	0.2816	0.0760
z_{12}	0.0540	0.3940		

Table 3: Simulation results for the estimated coefficients of the selection Equation 49 with $N = 25$ and 500 iterations. Test size refers to the size of the *t-test* at the 5% significance level.The estimator Probit Perfect Prediction is only obtained for designs with fixed effects in the selection equation. Empty cells refer to cases where the estimator is still not available.

	Mean Bias	Mean Std. Deviation	Mean Std. Error	Test Size
Design 7				
Probit				
β_{21}^*	0.3071	0.2439	0.1733	0.3564
β_{22}^*	0.3862	0.3695	0.2894	0.2311
β_{23}^*	0.7637	0.5344	0.3816	0.4816
ξ_1^*	0.4155	2.9246		
ζ_2^*	-0.0195	2.7139		
z_{12}	1.5677	5.3488		
Probit Perfect Prediction				
β_{21}^*	0.2968	0.2613	0.1724	0.3440
β_{22}^*	0.4048	0.3978	0.2897	0.2600
β_{23}^*	0.8006	0.6244	0.3855	0.4960
ξ_1^*	-0.3102	1.9542		
ζ_2^*	0.0994	1.8887		
z_{12}	-0.0445	4.0003		
Logit Charbonneau				
β_{21}^*	0.0398	0.1811	0.1729	0.0540
β_{22}^*	0.0645	0.3357	0.3198	0.0700
β_{23}^*	0.0948	0.3746	0.3548	0.0620
ξ_1^*	-0.1117	0.8924		
ζ_2^*	0.0073	0.9406		
z_{12}	-0.0854	1.2684		

Table 4: Simulation results for the estimated coefficients of the selection Equation 49 with $N = 25$ and 500 iterations. Test size refers to the size of the *t-test* at the 5% significance level.The estimator Probit Perfect Prediction is only obtained for designs with fixed effects in the selection equation. Empty cells refer to cases where the estimator is still not available.
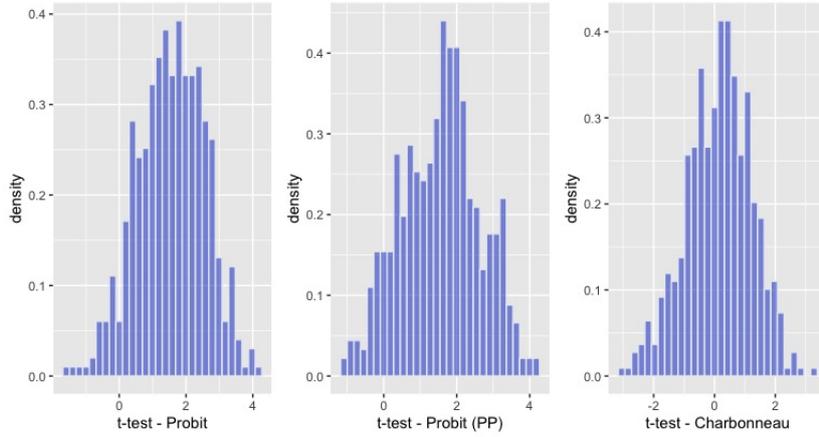
Figure 1: Histogram of the *t-test* for estimated β_{21}^* in Design 5
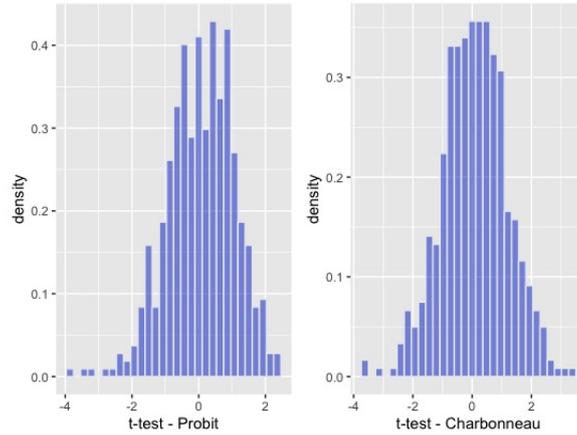
Figure 2: Histogram of the *t-test* for estimated β_{21}^* in Design 6

The results are even more improved for cases where the fixed effects are correlated with the explanatory variable x_1. While the biases in Charbonneau do not increase much when the fixed effects are correlated, indicating that this estimator is robust to such correlations, the biases in the Probit and Probit (PP) increase substantially. Not only that, but the standard deviations for the Probit estimators increases by a lot when the fixed effects are correlated. Note that the biases are even more severe for the parameters of the binary regressors.

The biases of the Probit and Probit (PP) estimators are also evident through the histogram for the *t-tests* of the estimated β_{21}^*. For this analysis we focus on Designs 5, 6 and 7, which includes cases where the fixed effects are correlated with x_1, and the improvements of Charbonneau are the bigger. In Design 6, where there are no fixed effects, the distribution of both Probit

48
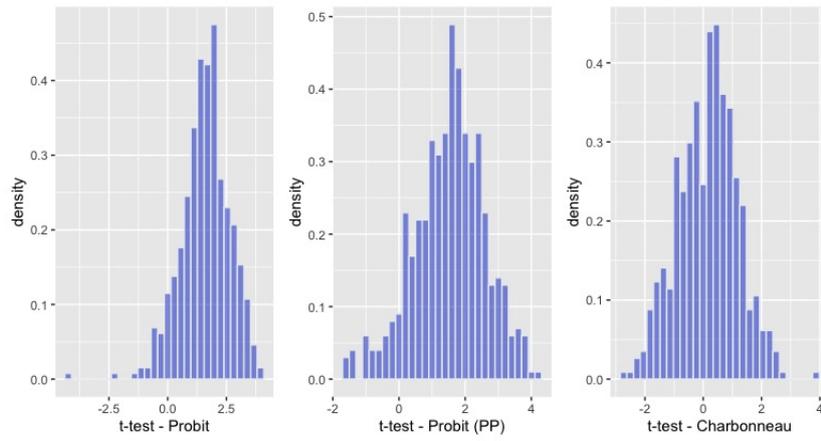
Figure 3: Histogram of the *t-test* for estimated β_{21}^* in Design 7
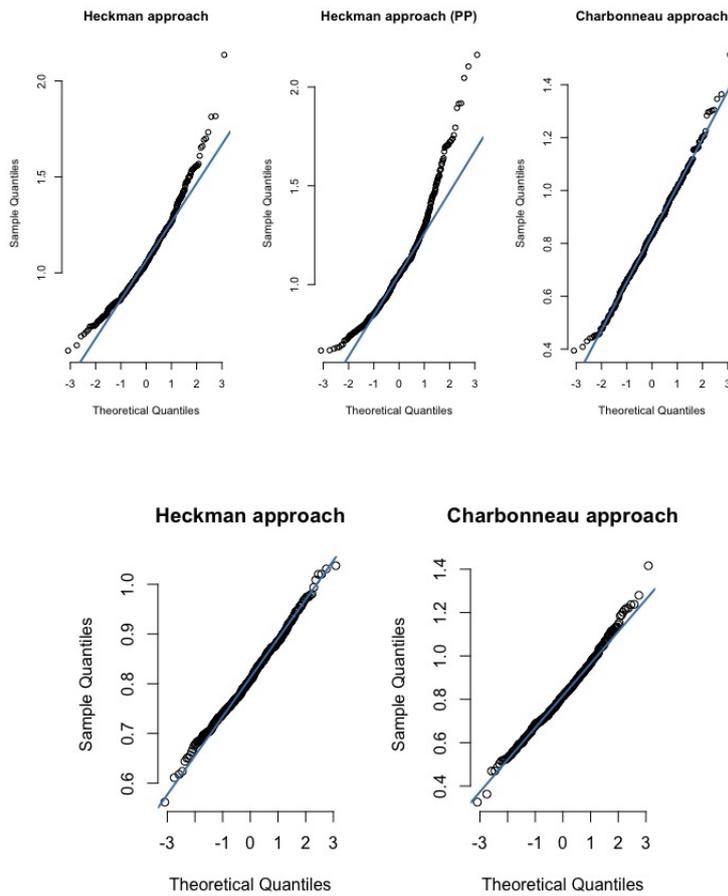
Figure 5: QQ plot of estimated β_{21}^* in Design 6

49
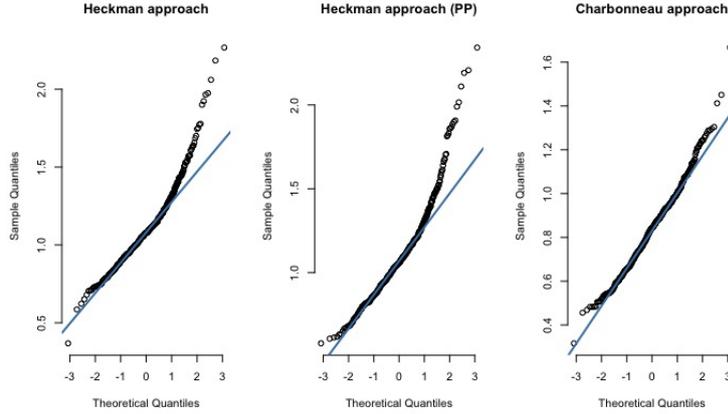
Figure 6: QQ plot of estimated β_{21}^* in Design 7

and Charbonneau *t-tests* are centered around zero, as illustrated in Figure 2, indicating that the tests have the right size. Also, from the shape of the distribution it indicates further that the estimated parameters are normally distributed. When there are fixed effects (Designs 5 and 7, in Figures 1 and 3), while the *t-tests* for Charbonneau are centered around zero, the ones for the Probit and Probit (PP) are not, indicating that the estimates are biased and the *t-tests* do not have the correct size.

The test sizes are also improved by the Charbonneau estimator. The significance level of the Probit and Probit (PP) estimators are much larger than 5%, in most cases ranging between 20% and 40%.

The distribution of the estimates are further assessed through the QQ-plots in Figures 4 - 6. In Design 6, the plot indicates that the distributions for both estimators are reasonably well approximated by a normal distribution. In Designs 5 and 7, while the Charbonneau estimates the distributions are also well approximated by a normal, the Probit and Probit (PP) estimates have distributions that are more skewed. The same argument holds for the estimates of β_{22}^* and β_{23}^*, which graphs are in the Appendix B.1.

We now analyze the estimated fixed effects ξ_i^* for $i = 1$ throughout the simulations. For both Designs 5 and 7, the estimates obtained by the Hybrid approach are much more closely distributed to the true fixed effects than the Probit and Probit (PP) estimates. The Probit estimates are off due to two reasons: (i) they are not corrected for the perfect prediction problem, leading to extreme estimated values; and (ii) even for not extreme estimated values, the distributions are not similar once the estimates are contaminated by the (asymptotic) bias of the structural parameters. The latter problem is evidenced when we look at the estimates of the fixed effects for the Probit (PP) estimator. While the estimates vary over a smaller range once we correct for the perfect prediction problem, leading to less biased estimates in general for the fixed effects, their distribution is still quite different from the distribution of its true values, still presenting a larger range than the

50
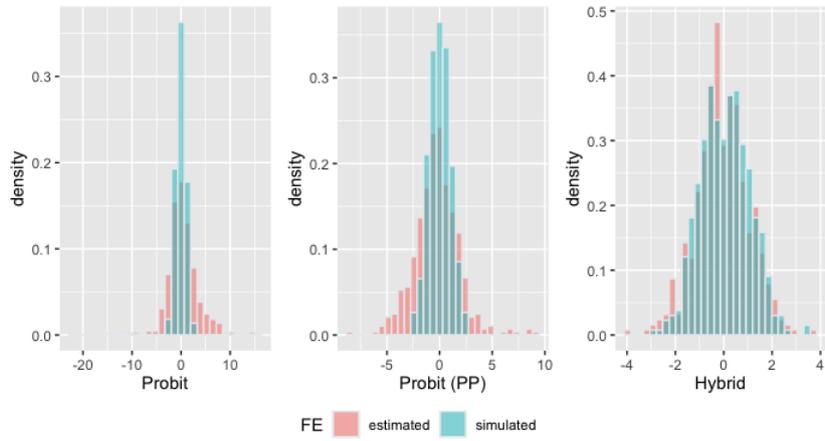
Figure 7: Histogram of estimated ξ_1^* in Design 5
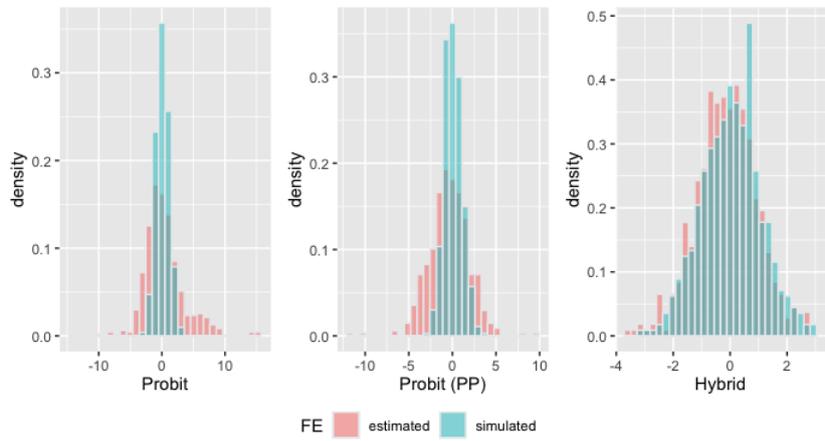
Figure 8: Histogram of estimated ξ_1^* in Design 7

true distribution. This indicates that even when correcting for the perfect prediction, the estimates of the fixed effects are still contaminated by the bias in the structural parameters. In Tables 1 - 4 we can see that the estimates of both fixed effects ξ_i^* and ζ_j^* have, in general, a much higher bias in the Probit estimators than in the Hybrid.
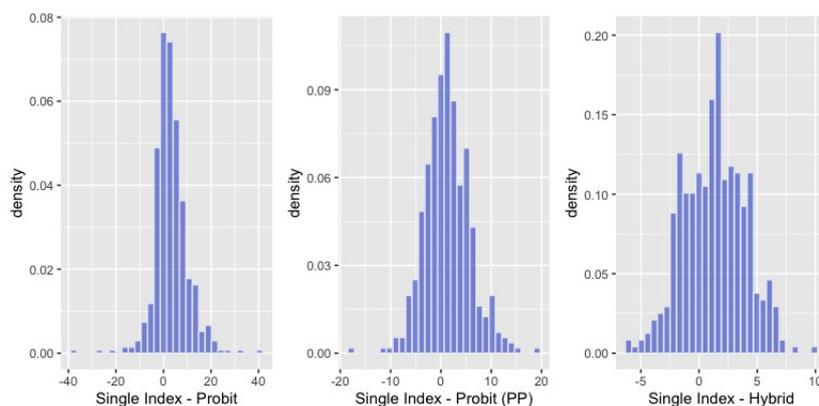
Figure 9: Histogram of $\hat{z}_{12} - z_{12}$ in Design 5
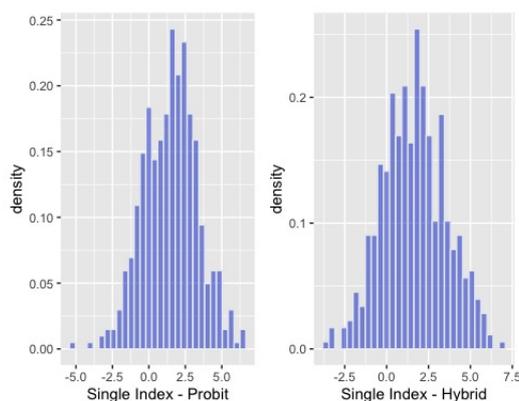
Figure 10: Histogram of $\hat{z}_{12} - z_{12}$ in Design 6

Both the structural and fixed effects parameters estimated by the Probit and Probit (PP) are biased when fixed effects are present in the DGP. The structural parameters through the incidental parameters problem, and the fixed effects parameters through the biased carried over from the estimation of the β_2^*'s and through the perfect prediction problem in the Probit. Both sources of biases are carried over to the estimates of the single index z_{ij}. This is seen in the previous tables, where in the designs with fixed effects, the bias of the estimates of z_{ij} are generally much larger for the Probit and Probit (PP). Even though the Probit (PP) produces less bias compared to
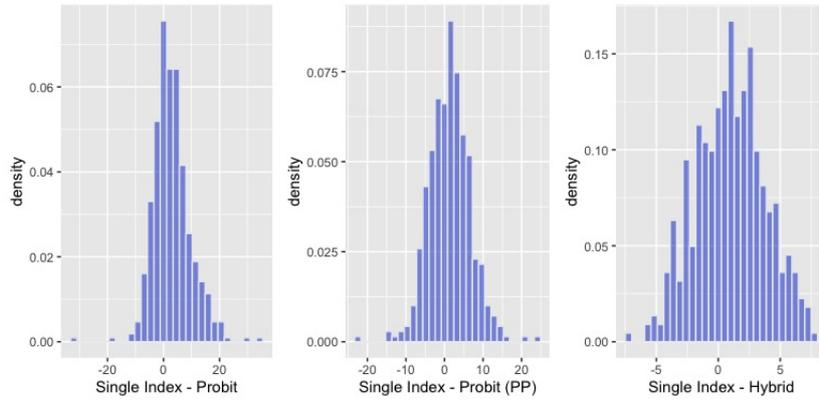
Figure 11: Histogram of $\hat{z}_{12} - z_{12}$ in Design 7
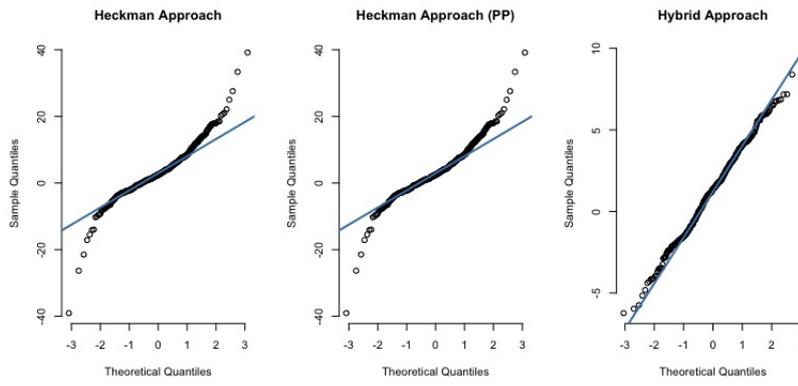
Figure 12: QQ plot of $\hat{z}_{12} - z_{12}$ in Design 5
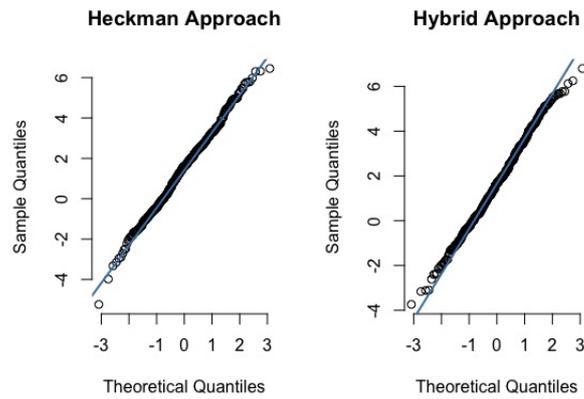
Figure 13: QQ plot of $\hat{z}_{12} - z_{12}$ in Design 6
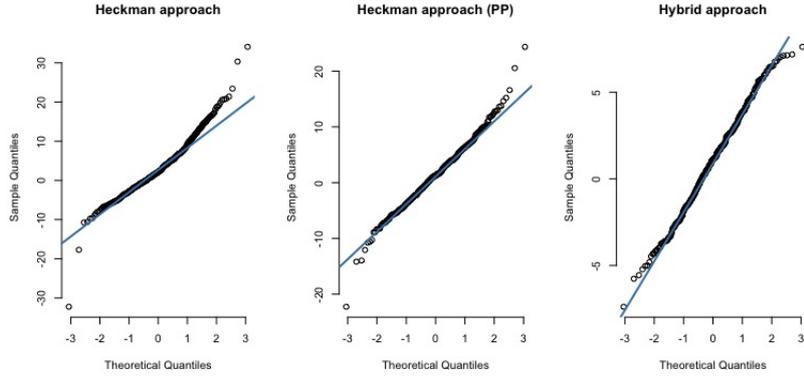
Figure 14: QQ plot of $\hat{z}_{12} - z_{12}$ in Design 7

the Probit estimator, since the perfect prediction problem is corrected, for most designs (with the exception of Design 7) the biases in the index z_{ij} are still larger than that of the Hybrid approach.

We can also see that through the histograms of the quantities $\hat{z}_{12} - z_{12}$, i.e, the difference of the estimated single index for the dyad $i = 1, j = 2$ and its true value, given by Figures 9 - 11 . For Designs 5 and 7, the range of the histograms for the Probit estimates is much larger than that of the Probit (PP) or the Hybrid estimates. The Hybrid estimator presents an even lower range than the Probit (PP). In the case where fixed effects are not present, in Design 6, this is not true anymore.

Besides, the QQ-plots for such quantities, given by Figures 12 - 14, show that in cases where there are fixed effects, the distribution of estimated single index for the Probit and the Probit (PP) are not well approximated by a normal, whereas for the Hybrid it is. There are two sources for that problem: (i) the fact that the structural parameters estimated by the Probit have a more skewed distribution than a normal, and (ii) the very extreme values estimated by the fixed effects. While the latter is aliviated for the Probit (PP) estimator, we notice that the estimations of the fixed effects are still varying over a wider range than those of the Hybrid approach. The QQ-plots indicates that for Design 5, the latter source dominates (as the figure indicates that the distribution of the single index has heavy tails), while for Design 7, the former source dominates. In the case of no fixed effects, all estimates are well behaved.

10.3   Second stage (estimation of the observation equation)

In this section, we present the results obtained from the Monte Carlo simulations for the second stage (observations equation). We denote by Heckman the estimators obtained through the standard approach, that take into ac-

54

count the Probit estimates in the first stage. Analogously, we denote by Heckman (PP) the estimators given by the standard approach, taking into account the Probit (PP) estimates in the first stage (we employ this estimator only in designs where there are fixed effects in the selection equation). We denote by Hybrid the estimators obtained by using the transformation proposed by Lee (1983), and finally we denote by Kyriazidou the estimators obtained through our modified methodology based on Kyriazidou (1997).

As we still do not have an analytical expression for the standard errors of the latter two estimators, we will present only the mean bias and the mean standard deviation of the estimators. Moreover, for the Kyriazidou approach we also provide the estimators obtained through the asymptotic bias correction highlighted in Section 9.2, and we denote them by Kyriazidou(corrected). Following the parameters chosen in the simulations in Kyriazidou (1997), we set $r = 1$ and $\delta = 0.1$. We choose the kernal density to be a standard normal density, and we experiment with several values of h, and its asymptotically bias corrections counterparts.

These results can be seen in Tables 5 and 6 for the seven proposed designs.

We can see that for designs where there were no fixed effects in the selection equation (Designs 1, 3 and 6) all proposed estimators present essentially no bias. The exceptions are given by (i) the estimates of the coefficient of the inverse Mills-ratio (which true value is essentially the correlation between the errors in the equations) in the Hybrid approach, and (ii) the estimates of β_{12} given by the Kyriazidou and Kyriazidou(corrected) estimators for $h = 10$. The latter is expected, since, as argued by Kyriazidou (1997), the bias of the estimates increases when the chosen bandwidth is increased.

The fact that such estimates are mostly unbiased for all proposed estimators also points out that even when fixed effects in the observation equation are present (as in Designs 3 and 6), the methods that rely on including dummy variables for them, namely, Heckman and Hybrid, deliver estimates with nice properties. Moreover, in this case, Kyriazidou, which relies on differencing out those fixed effects together with the sample selection effects also performs well. However, notice that the precision of the Hybrid and Kyriazidou approaches are slightly lower than the Heckman approach.

When fixed effects in the selection equation are included (Designs 2, 4, 5, 7), the Heckman and Heckman (PP) estimators present biases, which is aligned with Theorem 1. Since the single indices are biased in such cases, the estimates of the observation equation are likely to be also biased. Such biases are more pronounced for the coefficient of the binary variable x_2. Also, for this variable, the biases are not always in the same direction. However, it is striking that when comparing both estimators, the Heckman (PP) delivers even more biased estimates.

We note that for such designs, the Hybrid approach reduces the biases in general, specially for the coefficient of the binary variable. However, for

	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7
Heckman							
β_{11}	0.0002	0.0092	0.0014	0.0204	0.0160	0.0025	0.0161
	(0.0482)	(0.0466)	(0.0527)	(0.0516)	(0.0343)	(0.0536)	(0.0573)
β_{12}	-0.0004	-0.0257	0.0018	0.0288	-0.0398	-0.0025	0.0204
	(0.0628)	(0.0612)	(0.0957)	(0.1063)	(0.0687)	(0.1005)	(0.1114)
Inverse Mills-Ratio	0.0038	0.0613	0.0037	0.1317	0.0606	0.0043	0.1310
	(0.1203)	(0.1268)	(0.1659)	(0.1732)	(0.1387)	(0.1754)	(0.1877)
Heckman Perfect Prediction							
β_{11}		0.0174		0.0291	0.0427		0.0246
		(0.0531)		(0.0628)	(0.0399)		(0.0645)
β_{12}		-0.0495		0.0426	-0.0676		0.0418
		(0.08513)		(0.1131)	(0.0871)		(0.1218)
Inverse Mills-Ratio		-0.0997		0.1318	0.1016		0.4240
		(5.1842)		(2.7994)	(1.9824)		(4.0759)
Hybrid							
β_{11}	-0.0049	0.0109	0.0001	0.0128	0.0119	0.0035	0.0136
	(0.0587)	(0.0764)	(0.0705)	(0.0766)	(0.0925)	(0.0787)	(0.0987)
β_{12}	0.0001	0.0183	0.0038	-0.0044	0.0150	0.0048	-0.0051
	(0.0854)	(0.1381)	(0.1304)	(0.1475)	(0.1601)	(0.1372)	(0.1743)
Inverse Mills-Ratio	0.0050	0.0887	0.0189	0.0716	0.0968	0.0288	0.0781
	(0.1332)	(0.2597)	(0.2644)	(0.2695)	(0.3198)	(0.2742)	(0.3329)
Kyriazidou $h = 0.5$							
β_{11}	-0.0093	-0.0065	-0.0071	-0.0059	-0.0058	-0.0056	-0.0052
	(0.0820)	(0.0891)	(0.0811)	(0.0840)	(0.0941)	(0.0870)	(0.0903)
β_{12}	-0.0063	-0.0022	-0.0026	-0.0072	0.0044	0.0031	-0.0224
	(0.1421)	(0.1465)	(0.1412)	(0.1494)	(0.1657)	(0.1557)	(0.1633)
Kyriazidou $h = 0.5$, corrected							
β_{11}	-0.0093	-0.0066	-0.0071	-0.0059	-0.0058	-0.0056	-0.0053
	(0.0823)	(0.0894)	(0.0814)	(0.0843)	(0.0944)	(0.0873)	(0.0907)
β_{12}	-0.0064	-0.0022	-0.0026	-0.0071	0.0044	0.0030	-0.0225
	(0.1425)	(0.1471)	(0.1418)	(0.1499)	(0.1664)	(0.1563)	(0.1640)
Kyriazidou $h = 1$							
β_{11}	-0.0081	-0.0055	-0.0069	-0.0059	-0.0057	-0.0057	-0.0040
	(0.0736)	(0.0803)	(0.0729)	(0.0746)	(0.0862)	(0.0790)	(0.0809)
β_{12}	-0.0039	-0.0012	-0.0017	-0.0101	0.0034	0.0019	-0.0195
	(0.1290)	(0.1327)	(0.1264)	(0.1337)	(0.1461)	(0.1391)	(0.1445)
Kyriazidou $h = 1$, corrected							
β_{11}	-0.0082	-0.0056	-0.0071	-0.0060	-0.0058	-0.0058	-0.0041
	(0.0738)	(0.0805)	(0.0731)	(0.0748)	(0.0864)	(0.0792)	(0.0812)
β_{12}	-0.0041	-0.0013	-0.0018	-0.0102	0.0033	0.0017	-0.0197
	(0.1294)	(0.1331)	(0.1267)	(0.1341)	(0.1465)	(0.1395)	(0.1449)

Table 5: Simulation results for the estimated coefficients for the observation Equation 47 with $N = 25$ and 500 iterations. The values correspond to the mean bias of estimates, and the standard deviation is in parenthesis. The estimator Heckman Perfect Prediction is only obtained for designs with fixed effects in the selection equation.

	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7
Kyriazidou $h = 2$							
β_{11}	-0.0070	-0.0045	-0.0063	-0.0044	-0.0045	-0.0051	-0.0035
	(0.0692)	(0.0751)	(0.0690)	(0.0701)	(0.0812)	(0.0735)	(0.0750)
β_{12}	-0.0017	0.0000	-0.0021	-0.0135	0.0039	0.0030	-0.0175
	(0.1210)	(0.1240)	(0.1175)	(0.1253)	(0.1371)	(0.1287)	(0.1346)
Kyriazidou $h = 2$, corrected							
β_{11}	-0.0073	-0.0049	-0.0066	-0.0048	-0.0048	-0.0055	-0.0037
	(0.0694)	(0.0753)	(0.0692)	(0.0703)	(0.0814)	(0.0737)	(0.0753)
β_{12}	-0.0022	0.0001	-0.0026	-0.0138	0.0035	0.0026	-0.0178
	(0.1213)	(0.1243)	(0.1177)	(0.1256)	(0.1374)	(0.1290)	(0.1349)
Kyriazidou $h = 3$							
β_{11}	-0.0064	-0.0037	-0.0059	-0.0032	-0.0037	-0.0043	-0.0028
	(0.0678)	(0.0731)	(0.0676)	(0.0682)	(0.0791)	(0.0710)	(0.0722)
β_{12}	0.0000	0.0000	-0.0020	-0.0142	0.0048	0.0043	-0.0157
	(0.1178)	(0.1205)	(0.1142)	(0.1218)	(0.1331)	(0.1246)	(0.1302)
Kyriazidou $h = 3$, corrected							
β_{11}	-0.0069	-0.0042	-0.0065	-0.0037	-0.0041	-0.0048	-0.0032
	(0.0680)	(0.0734)	(0.0678)	(0.0684)	(0.0794)	(0.0713)	(0.0724)
β_{12}	-0.0001	-0.0010	-0.0028	-0.0148	0.0043	0.0037	-0.0162
	(0.1181)	(0.1208)	(0.1145)	(0.1221)	(0.1334)	(0.1249)	(0.1305)
Kyriazidou $h = 5$							
β_{11}	-0.0044	-0.0013	-0.0039	-0.0001	-0.0020	-0.0023	-0.0014
	(0.0658)	(0.0711)	(0.0656)	(0.0663)	(0.0768)	(0.0683)	(0.0693)
β_{12}	0.0033	0.0010	0.0000	-0.0130	0.0071	0.0066	-0.0124
	(0.1142)	(0.1167)	(0.1115)	(0.1186)	(0.1286)	(0.1204)	(0.1262)
Kyriazidou $h = 5$, corrected							
β_{11}	-0.0052	-0.0019	-0.0047	-0.0016	-0.0025	-0.0030	-0.0019
	(0.0661)	(0.0713)	(0.0659)	(0.0666)	(0.0771)	(0.0685)	(0.0695)
β_{12}	0.0023	0.0000	-0.0011	-0.0138	0.0064	0.0057	-0.0130
	(0.1146)	(0.1171)	(0.1118)	(0.1189)	(0.1289)	(0.1208)	(0.1266)
Kyriazidou $h = 10$							
β_{11}	0.0045	0.0073	0.0056	0.0074	0.0043	0.0062	0.0045
	(0.0614)	(0.0668)	(0.0617)	(0.0630)	(0.0722)	(0.0637)	(0.0656)
β_{12}	0.0153	0.0113	0.0108	-0.0039	0.0156	0.0168	-0.0046
	(0.1089)	(0.1116)	(0.1071)	(0.1141)	(0.1226)	(0.1144)	(0.1210)
Kyriazidou $h = 10$, corrected							
β_{11}	0.0036	0.0066	0.0047	0.0067	0.0038	0.0054	0.0040
	(0.0616)	(0.0670)	(0.0619)	(0.0632)	(0.0725)	(0.0640)	(0.0658)
β_{12}	0.0142	0.0104	0.0097	-0.0048	0.0150	0.0158	-0.0052
	(0.1092)	(0.1119)	(0.1074)	(0.1144)	(0.1229)	(0.1147)	(0.1213)

Table 6: Simulation results for the estimated coefficients for the observation Equation 47 with $N = 25$ and 500 iterations. The values correspond to the mean bias of estimates, and the standard deviation is in parenthesis.

Designs 2, 4 and 5, the Kyriazidou and Kyriazidou(corrected) estimators reduces the biases even further, specially for the initial choice of h ranging between 0.5 and 5. Moreover, looking at the standard deviations, in general, the Kyriazidou estimators reduce the bias while not sacrificing precision.

The exception to this observation is in Design 7. For this design we have that, while Kyriazidou estimators reduce further the biases in the coefficients of the continuous regressor x_1, the Hybrid approach does a better job for the binary variable x_2. For the coefficients of this variable the Kyriazidou estimators only reduces the biases observed in the Heckman estimator for $h > 2$. Then, the bias is basically eliminated only for $h = 10$. This contradicts the argument of Kyriazidou (1997) that the bias of the estimates increases when the chosen bandwidth is increased. Also, it points out that the estimates are sensitive to the choice of the initial h. However, note that the Kyriazidou estimators always reduces the biases when compared to the Heckman (PP) estimator, for all the designs and all the initially chosen values of h.

In appendix B.1 the estimates for all designs and approaches are given such that the parameters of the first stage estimation are set to their true value (for the Hybrid approach it boils down to only applying Lee's transformation on the variables). We note that in this case, the Kyriazidou estimators delivers unbiased estimates for all designs (except for the case when $h = 2$). Therefore, there is evidence that such remaining biases might come from the estimators of the first stage. We saw previously that the biases in the first stage Charbonneau estimates still did not vanish due to the reduced sample size. This indicates that possibly the remaining biases seen here are not originated by the Kyriazidou estimator itself. An intriguing fact is that the Hybrid estimator in this case presents some bias for the estimate of β_{12}.

Moreover, from the tables we can see that the estimators provided by Kyriazidou(corrected) do not reduce biases further compared to Kyriazidou. Even though that is contractory to Corollary 1, the same is observed in the simulations provided in Kyriazidou (1997). There are two possible explanations for this: (i) the small sample bias is more important than the asymptotic bias for these designs, or (ii) other methods for choosing an appropriate bandwith given an initial chosen h should be provided.

In Figures 15 - 17, we provide the QQ-plots for the estimates obtained by the different approaches for the parameter β_{11}. While it is expected that the estimates of the Hybrid and the Kyriazidou approaches are approximatelly normal, it is striking that the estimates of the Heckman methods are also approximately normally distributed, since the distribution of single indexes for these approaches deviates from a normal distribution. The same is observed for the estimates of the parameter β_{12}, which plots are presented in Appendix B.1.
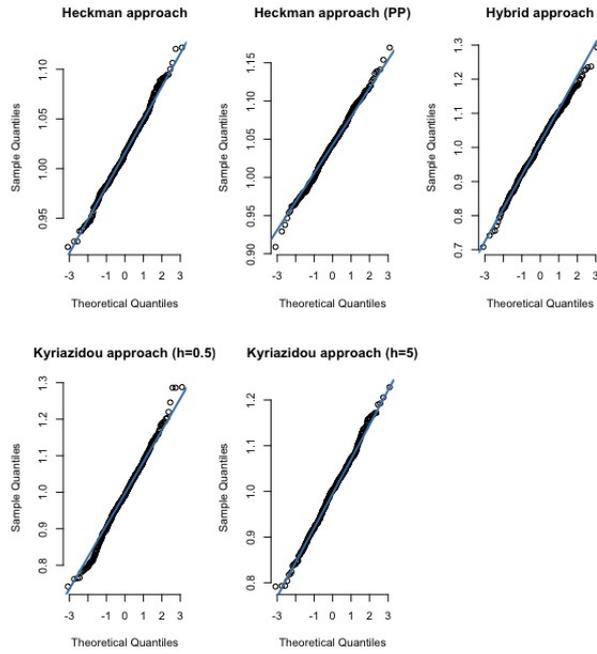
Figure 15: QQ plot of estimated β_{11} in Design 5
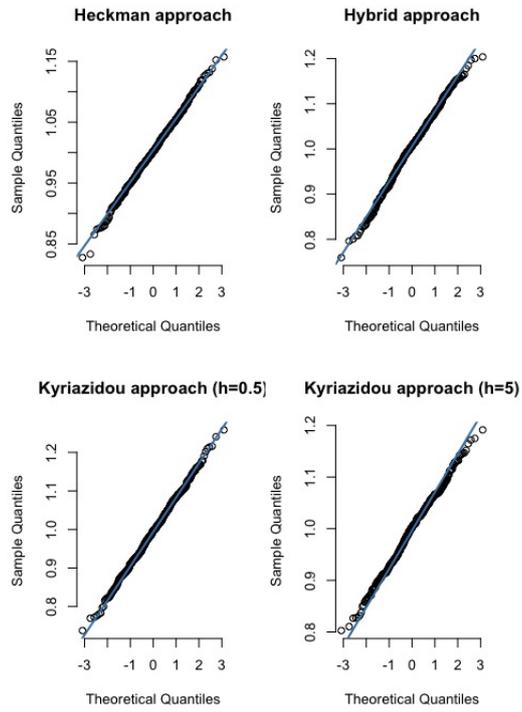
Figure 16: QQ plot of estimated β_{11} in Design 6
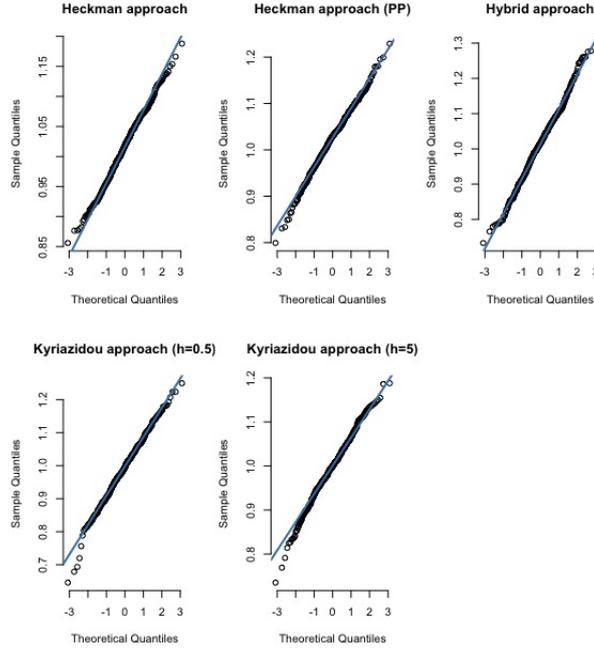
Figure 17: QQ plot of estimated β_{11} in Design 7

11 An application to the gravity model for international trade

In this section we apply the proposed methods for estimating gravity models to international trade flows. We aim to estimate how trade barriers affect both the decision of country i to export to country j, and the volume of trade. We will estimate the model outlined in Section 3, with the only difference being that we do not include w_{ij} as a regressor in the second stage estimation. This variable accounts for the fraction of firms that export from country i to country j in country i, being related to the extensive margins of trade. When this variable is omitted, the estimation confounds the effects of trade barriers on the intensity of trade (intensive margin) with the effects on the proportion of exporting firms (extensive margin). The sample selection effect is introduced in this model when only countries with positive trade flows are taken into account in the observation equation.

Therefore, while the estimation proposed by Helpman et al. (2008) corrects both for the effect of this variable w_{ij} and the sample selection effect in the observation equation, we will focus here on the latter effect.

The reason to not take into account w_{ij} for now is that this is an estimated variable, that is obtained through the estimated inverse Mills-ratio. As demonstrated in Helpman et al. (2008), it introduces a further non-linearity in the observation equation. This non-linearity needs to be taken

into account by employing a non-linear least squares (NLS) estimator when applying the standard Heckman approach. Therefore, for simplicity, and to clearly explore the differences between our proposed estimators and the standard Heckman procedure, we will focus only on correcting for the sample selection bias.

We use the dataset provided by Helpman et al. (2008), which is the same used by Charbonneau (2017) in her application to estimate the parameters for the first stage estimation, that refers to the decision of countries to trade (the probability that country i exports to country j). Therefore, we can compare our results in the first stage estimation with both results obtained by Helpman et al. (2008) and Charbonneau (2017), determining if we have a successfull replication of their studies.

The dataset provides information on directed trade flows and country characteristics for 158 countries in 1986. The country characteristics boil down to attributes on the country pairs (dyads). The variables used are, as described by Helpman et al. (2008):

- $Distance_{ij}$: the distance (km) between the exporter's i and importer's j capitals in logs.

- $Common\ Border_{ij}$: a binary variable that equals one if the exporter i and importer j share borders (are neighbors).

- $Colonial\ ties_{ij}$: a binary variable that equals one if importing country j ever colonized in the history the exporting country i or vice versa, and zero otherwise.

- $Currency\ Union_{ij}$: a binary variable that equals one if the exporter i and the importer j have a common currency or if their own currencies had a 1:1 exchange rate for an extended period of time, and zero otherwise.

- $Common\ Legal\ system_{ij}$: a binary variable that equals one if the exporting country i and importing country j have the same legal origin, and zero otherwise.

- $Common\ Religion_{ij}$: (% Protestants in country i. % Protestants in country j) +(% Catholics in country $i \cdot$ % Catholics in country j) +(% Muslims in country i. % Muslims in country j).

- FTA_{ij}: a binary variable that equals one if the exporter i and the importer j belong to a common regional trade agreement, and zero otherwise.

In our estimation of the first stage equation we disconsider the observations with Congo as an exporter, as for those, the decision of trade does

not change (Congo never exports to any country). This is done to avoid the problem of perfect predictability.

Both Helpman et al. (2008) and Charbonneau (2017) take into account variables related to whether both countries are islands or if both do not have direct access to the sea. However, when we implement the Charbonneau approach in this application, such transformed variables lack of variation for the identification of the parameters. For obtaining the coefficients, we apply a standard logit estimation to the transformed variable (note that the standard errors have a more involved estimation, obtained as Jochmans (2018) proposes). The reason why Charbonneau (2017) is able to incorporate such variables might come from two sources: (i) instead of applying a standard logit, the author maximizes numerically the Equation (32), or (ii) the author takes into account the observations with Congo as exporter. However, this is not a crucial issue, once other studies also disregarded such variables (for instance, Jochmans (2018)).

For more details on the construction of the variables in the dataset, refer to Appendix I in Helpman et al. (2008).

We estimate the parameters β_2^* of the following equation in the first stage (selection equation):

$$y_{2,ij} = \mathbb{1}(x'_{2,ij}\beta_2^* + \xi_i^* + \zeta_j^* > \eta_{ij}^*) \tag{50}$$

where: (i) $y_{2,ij}$ is a binary variable, being one if the i exports to j and zero otherwise, (ii) ξ_i^* is the exporter fixed effect, (iii) ζ_j^* is the importer fixed effect, (iv) $x_{2,ij}$ is the vector that collects the variables $Distance_{ij}$, $Common\ Border_{ij}$, $Colonial\ Ties_{ij}$, $Currency\ Union_{ij}$, $Common\ Legal\ System_{ij}$, FTA_{ij}, $Common\ Religion_{ij}$.

The following is the specified model for the second stage (observation equation):

$$y_{1,ij,t} = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + u_{ij,t} \tag{51}$$

where: (i) $y_{1,ij}$ is the log of the value of the exports from i to j, (ii) ϑ_i is the exporter fixed effect, (iii) χ_j is the importer fixed effect, (iii) $x_{1,ij,t}$ is the vector that collects the same variables as $x_{2,ij}$, with the exception of $Common\ Religion_{ij}$ which is taken into account only in the first estimates of the standard Heckman approach as we will explain later. We only use dyads with positive exports to estimate this second stage equation. Therefore, a sample selection correction must be employed.

11.1 Estimation of selection equation (decision to trade)

We estimate the selection equation using a standard Probit with dummies for the fixed effects (which here boils down to the Probit (PP) in our simulations, as we excluded Congo as an exporter) and the Charbonneau estimator. The results can be seen in Table 7.

	Probit	Charbonneau
Common language	0.2903***	0.4248***
	(0.0379)	(0.0732)
Common legal system	0.0972**	0.1822***
	(0.0296)	(0.0597)
Common Religion	0.2647***	0.4979***
	(0.0585)	(0.1125)
Common Border	−0.3798***	−0.5164**
	(0.0946)	(0.2300)
Currency Union	0.4883***	1.0459***
	(0.1306)	(0.2362)
Distance	−0.6626***	−1.0001***
	(0.0208)	(0.0541)
FTA	2.0170***	3.5565***
	(0.3085)	(0.5237)
Colonial Ties	0.3337	1.1432*
	(0.2852)	(0.6473)

Table 7: Estimates for the first stage Equation 50. For the Probit estimates we consider importer and exporter fixed effects, correcting for the perfect prediction problem. Standard errors are in parenthesis.
* indicates that the coefficient is significant at the 10 % level
** indicates that the coefficient is significant at the 5 % level
*** indicates that the coefficient is significant at the 1 % level.

Even though we do not take into account two variables considered by Charbonneau (2017), our estimates for the remaining parameters are remarkably similar for both estimators and for all remaining coefficients, when comparing to the results she presents in her paper.

We can see that our Charbonneau estimates have a higher magnitude for all the coefficients when compared to our Probit estimates. Thus, the proposed correction for the asymptotic biases in the standard Probit model plays a practical role in this application.

All the estimates also have the same sign as the ones obtained by Helpman et al. (2008) (where the Probit estimator is applied in the first stage). As expected, geographic distance decreases the probability of countries to trade, while the coefficients for the other characteristics that reflect how similar countries i and j are in general positive (*Common Language*, *Common Religion*, *Colonial Ties*, *Currency Union*, *Common Legal System*), indicating that there is homophily in how trade networks are formed (similar countries have a higher propensity to trade). Countries that belong to a common regional trade agreement also have a higher probability of trading, as expected.

The estimates for both methods point out that countries with a common border are likely to trade less. As explained by Helpman et al. (2008), this negative effect might be due to territorial border conflicts that harms trade

between countries that are neighbors.

For the Charbonneau method, we estimate the standard errors as proposed by Jochmans (2018). Even though Charbonneau (2017) does not clearly state how the estimates of the standard errors are obtained in her paper, when comparing, we find that our estimates are similar to hers. The exceptions are only for the variables *Colonial Ties* and *FTA*. In terms of significance levels, the variable *FTA* is significant at the 1% level both in our application and in hers, while the variable *Colonial Ties* is not significant even at the 10% level in her application and here it is significant at this level.

When comparing our Charbonneau estimates to our Probit estimates, the levels of significance of the variables are also very similar, with the exceptions of (i) *Common legal system*, which in these estimations it is significant at the 5% level, while in our Charbonneau estimation is significant at the 1% level; and (ii) *Colonial Ties*, which is not significant even at the 10% level in the Probit estimations, but it is significant at this level for the Charbonneau estimator, as mentioned before.

11.2 Estimation of observation equation (directed trade flows)

Once the estimates of the first stage are obtained, we follow to the estimation of the second stage, which measures the effects of these variables on the volumes of exports from country i to j.

We apply the same methods outlined in Section 10. Namely, given the Probit estimates we employ the standard Heckman approach, and given the Charbonneau estimates, we employ the Hybrid and modified Kyriazidou approaches (with and without the asymptotic corrections). All the methods propose to correct for sample selectivity, and thus, allow for correlation between the errors of Equations (50) and (51).

For the modified Kyriazidou approach, we follow the parameters chosen in Section 10, setting $r = 1$, $\delta = 0.1$, and experimenting throughout a range of possible values of the initial bandwidth h. Moreover, this approach requires a variable that satisfies an exclusion restriction in the sense that it affects the decisions to trade (affects the fixed costs of trading), but it does not affect the volume of trade flows (does not affect the variable costs). We follow Helpman et al. (2008) and use *Common Religion* as such variable. That *Common Religion* affects the decision of trade is clear from the estimates of the first stage. For the second requirement, the authors argue that when including variables related to regulation costs of firm entry, the coefficient of the variable *Common Religion* becomes not significant in the second stage. Note that variables related to regulation costs satisfy the exclusion restriction, since by construction they do not affect the variable costs.

In this study, we deliberately chose not to include the variables that refer to the regulation costs, since they reduce the sample size drastically.

In our case, this is aggravated since both the Charbonneau and the modified Kyriazidou estimators already reduce the amount of information used from the dataset due to all the conditions that must be met on quadruples.

However, as it is evident on the results presented in Table 8, the variable *Common Religion* does not completely satisfy the conditions of exclusion restriction, since its coefficient in the estimates of the standard Heckman approach is still significant at the 5% level. Thus, one should be careful when looking at the estimates of the modified Kyriazidou approach. We could not verify the significance level in the new proposed approaches since the estimates of the standard errors are still not available.

From Tables 8 and 9, we can see that regardless of the employed method, all coefficients except for *Distance* are positive. In the case of the variable *Common Border* this means that, while being neighbors reduces the probability of trading, when trade is made, being neighbors increases the volume of trade.

The estimates obtained with the Hybrid and the modified Kyriazidou approach are different from the ones obtained with the standard Heckman approach. Thus, there is evidence that the latter is biased. However, we should point out that in order to properly claim that there are biases in the estimates, we would need to test whether the estimated coefficients are significantly different. As we still do not have estimates for the standard errors in the proposed approaches, this test cannot be done.

Moreover, while for some variables these possible biases have a clear direction when looking at the estimates provided by the two proposed methods, for others it does not have.

The coefficient of *Common Border* is higher for all estimates obtained from the Hybrid and the Kyriazidou approach. For the *FTA* variable, the same holds, except for the estimator obtained with the Kyriazidou approach when setting the initial h to be ten. Finally, for *Currency Union* the coefficients are also higher for the Hybrid approach and for the Kyriazidou with higher levels of the initial h. For the remaining variables, the two proposed methods would imply various directions for the bias.

As in the results obtained in the simulation exercise, the asymptotic correction for the estimates of the modified Kyriazidou approach does not deliver different estimates compared to the initial estimation. Finally, it is evidenced in this application that the estimates for this approach are sensitive to the initial chosen level of h.

12   Conclusion

In this study we showed that accounting for sample selection bias in dyadic data settings is not straightforward and requires more involved methods than the standard Helpman et al. (2008) two-step approach. The standard

	Heckman (1)	Heckman (2)	Hybrid	K (h=0.5)	K,corrected (h=0.5)	K (h=1)	K,corrected (h=1)
Common Language	0.2090***	0.2333***	0.3566	0.1094	0.1069	0.1465	0.1448
Common Legal System	0.4797***	0.4914***	0.4975	0.3805	0.3791	0.4054	0.4047
Common Border	0.4031***	0.4318***	0.5836	0.6114	0.6127	0.6106	0.6131
Currency Union	1.3537***	1.3368***	1.4456	1.2315	1.2248	1.3186	1.3158
Distance	−1.2087***	−1.2190***	-1.3470	-1.1371	-1.1329	-1.2185	-1.2187
FTA	0.7774***	0.7719***	2.0569	1.2835	1.2835	1.2475	1.2636
Colonial Ties	1.3418***	1.3434***	1.4209	1.2594	1.2560	1.3524	1.3563
Common Religion	0.2389**						
Inverse Mills-Ratio	0.2784***	0.2671***	0.7130				

Table 8: Estimates for the second stage Equation 51. For the Heckman and Hybrid estimates we consider importer and exporter fixed effects, correcting for the perfect prediction problem. We denote by Kyriazidou by K.
* indicates that the coefficient is significant at the 10 % level
** indicates that the coefficient is significant at the 5 % level
*** indicates that the coefficient is significant at the 1 % level.

	K(h=2)	K,corrected (h=2)	K(h=3)	K,corrected (h=3)	K(h=5)	K,corrected (h=5)	K(h=10)	K,corrected (h=10)
Common Language	0.1696	0.1685	0.1830	0.1822	0.1933	0.1926	0.1994	0.1987
Common Legal System	0.4154	0.4151	0.4206	0.4204	0.4242	0.4240	0.4255	0.4252
Common Border	0.5920	0.5950	0.5673	0.5708	0.5436	0.5486	0.5129	0.5193
Currency Union	1.4043	1.4035	1.4164	1.4159	1.3955	1.3941	1.4213	1.4208
Distance	-1.2525	-1.2555	-1.2503	-1.2538	-1.2172	-1.2198	-1.1721	-1.1730
FTA	1.3033	1.3415	1.2408	1.2826	0.8675	0.8972	0.2901	0.2981
Colonial Ties	1.3729	1.3794	1.3387	1.3441	1.2552	1.2576	1.1945	1.1947

Table 9: Estimates for the second stage Equation 51. We denote by Kyriazidou by K.

Heckman approach involves a first stage where the probability of observing an outcome is estimated through a probit model with dummys for fixed effects, and a second stage that estimates the effects of different variables on such outcome, taking into account the estimated probabilities in the first stage.

The motivation to study dyadic interactions came from the gravity models for international flows. We specifically consider the model proposed by Helpman et al. (2008), since it explicitly takes into account how the sample selectivity arises when estimating how trade barriers affect trade flows, when considering in the estimation only countries with positive trade flows. Another key aspect in this model is the presence of multilateral resistance terms, that boils down to the inclusion of two-way fixed effects for both countries involved in the trade.

Even though the motivation emerged from the trade literature, the approaches outlined in this study can be applied to any dyadic interaction where sample selectivity might be present (given that the model satisfies the outlined assumptions throughout the study). More generally, the first stage estimations, which outline the probability of observing a pairwise interaction that generates an outcome (in the case above, the probability of trading with each other), can be seen as an estimation of a network formation model, as highlighted in Jochmans (2018). The presence of the two-way fixed effects also indicates a close relationship to the β-models in the networks literature.

The difficulty in correcting for sample selectivity in such models arises as the two-way fixed effects are incorporated in the first stage estimation. These effects leads to the incidental parameter problem (Neyman and Scott (1948)), and the estimates of structural parameters obtained through the standard Probit are asymptotically biased (Fernández-Val and Weidner (2016)).

We showed more formally how this problem arises, and proposed the conditional logit estimates of Charbonneau (2017) to estimate the structural parameters of the first stage equation. Such method relies on a clever way of differencing out both fixed effects in the equation over quadruples of dyads. The downside of this method is that it evidently does not provide estimates of the fixed effects, and therefore, the estimation of the inverse Mills-ratio for the traditional Heckman approach are not readily available.

To bypass this problem, we proposed two approaches based on existing methods, but providing the suitable modifications to this setting of dyadic regressions. The first method, denoted by Hybrid, relies on retrieving the estimates of the fixed effects through the unconditional log-likelihood, restricting the structural parameters to be the same as the estimates obtained by the Charbonneau estimates. Once the fixed effects estimates are collected, it is possible to transform the variables, as proposed by Lee (1983) such that the Heckman two-step approach can be employed even when the errors in the selection equation are logistically distributed. The advantages

of this approach is that there is no need for a variable that satisfies an exclusion restriction, since the sample selection effects have a defined functional form in the observation equation, guaranteeing the identification of its structural parameters. The downsides of this method are that (i) one still needs to obtain estimates for the fixed effects in the first stage equation, pointing out that this approach may not be suitable for sparse networks, and that (ii) the distribution of the errors need to be specified.

The second method is based on Kyriazidou (1997), and involves differencing the second stage equation over quadruples of dyads. The form of differencing guarantees that the fixed effects in the second stage equation are completely differenced out, however it does not guarantee that the sample selectivity effects are completely differenced out as well. To overcome this problem, the main idea is then that weights are applied to the transformed observations such that a higher weight is attributed to quadruples for which the differences in the sample selectivity effects are smaller. Such differences in sample selectivity effects take into account the estimates provided by Charbonneau of the first stage. The benefits of this method is that there is no need to estimate any of the fixed effects in both equations, and to specify the distribution of the errors in the observation equation. The deficiencies of this approach are that the estimates are sensitive to the choices of bandwidths h, as shown in the application and in the Monte Carlo simulations. Also, the proposed method to choose an optimal bandwidth seems to not show improvements. One aspect of both methods is that they are computationally costly, since they boil down to combinatoric problems over quadruples of dyads that must satisfy some given restrictions.

Our Monte Carlo simulation exercise confirms the theoretical predictions that under the normality assumption of both error terms, the estimators of the first stage equation given by the Probit model with fixed effects are biased. On top of that, such biases are carried over to the estimations of the second stage equation, given by the standard Heckman approach. The exercise also demonstrates that,as predicted by the theory, the proposed methods (under their respective assumptions) deliver reductions in the biases the estimators of both equations.

However, we must point out that in our empirical application to the gravity model for international trade flows, it is not straightforward to draw conclusions from the obtained estimates from the different approaches for some variables. One of the potential sources for this limitation is that the method based on Kyriazidou (1997) requires a variable that satisfies an exclusion restriction. In our application it seems to be the case that the proposed variable does not satisfy the requirement that it should not affect the outcome of the second stage equation.

Finally, further research topics related to this study are: (i) obtaining the asymptotic properties of both proposed estimators, (ii) allow for a setting with several time periods, (iii) investigate whether a better method for

choosing the bandwidth h in the second approach can be obtained and (iv) study the possibility of extending the methods such that it is possible to be employed in networks where transitivy is present (which involves relaxing assumptions related to errors being i.i.d.).

References

AHN, H. AND J. L. POWELL (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

ANDERSON, J. E. (1979): "A theoretical foundation for the gravity equation," *The American economic review*, 69, 106–116.

ANDERSON, J. E. AND E. VAN WINCOOP (2003): "Gravity with gravitas: A solution to the border puzzle," *American economic review*, 93, 170–192.

ARELLANO, M. AND J. HAHN (2007): "Understanding bias in nonlinear panel models: Some recent developments," *Econometric Society Monographs*, 43, 381.

ARELLANO, M. AND B. HONORÉ (2001): "Panel data models: some recent developments," in *Handbook of econometrics*, Elsevier, vol. 5, 3229–3296.

BARTOLUCCI, F. AND C. PIGINI (2019): "Partial effects estimation for fixed-effects logit panel data models," *Working Papers 431, Universita' Politecnica delle Marche (I), Dipartimento di Scienze Economiche e Sociali.*

BIERENS, H. J. (1987): "Kernel estimators of regression functions," in *Advances in econometrics: Fifth world congress*, Cambridge University Press New York, vol. 1, 99–144.

CHARBONNEAU, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13.

DZEMSKI, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101, 763–776.

FERNÁNDEZ-VAL, I. AND M. WEIDNER (2016): "Individual and time effects in nonlinear panel models with large N, T," *Journal of Econometrics*, 192, 291–312.

GRAHAM, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

——— (2020): "Dyadic regression," in *The Econometric Analysis of Network Data*, Elsevier, 23–40.

HECKMAN, J. J. (1979): "Sample selection as a specification error," *Econometrica*, 47, 153–161.

HELPMAN, E., M. MELITZ, AND Y. RUBINSTEIN (2008): "Estimating trade flows: Trading partners and trading volumes," *The quarterly journal of economics*, 123, 441–487.

JOCHMANS, K. (2018): "Semiparametric analysis of network formation," *Journal of Business & Economic Statistics*, 36, 705–713.

——— (2019): "Modified-likelihood estimation of fixed-effect models for dyadic data," *Cambridge Working Papers in Economics*.

KRUGMAN, P. (1991): "Increasing returns and economic geography," *Journal of political economy*, 99, 483–499.

KUNZ, J., K. E. STAUB, AND R. WINKELMANN (2017): "Estimating Fixed Effects: Perfect Prediction and Bias in Binary Response Panel Models, with an Application to the Hospital Readmissions Reduction Program," *Available at SSRN 3074193*.

KYRIAZIDOU, E. (1997): "Estimation of a panel data sample selection model," *Econometrica: Journal of the Econometric Society*, 1335–1364.

LEE, L.-F. (1983): "Generalized econometric models with selectivity," *Econometrica: Journal of the Econometric Society*, 507–512.

MADDALA, G. S. (1983): "Qualitative and limited dependent variable models in econometrics," *Cambridge: Cambridge*, 498.

NEYMAN, J. AND E. L. SCOTT (1948): "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, 1–32.

RASCH, G. (1960): *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.*, Nielsen & Lydiche.

TINBERGEN, J. (1962): *Shaping the World Economy; Suggestions for an International Economic Policy*, Twentieth Century Fund, New York.

WOOLDRIDGE, J. M., D. W. KWAK, AND R. S. MARTIN (2018): "The robustness of conditional logit for binary response panel data models with serial correlation," *August. US Bureau of Labor Statistics Working Paper*, 502.

A Appendix: Derivations

A.1 Derivation of Helpman et al. (2008) gravity equation

In this subsection we follow the derivations of Helpman et al. (2008) for their gravity model. We consider a model with an economy comprising I countries, indexed by $i = 1, 2, ..., I$. All countries consume and produce a continuum of products.

Country i's utility function is given by:

$$u_i = \left[\int_{l \in B_i} x_i(l)^\alpha dl \right]^{1/\alpha}, 0 < \alpha < 1 \tag{52}$$

where $x_i(l)$ is its consumption of product i, B_i is the set of products available for consumption in country i and α determines the elasticity of substitution across products, given by $\sigma = \frac{1}{1-\alpha}$. Denoting Y_i the income of country i, which equals its expenditure levels, we can write country i's demand for product l as:

$$x_i(l) = \frac{\breve{p}_i(l)^{-\sigma} Y_i}{P_i^{1-\sigma}} \tag{53}$$

Where $\breve{p}_i(l)$ is the price of product l in country i, and P_i is the country's price index, given by:

$$P_i = \left[\int_{l \in B_i} \breve{p}_i(l)^{1-\sigma} dl \right]^{1/(1-\sigma)} \tag{54}$$

Note here that σ is not only constant across countries, but also across products. Some of the products in B_i are domestically produced while others are imported. Remembering that a given country i has a measure of N_i firms, and that the products are differentiated by the origin country, we have that there are $\sum_i^I N_i$ products in the economy.

As mentioned in Section 2, the firms of country i produces one unit of output with a combination of inputs given by $c_i a$ that minimizes costs, where a is firm-specific and denotes the number of bundles of inputs used per output, and c_i is a country-specific cost of the bundle. Moreover, $1/a$ reflects the productivity of a firm, and a follows a cumulative distribution function $G(a)$ with support $[a_L, a_H]$. The functional form of this distribution is the same across countries.

When a firm sells a product in the home market it bears only the production costs, but if the firm sells its product in country j, there are two additional costs: the fixed cost of serving j, and a transport (variable) cost that takes the form of a "melting iceberg" cost, as mentioned in Section 2.

By assuming monopolistic competition in the final goods, and given that every single firm has a measure zero, the demand function implies that a

firm from country i maximizes its profits by charging the mill-price, which is a standard mark-up pricing equation:

$$\breve{p}_i(l) = \tau_{ij}\frac{c_i a}{\alpha} \tag{55}$$

If this firm sells to consumers in country j, then it sets a delivered price equal to it. As a result, by combining the demand function, the prices and the associated costs, we have that the profits that a firm in country i obtains by exporting to country j is given by:

$$\pi_{ij}(a) = (1-\alpha)\left(\frac{\tau_{ij}c_i a}{\alpha P_j}\right)^{1-\sigma}Y_j - c_i f_{ij} \tag{56}$$

For a firm in country i, exporting to a country j is only profitable for $a \leq a_{ij}$, where a_{ij} can be defined by $\pi_{ij}(a_{ij}) = 0$, or:

$$(1-\alpha)\left(\frac{\tau_{ij}c_i a_{ij}}{\alpha P_j}\right)^{1-\sigma}Y_j = c_i f_{ij} \tag{57}$$

Which is equation 1. Therefore, this equation determines the fraction of firms in country i that will export to country j, given by $G(a_{ij})$, which, as highlighted before, can be zero. Therefore, the set B_j of products available in country j will be smaller than the set of products in the economy.

Then, the bilateral trade volumes is given by:

$$V_{ij} = \begin{cases} \int_{a_L}^{a_{ij}} a^{1-\sigma}dG(a) & \text{for } a_{ij} \geq a_L \\ 0 & \text{otherwise} \end{cases} \tag{58}$$

Note that this equation is almost simply obtaining the fraction of firms exporting to country j. The difference is the term $1-\sigma$, which was included just to simplify the remaining derivations of the model.

From the demand function, the pricing equation and by taking into account that country i has a measure of N_i firms, we have that the value of country i's exports to country j is:

$$Y_{1,ij} = \left(\frac{c_i\tau_{ij}}{\alpha P_j}\right)^{1-\sigma}Y_j N_i V_{ij} \tag{59}$$

It is clear that this bilateral trade volume is equal to zero if $a_{ij} \leq a_L$, since in this case $V_{ij} = 0$. Moreover, using the equation of the price indices, Equation (54), and the definition of V_{ij}, we finally obtain the last equation that characterizes the model:

$$P_j^{1-\sigma} = \sum_{i=1}^{I} \left(\frac{c_i \tau_{ij}}{\alpha} \right)^{1-\sigma} N_i V_{ij} \tag{60}$$

Equations (57)-(60) show that the bilateral trade flows $Y_{1,ij}$ can be obtained from the income levels Y_j, the number of firms N_i, the unit costs c_i, the fixed costs f_{ij} and the variable costs τ_{ij}.

We now show that under the assumptions that the variable costs are symmetric ($\tau_{ij} = \tau_{ji}$) and that V_{ij} can be multiplicatively decomposed as a deterministic function of three components (one that depends only on the exporter, one that depends only on the importer and one that depends on country-pair characteristics that are symmetric for every country-pair), we can obtain the gravity model detailed in Anderson and Van Wincoop (2003). This indicates that the model presented here is a generalization of the traditional gravity model, where such assumptions were relaxed.

First, we use the equality of income and expenditures of country i, given by $Y_i = \sum_{j=1}^{J} Y_{1,ij}$ (note that this summation also takes into account the sales to home residents, $Y_{1,ii}$). Then, we can use Equation (59) to write Y_i as:

$$Y_i = \left(\frac{c_i}{\alpha} \right)^{1-\sigma} N_i \sum_h \left(\frac{\tau_{ih}}{P_h} \right)^{1-\sigma} Y_h V_{ih} \tag{61}$$

We can obtain an analogous expression for Y_j, and substitute in Equation (59) to obtain:

$$Y_{1,ij} = \frac{Y_i Y_j}{Y^W} \frac{\left(\frac{\tau_{ij}}{P_j} \right)^{1-\sigma} V_{ij}}{\sum_{h=1}^{J} \left(\frac{\tau_{ih}}{P_h} \right)^{1-\sigma} V_{ih} s_h} \tag{62}$$

where $Y^W = \sum_{j=1}^{J} Y_j$ is the world income and $s_h = Y_h / Y^W$ is the share of country h in the world income.

We now assume the following:

ASSUMPTION 17. The following restrictions should hold:
(i) The variable costs are symmetric $\tau_{ij} = \tau_{ji}$.
(ii) V_{ij} is decomposable as follows

$$V_{ij} = (\varphi_{IM,j} \varphi_{EX,i} \varphi_{ij})^{1-\sigma}$$

where $\varphi_{IM,i}$ depends only on the parameters of the importing country, $\varphi_{EX,j}$ depends only on the parameters of the exporting country, and $\varphi_{ij} = \varphi_{ji} \ \forall i, j$.

We can then use Equation (62) to obtain:

$$\frac{Y_{1,ij}}{Y^W} = s_i s_j \left(\frac{\tau_{ij}\varphi_{ij}}{Q_i Q_j}\right)^{1-\sigma} \tag{63}$$

Where $Q_j = P_j/\varphi_{IM,j}$, and the values of Q_i are solved from:

$$Q_i^{1-\sigma} = \sum_h \left(\frac{\tau_{ih}\varphi_{ih}}{Q_h}\right)^{1-\sigma} s_h \tag{64}$$

Which is implied by Equations (4) and (61). Therefore, we can see that Equations (63) and (64) are essentially the system of equations derived by Anderson and Van Wincoop (2003). Note that in this case, the heterogeneous productivity of firms do not play a role in the volume of trade flows, and neither the sample selection problem is explicitly stated in the equation.

A.2 Helpman et al. (2008) estimation equations

A.2.1 Definitions of latent variable $Y_{2,ij,t}^*$ and $W_{ij,t}$

We define the latent variable $Y_{2,ij,t}^*$ as:

$$Y_{2,ij,t}^* = \frac{(1-\alpha)\left(P_j \frac{\alpha}{c_i \tau_{ij}}\right)^{\sigma-1} Y_j a_L^{1-\sigma}}{c_i f_{ij}} \tag{65}$$

Which represents the ratio of variable export profits for the most productive firm (with productivity $1/a_L$) to the fixed export costs (common to all exporters) for exports from i to j. Then, in this case, positive exports are observed if $Y_{2,ij,t}^* \geq 1$, and one can verify that W_{ij} is a monotonic function of $Y_{2,ij,t}^*$ in the case of positive exports: $W_{ij} = Y_{2,ij,t}^{*(k-\sigma+1)/(\sigma-1)} - 1$.

To see this last equality, by isolating a_{ij} from equation 1, we have that:

$$a_{ij} = \left(\frac{1-\alpha}{c_i f_{ij}} Y_j\right)^{1/\sigma-1} \left(\frac{\tau_{ij} c_i}{\alpha P_j}\right)^{\frac{1-\sigma}{\sigma-1}} \tag{66}$$

In case of positive exports, we have that:

$$W_{ij} = \left(\frac{a_{ij}}{a_L}\right)^{k-\sigma+1} - 1 \tag{67}$$

Then, by substituting Equation (66) in this later expression, we find that $W_{ij} = Y_{2,ij,t}^{*(k-\sigma+1)/(\sigma-1)} - 1$.

A.2.2 Sample selection as an omitted variable bias

Now, following Heckman (1979), we can formulate how the sample selection bias arises in this framework for the gravity model. Recording the form of the equation for trade flows (5), if we consider the its population regression function, we can write:

$$\mathbb{E}[y_{1,ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}] = x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + w_{ij} \tag{68}$$

However, the regression function for the subsample of the data that is self-selected into trade is given by:

$$\mathbb{E}[y_{1,ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}, \text{sample selection rule}] \tag{69}$$
$$= x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + w_{ij} + \mathbb{E}[u_{ij}|\text{sample selection rule}]$$

Therefore, the assumption that $\mathbb{E}[u_{ij}] = 0$ holds for the population, but not necessarily for the subsample. Only if the conditional expectation of u_{ij} is zero, then the assumption also holds for the subsample. However, we will see that in our framework it does not hold. As we know, we observe trade flows if and only if $y_{2,ij,t}{}^* > 0$. Therefore, we have that:

$$\mathbb{E}[u_{ij}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}, \text{sample selection rule}] \tag{70}$$
$$= \mathbb{E}[u_{ij}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}, y_{2,ij,t}{}^* > 0]$$

Given equation 70, and the definition of $y_{2,ij,t}$, we refine it down to:

$$\mathbb{E}[u_{ij}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}, \text{sample selection rule}] \tag{71}$$
$$= \mathbb{E}[u_{ij}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}, \eta_{ij} > -x'_{2,ij,t}\beta_2 - \xi_i - \zeta_j]$$

This highlights the fact that in case of independence between u_{ij} and η_{ij}, data on x_{ij} is missing randomly, and the conditional mean of u_{ij} is zero. However, this is not the case since they are correlated, as mentioned before. Therefore, the subsample regression function translates to:

$$\mathbb{E}[y_{1,ij,t}|x_{1,ij,t}, \vartheta_i, \chi_j, w_{ij}, y^*_{2,ij,t} > 0] \tag{72}$$
$$= x'_{1,ij,t}\beta_1 + \vartheta_i + \chi_j + w_{ij} + \mathbb{E}[u_{ij}|\eta_{ij} > -x'_{2,ij,t}\beta_2 - \xi_i - \zeta_j]$$

One can see that the sample selection function depends not only on the variable trade costs, but also on the fixed trade costs. Therefore, the regression estimators of the parameters of Equation (5) fitted on the selected sample omit the final term of Equation (73) as a regressor. Thus, essentially, the bias that results from using nonrandomly selected samples arise from a problem of omitted variables.

A.3 Sketch of the derivation of the asymptotic expansion

In this section we provide a more formalized version of the argument provided for the derivation of the asymptotic expansion given by Equation (27). The basis for it is found in Remark 2 in Fernández-Val and Weidner (2016).

Taking a first-order Taylor expansion of the first order conditions of Equation (25) around $\beta_{2,0}^*$, gives:

$$
0 = \frac{\partial \mathcal{L}(\hat{\beta}_2^*, \hat{\omega}_{NN}^*(\beta_2^*))}{\partial_{\beta_2^*}} \approx \frac{\partial \mathcal{L}(\beta_{2,0}^*, \hat{\omega}_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}}
$$
$$
- \bar{W}_\infty \sqrt{N(N-1)T}(\hat{\beta}_2^* - \beta_{2,0}^*) \tag{73}
$$

Then, we apply a second-order Taylor expansion to approximate the above term $\frac{\partial \mathcal{L}(\beta_{2,0}^*, \hat{\omega}_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}}$ around $\omega_{NN}^*(\beta_{2,0}^*)$, such that the estimates of the fixed effects are taken into account.

$$
\frac{\partial \mathcal{L}(\beta_{2,0}^*, \hat{\omega}_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}} \approx \frac{\partial \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}} \tag{74}
$$
$$
+ \frac{\partial^2 \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*} \partial_{\omega_{NN}'}}[\hat{\omega}_{NN}^*(\beta_{2,0}^*) - \omega_{NN}^*(\beta_{2,0}^*)]
$$
$$
+ \sum_{k=1}^{\dim \omega_{NN}} \frac{\partial^3 \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*} \partial_{\omega_{NN}'} \partial_{\omega_{NN,k}}}[\hat{\omega}_{NN}^*(\beta_{2,0}^*) - \omega_{NN}^*(\beta_{2,0}^*)][\hat{\omega}_{NN,k}^*(\beta_{2,0}^*) - \omega_{NN,k}^*(\beta_{2,0}^*)]/2
$$

Under regularity conditions, since the first term in this expression is the score vector, it has mean zero and it generates the asymptotic variance. By the information matrix equality and the Central Limit Theorem, we have:

$$
\frac{\partial \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}} \xrightarrow{d} N(0, \bar{W}_\infty) \tag{75}
$$

For some variance $\bar{W}_\infty$. According Fernández-Val and Weidner (2016), the second and the third term satisfies:

$$
\frac{\partial^2 \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*} \partial_{\omega_{NN}'}}[\hat{\omega}_{NN}^*(\beta_{2,0}^*) - \omega_{NN}^*(\beta_{2,0}^*)]
$$
$$
+ \sum_{k=1}^{\dim \omega_{NN}} \frac{\partial^3 \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*} \partial_{\omega_{NN}'} \partial_{\omega_{NN,k}}}[\hat{\omega}_{NN}^*(\beta_{2,0}^*) - \omega_{NN}^*(\beta_{2,0}^*)][\hat{\omega}_{NN,k}^*(\beta_{2,0}^*) - \omega_{NN,k}^*(\beta_{2,0}^*)]/2
$$
$$
\approx \sqrt{N(N-1)T}\Big(\frac{\bar{B}_\infty^\beta}{(N-1)T} + \frac{\bar{D}_\infty^\beta}{(N-1)T}\Big) \tag{76}
$$

The analytical form of terms $\bar{B}_\infty^\beta$ and $\bar{D}_\infty^\beta$ can be obtained from the second-order Taylor expansion as shown in Fernández-Val and Weidner (2016).

However, as mentioned before, in this study we do not aim to do so. Those terms originate from elements corresponding to the two-way fixed effects.

Plugging in the expression (76) into the equation for the first-order Taylor expansion, we have, as $N \to \infty$:

$$\bar{W}_\infty \sqrt{N(N-1)T}(\hat{\beta}_2^* - \beta_{2,0}^*) = \frac{\partial \mathcal{L}(\beta_{2,0}^*, \omega_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}} + \frac{\bar{B}_\infty^\beta}{\sqrt{T}} + \frac{\bar{D}_\infty^\beta}{\sqrt{T}} \quad (77)$$

By Slutsky Theorem, we have, given (75):

$$\sqrt{N(N-1)T}(\hat{\beta}_2^* - \beta_{2,0}^*) \xrightarrow{d} \bar{W}_\infty^{-1} N\left(\frac{\bar{B}_\infty^\beta}{\sqrt{T}} + \frac{\bar{D}_\infty^\beta}{\sqrt{T}}, \bar{W}_\infty\right) \quad (78)$$

Therefore, compared to the expression given by (29) in Section 6, we have that:

$$\bar{W}_\infty^{-1} \bar{B}_\infty^\beta = \bar{B}_\infty$$
$$\bar{W}_\infty^{-1} \bar{D}_\infty^\beta = \bar{D}_\infty$$

The main point of this sketch of the derivation is to show that the score $\frac{\partial \mathcal{L}(\beta_{2,0}^*, \hat{\omega}_{NN}^*(\beta_{2,0}^*))}{\partial_{\beta_2^*}}$ is not centered around zero when $\beta_2^* = \beta_{2,0}^*$, which originates the asymptotic biases. Moreover, the score not being centered around zero comes from the introduction of the nuisance parameters, which converge only at a slower rate compared to β_2^*.

A.4 A simple example for the perfect prediction problem

This Appendix outlines a simple example provided by Kunz et al. (2017) that demonstrates the perfect prediction problem.

Consider a standard panel Probit model:

$$\Pr(y_{it} = 1 | \alpha_i, x_{it}) = \Phi(\alpha_i + x_{it}'\beta)$$

$$(i = 1, ...N; t = 1, ...T)$$

where $y_{it} \in \{0, 1\}$, a binary variable, α_i is an individual fixed effect, x_{it} is a vector of K explanatory variables, $\beta \in \mathbb{R}^K$ is a vector of coefficients, and Φ is the standard normal distribution function.

The log-likelihood is given by:

$$\mathcal{L}(\beta) = \sum_{i=1}^N \sum_{t=1}^T \{y_{it} \ln \Phi(\alpha_i + x_{it}'\beta) + (1 - y_{it}) \ln(1 - \Phi(\alpha_i + x_{it}'\beta))\}$$

We show that the first order conditions of this log-likelihood does not have a finite solution if $\sum_{t=1}^T y_{it} = 0$ or $\sum_{t=1}^T y_{it} = T$ for a given i. Thus, the estimates of α_i do not exist in those cases.

The first order conditions are given by:

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \Phi(\alpha_i + x'_{it}\beta)) \frac{\phi(\alpha_i + x'_{it}\beta)}{\Phi(\alpha_i + x'_{it}\beta)(1 - \Phi(\alpha_i + x'_{it}\beta))} x_{k,it} = 0$$

$$\forall k = 1, ...K$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \sum_{t=1}^{T} (y_{it} - \Phi(\alpha_i + x'_{it}\beta)) \frac{\phi(\alpha_i + x'_{it}\beta)}{\Phi(\alpha_i + x'_{it}\beta)(1 - \Phi(\alpha_i + x'_{it}\beta))} = 0 \quad (79)$$

$$\forall i = 1, ...N$$

In the first case, suppose that $\sum_{t=1}^{T} y_{it} = 0$ for some i. Then, from Equation (79) it follows that:

$$-\sum_{t=1}^{T} \frac{\phi(\alpha_i + x'_{it}\beta)}{(1 - \Phi(\alpha_i + x'_{it}\beta))} = 0$$

From Section 5 we see that this is a summation over inverse Mills-ratios. One of the properties of the inverse Mills ratio is that $\lambda_{it} = \frac{\phi(\alpha_i + x'_{it}\beta)}{(1 - \Phi(\alpha_i + x'_{it}\beta))} > 0$ for a finite index $\alpha_i + x'_{it}\beta$. Then, this summation cannot be equal to zero, and this equation does not have a solution.

In the second case, suppose that $\sum_{t=1}^{T} y_{it} = T$, for some i. Then, from Equation (79) it follows that:

$$-\sum_{t=1}^{T} \frac{\phi(\alpha_i + x'_{it}\beta)}{\Phi(\alpha_i + x'_{it}\beta)} = 0$$

This summation also cannot be equal to zero from the properties of the standard normal density function for a finite $\alpha_i + x'_{it}\beta$.

B Appendix: Aditional figures and tables for simulations

B.1 First stage estimates
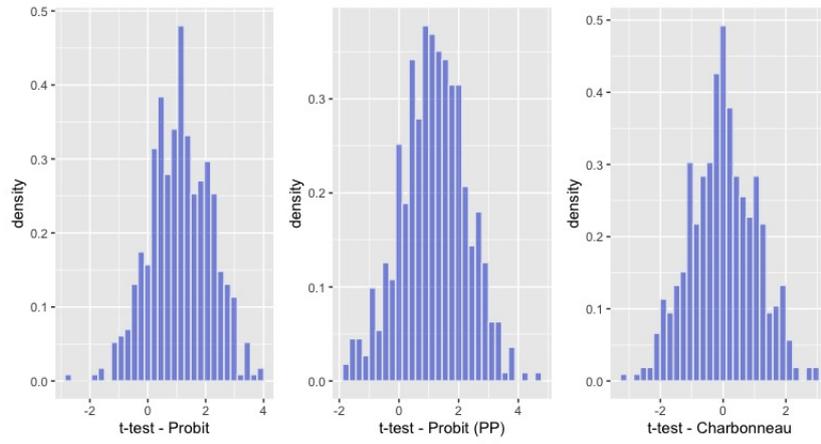
Figure 18: Histogram of the *t-test* for estimated β_{22}^* in Design 5
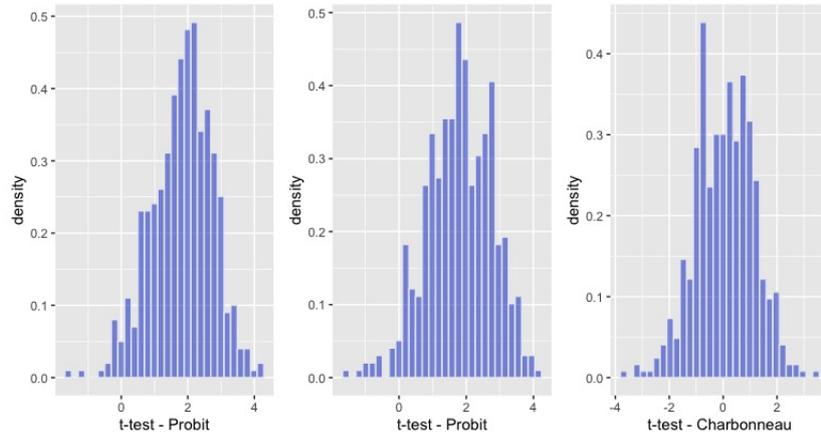
Figure 19: Histogram of the *t-test* for estimated β_{23}^* in Design 5
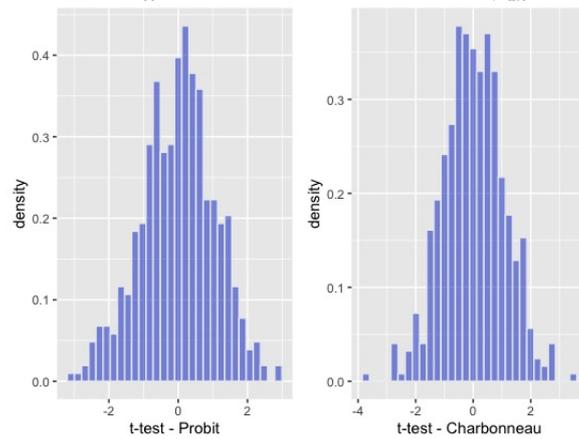
Figure 20: Histogram of the *t-test* for estimated β_{22}^* in Design 6
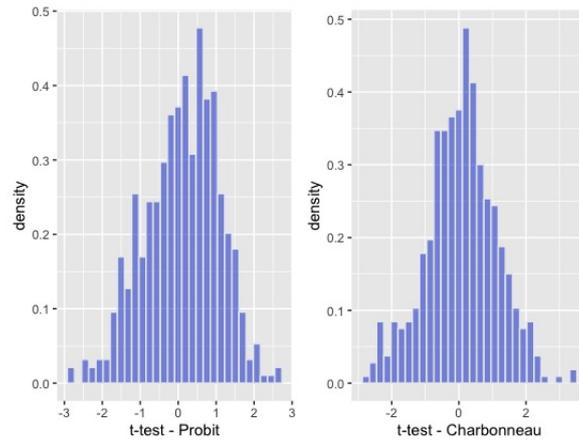
Figure 21: Histogram of the *t-test* for estimated β_{23}^* in Design 6
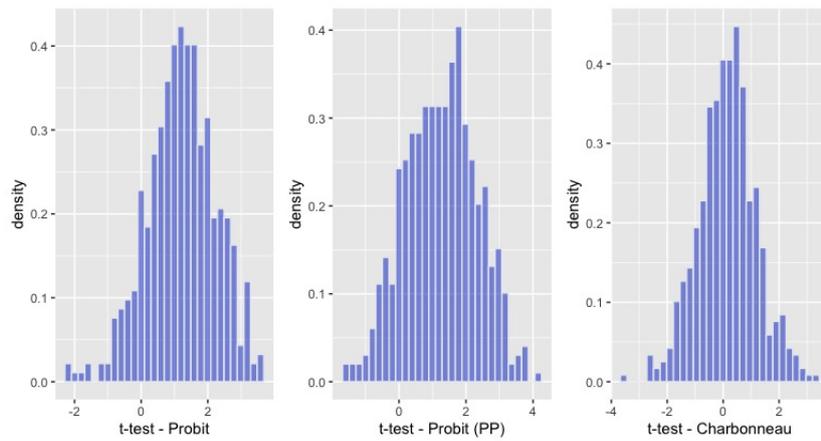
Figure 22: Histogram of the *t-test* for estimated β_{22}^* in Design 7
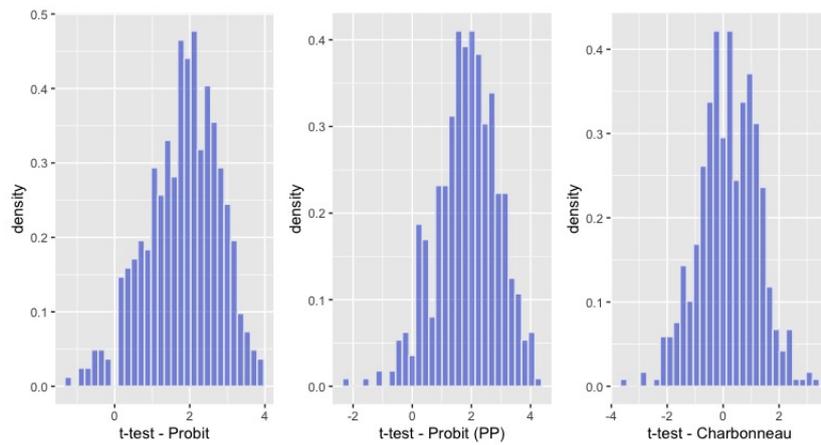
Figure 23: Histogram of the *t-test* for estimated β_{23}^* in Design 7
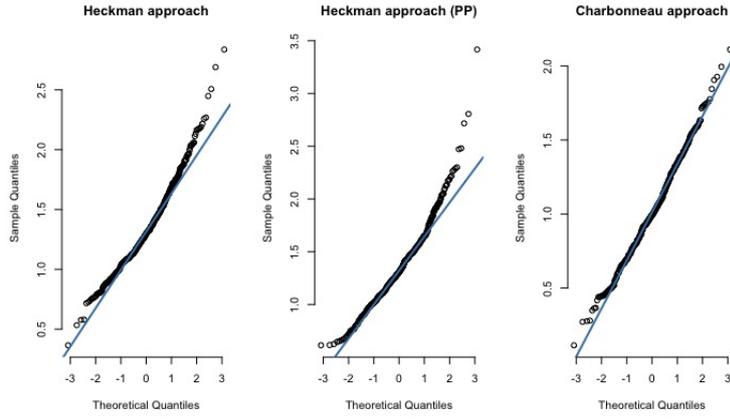
Figure 24: QQ plot of estimated β_{22}^* in Design 5
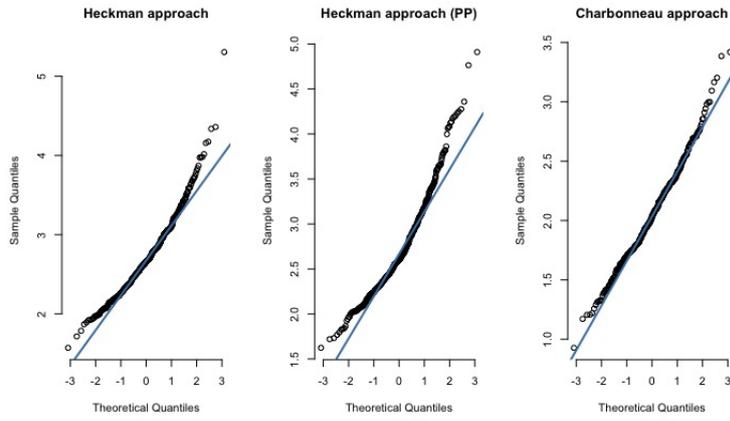
Figure 25: QQ plot of estimated β_{23}^* in Design 5
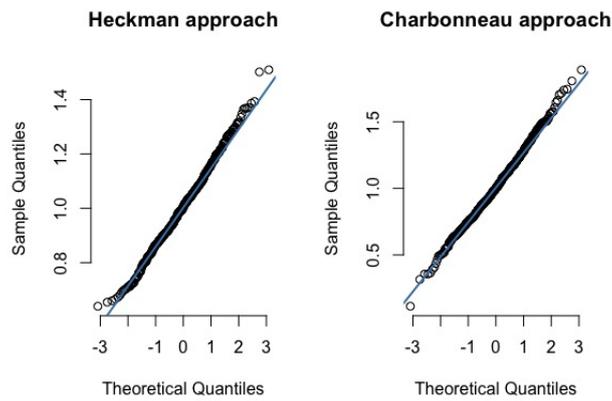
Figure 26: QQ plot of estimated β_{22}^* in Design 6
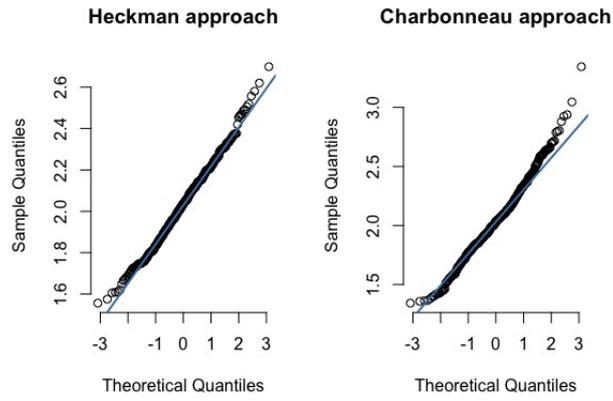
Figure 27: QQ plot of estimated β_{23}^* in Design 6
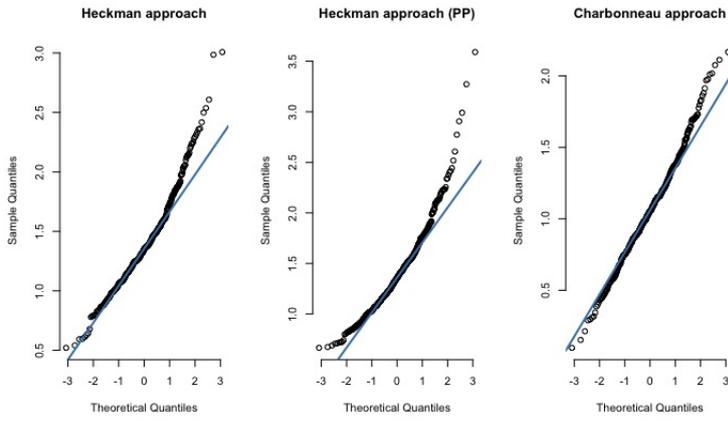
Figure 28: QQ plot of estimated β_{22}^* in Design 7
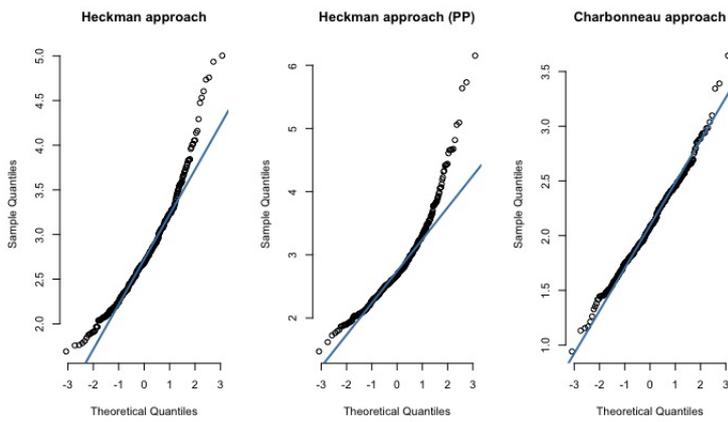
Figure 29: QQ plot of estimated β_{23}^* in Design 7

B.2 Second stage estimates
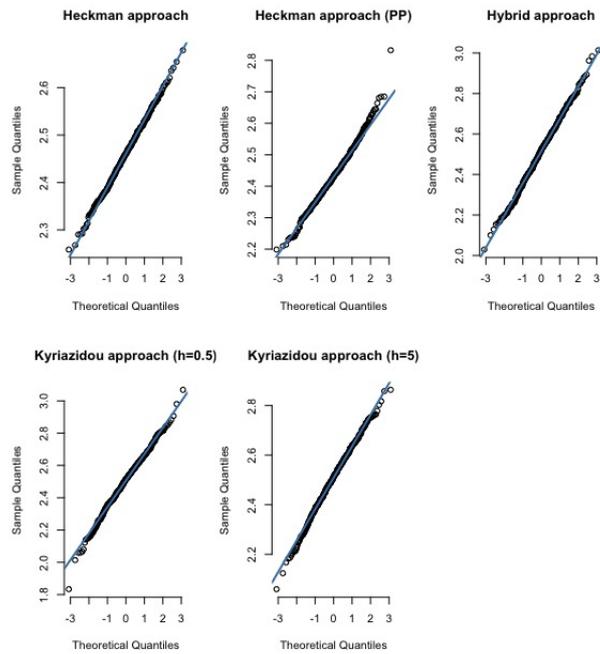
Figure 30: QQ plot of estimated β_{12} in Design 5

	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7
Hybrid							
β_{11}	-0.0052	0.0000	0.0001	0.0014	0.0014	0.0033	0.0001
	(0.0569)	(0.0747)	(0.0672)	(0.0750)	(0.0893)	(0.0745)	(0.0963)
β_{12}	0.0022	0.0042	0.0041	-0.0188	-0.0026	0.0036	-0.0194
	(0.0821)	(0.1336)	(0.1263)	(0.1426)	(0.1547)	(0.1313)	(0.1680)
Inverse Mills-Ratio	0.0013	0.0036	0.0143	-0.0112	0.0155	0.0212	-0.0030
	(0.1306)	(0.2508)	(0.2524)	(0.2738)	(0.3001)	(0.2627)	(0.2977)
Kyriazidou $h = 0.5$							
β_{11}	-0.0054	0.0011	-0.0060	-0.0046	0.0019	-0.0018	-0.0067
	(0.0759)	(0.0831)	(0.0779)	(0.0837)	(0.0935)	(0.0837)	(0.0887)
β_{12}	-0.0010	-0.0032	-0.0022	-0.0082	0.0075	0.0056	-0.0032
	(0.1305)	(0.1401)	(0.1304)	(0.1431)	(0.1587)	(0.1391)	(0.1551)
Kyriazidou $h = 0.5$, corrected							
β_{11}	-0.0054	0.0011	-0.0060	-0.0046	0.0019	-0.0018	-0.0068
	(0.0762)	(0.0834)	(0.0782)	(0.0840)	(0.0939)	(0.0840)	(0.0890)
β_{12}	-0.0010	-0.0033	-0.0023	-0.0082	0.0074	0.0056	-0.0031
	(0.1310)	(0.1407)	(0.1309)	(0.1437)	(0.1593)	(0.1396)	(0.1558)
Kyriazidou $h = 1$							
β_{11}	-0.0061	0.0011	-0.0057	-0.0042	0.0000	-0.0029	-0.0062
	(0.0685)	(0.0746)	(0.0697)	(0.0757)	(0.0841)	(0.0743)	(0.0788)
β_{12}	0.0001	-0.0017	-0.0013	-0.0124	0.0090	0.0071	-0.0083
	(0.1161)	(0.1228)	(0.1162)	(0.1254)	(0.1403)	(0.1270)	(0.1361)
Kyriazidou $h = 1$, corrected							
β_{11}	-0.0062	0.0010	-0.0058	-0.0043	0.0000	-0.0030	-0.0064
	(0.0687)	(0.0748)	(0.0699)	(0.0759)	(0.0843)	(0.0745)	(0.0791)
β_{12}	0.0001	-0.0018	-0.0015	-0.0125	0.0090	0.0070	-0.0083
	(0.1164)	(0.1231)	(0.1165)	(0.1257)	(0.1407)	(0.1273)	(0.1364)
Kyriazidou $h = 2$							
β_{11}	-0.0063	0.0010	-0.0048	-0.0029	-0.0012	-0.0036	-0.0050
	(0.0647)	(0.0699)	(0.0641)	(0.0699)	(0.0785)	(0.0686)	(0.0726)
β_{12}	0.0000	0.0001	0.0000	-0.0145	0.0084	0.0064	-0.0103
	(0.1098)	(0.1150)	(0.1097)	(0.1172)	(0.1324)	(0.1207)	(0.1256)
Kyriazidou $h = 2$, corrected							
β_{11}	-0.0066	0.0001	-0.0052	-0.0033	-0.0014	-0.0040	-0.0052
	(0.0649)	(0.0700)	(0.0643)	(0.0701)	(0.0787)	(0.0688)	(0.0728)
β_{12}	0.0000	0.0000	-0.0001	-0.0149	0.0082	0.0060	-0.0106
	(0.1100)	(0.1152)	(0.1099)	(0.1174)	(0.1327)	(0.1209)	(0.1258)

Table 10: Simulation results for the estimated coefficients for the observation Equation 47 with $N = 25$ and 500 iterations. For these results, we set the values of the parameters from the first stage Equation 49 to be equal to their true values. The values correspond to the mean bias of estimates, and the standard deviation is in parenthesis.

	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7
Kyriazidou $h = 3$							
β_{11}	-0.0060	0.0011	-0.0039	-0.0019	-0.0001	-0.0033	-0.0038
	(0.0632)	(0.0685)	(0.0622)	(0.0675)	(0.0763)	(0.0665)	(0.0700)
β_{12}	0.0000	0.0019	0.0000	-0.0139	0.0085	0.0065	-0.0097
	(0.1081)	(0.1128)	(0.1078)	(0.1146)	(0.1290)	(0.1176)	(0.1224)
Kyriazidou $h = 3$, corrected							
β_{11}	-0.0065	0.0001	-0.0045	-0.0024	-0.0013	-0.0038	-0.0042
	(0.0634)	(0.0686)	(0.0623)	(0.0677)	(0.0766)	(0.0667)	(0.0702)
β_{12}	0.0000	0.0013	0.0000	-0.0145	0.0081	0.0059	-0.0101
	(0.1083)	(0.1130)	(0.1080)	(0.1148)	(0.1293)	(0.1178)	(0.1226)
Kyriazidou $h = 5$							
β_{11}	-0.0042	0.0020	-0.0020	0.0001	0.0001	-0.0011	-0.0015
	(0.0615)	(0.0672)	(0.0606)	(0.0649)	(0.0740)	(0.0645)	(0.0673)
β_{12}	0.0029	0.0040	0.0021	-0.0117	0.0100	0.0082	-0.0080
	(0.1062)	(0.1103)	(0.1058)	(0.1123)	(0.1254)	(0.1141)	(0.1198)
Kyriazidou $h = 5$, corrected							
β_{11}	-0.0050	0.0014	-0.0028	0.0000	0.0000	-0.0018	-0.0020
	(0.0617)	(0.0674)	(0.0608)	(0.0651)	(0.0742)	(0.0647)	(0.0675)
β_{12}	0.0020	0.0032	0.0012	-0.0125	0.0094	0.0073	-0.0086
	(0.1065)	(0.1106)	(0.1060)	(0.1125)	(0.1257)	(0.1143)	(0.1201)
Kyriazidou $h = 10$							
β_{11}	0.0046	0.0090	0.0070	0.0091	0.0067	0.0073	0.0050
	(0.0583)	(0.0646)	(0.0578)	(0.0613)	(0.0702)	(0.0612)	(0.0637)
β_{12}	0.0147	0.0128	0.0123	-0.0026	0.0168	0.0180	-0.0016
	(0.1023)	(0.1074)	(0.1029)	(0.1090)	(0.1206)	(0.1096)	(0.1160)
Kyriazidou $h = 10$, corrected							
β_{11}	0.0037	0.0083	0.0061	0.0083	0.0062	0.0065	0.0045
	(0.0584)	(0.0648)	(0.0580)	(0.0615)	(0.0704)	(0.0614)	(0.0639)
β_{12}	0.0136	0.0120	0.0112	-0.0035	0.0162	0.0170	-0.0022
	(0.1025)	(0.1076)	(0.1031)	(0.1092)	(0.1208)	(0.1098)	(0.1162)

Table 11: Simulation results for the estimated coefficients for the observation Equation 47 with $N = 25$ and 500 iterations. For these results, we set the values of the parameters from the first stage Equation 49 to be equal to their true values. The values correspond to the mean bias of estimates, and the standard deviation is in parenthesis.
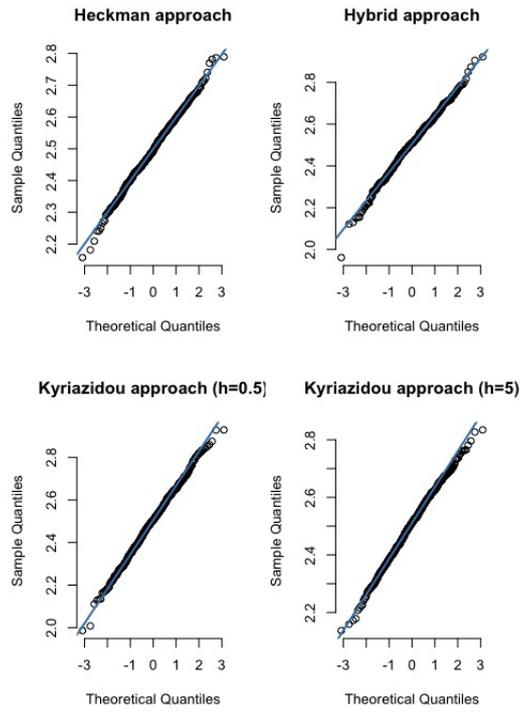
Figure 31: QQ plot of estimated β_{12} in Design 6
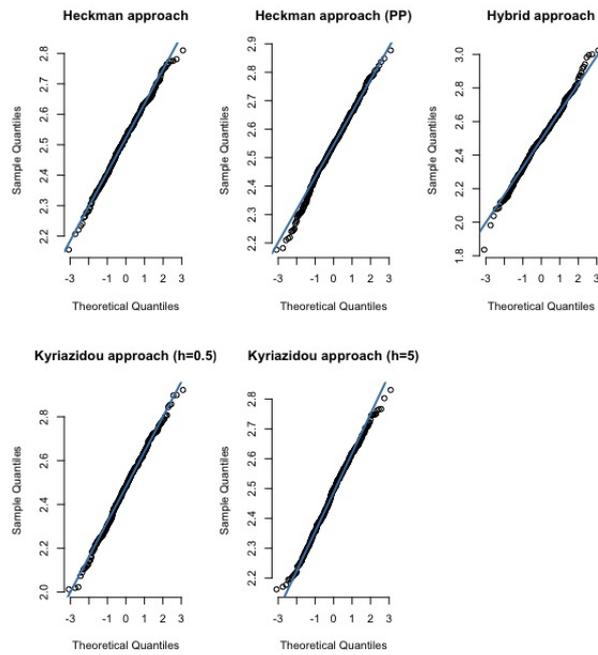
Figure 32: QQ plot of estimated β_{12} in Design 7