Estimating the extensive margin of trade

J.M.C. Santos Silva a,b,⁎, Silvana Tenreyro c, Kehai Wei a

a University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
b CEMAPRE, Portugal
c London School of Economics, CFM, CEP, and CEPRI, Department of Economics, s.579. St. Clement’s Building, Houghton St., London, WC2A 2AE, UK

A R T I C L E   I N F O

Article history:
Received 18 June 2013
Received in revised form 29 November 2013
Accepted 2 December 2013
Available online 25 December 2013

JEL classification:
C13
C25
C51
F11
F14

Keywords:
Bounded data
Estimation of trade models
Number of sectors

A B S T R A C T

Understanding and quantifying the determinants of the number of sectors or firms exporting in a given country is of relevance for the assessment of trade policies. Estimation of models for the number of exporting sectors, however, poses a challenge because the dependent variable has both a lower and an upper bound, implying that the partial effects of the explanatory variables on the conditional mean of the dependent variable cannot be constant. We argue that ignoring these bounds can lead to erroneous conclusions and propose a flexible specification that accounts for the doubly-bounded nature of the dependent variable. We empirically investigate the problem and the proposed solution, finding significant differences between estimates obtained with the proposed estimator and those obtained with standard approaches.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

In a landmark paper, Hummels and Klenow (2005) drew attention to the role of the extensive margin in explaining observed international trade patterns, giving origin to a burgeoning literature on its determinants and importance.

Building on Melitz’s (2003) model with heterogeneous firms, Helpman et al. (2008) and Chaney (2008), among others, developed trade models that explicitly consider the decision to export and therefore explicitly model the extensive margin of trade. In parallel, several authors have studied empirically how the extensive margin is affected by factors such as transportation costs, tariffs, or economic and political integration.

The extensive margin can be defined at different levels of aggregation and a variety of definitions have been used in empirical work. For example, Hillberry and Hummels (2008) work at the shipment level, Eaton et al. (2004), Berthou and Fontagné (2008) work at the firm level, Hillberry and McDaniel (2002), Hummels and Klenow (2005), and Dennis and Shepherd (2007) define the extensive margin at the sector-product level, and Helpman et al. (2008) consider data at the country level.

Naturally, the econometric methods used in the estimation of models for the extensive margin of trade depend on the level of aggregation that is considered and on the nature of the data available. For example, Berthou and Fontagné (2008), Baldwin and Di Nino (2006), and Helpman et al. (2008) use binary models to study whether a firm, or a sector, or a country exports, while Eaton et al. (2004), Hillberry and McDaniel (2002), Flam and Nordström (2006), and Dennis and Shepherd (2007) model the number of firms or sectors that export. While some of the models used in these studies are standard, the specification and estimation of models for the number of exporting sectors raises specific problems and is the focus of this paper.

The number of sectors exporting from origin country i to destination country j is a count and therefore it is a non-negative integer. Moreover, if the sectors or products are defined using a classification of economic...
activities such as the Harmonized Commodity Description and Coding System, the variate of interest has as an upper bound the number of classes in the system. That is, the variate of interest is bounded from below by zero and from above by the number of product categories.

The existence of these bounds implies that the partial effect of the regressors on the conditional mean of the dependent variable (the number of sectors) cannot be constant and must approach zero as the conditional mean approaches its bounds. Therefore, ignoring the nature of the data and simply using OLS, as in Flam and Nordström (2006), is likely to lead to erroneous conclusions because the linear model assumes that the partial effects are constant. Some authors have eliminated the lower bound by using the log of the number of sectors as the dependent variable, see, e.g., Eaton et al. (2004) and Hillberry and Hummels (2008). Alternatively, standard count data models, such as Poisson and negative binomial regressions have been used by Dennis and Shepherd (2007), Berthou and Fontagné (2008), and Persson (2013). However, all these approaches ignore the upper bound and therefore are also unsatisfactory; as we will illustrate, these estimators can lead to very misleading results.

In this paper we study the specification and estimation of models for the number of sectors exporting from country \( j \) to country \( i \). Building on the literature on fractional data (see Ramalho et al., 2011, for a recent survey), we suggest a flexible specification that takes into account the doubly-bounded nature of the data. The performance of the proposed estimator is evaluated both with simulations and in an empirical application. The advantage of the proposed approach over various alternatives previously used in the literature is clearly illustrated in both cases. The simulations show that the proposed specification is reasonably resilient and is capable of delivering fairly accurate results even in the presence of misspecification. In the application we find that the proposed model fits the data much better than standard alternatives and, more importantly, we find that while other methods yield economically implausible quantitative effects for various trade determinants (e.g., sharing a border, a common currency, or trade agreements) the new method yields economically reasonable effects.

2. The problem and the proposed solution

As in Armenter and Koren (2012), suppose that the goods in the economy are partitioned into \( S \) sectors according to some classification of economic activities, and let \( T_{ij} \) denote the number of sectors for which there are exports from country \( j \) to country \( i \). Our objective is to consider possible specifications and estimators for the conditional expectation \( E(T_{ij}|x_{ij}) \), where \( x_{ij} \) denotes a set of geographic and economic determinants of international trade measured at the country-pair level.

By construction, \( T_{ij} \) is such that \( 0 \leq T_{ij} \leq S \) and therefore its conditional expectation has the same non-stochastic bounds. This implies that it is always possible to write \( E(T_{ij}|x_{ij}) \) as the product of \( S \) by a function whose codomain is bounded by 0 and 1, such as one of the many specifications that have been used in binary choice models. That is, the expected value of the number of exporting sectors can be expressed as

\[
E(T_{ij}|x_{ij}) = SF(x_{ij})^\beta.
\]  

(1)

where \( \beta \) is a vector of parameters and \( F(x_{ij})^\beta \) can be interpreted as the probability that a randomly drawn sector in country \( j \) will export to destination \( i \).

To proceed it is necessary to specify a functional form for \( F(x_{ij}) \). The choice of this functional form is an empirical issue that has to be addressed on a case-by-case basis. Indeed, the shape of \( F(x_{ij}) \) will depend both on the classification that is used to define the sectors and on the set of regressors that is available in the particular application. Therefore, it is important to specify \( F(x_{ij}) \) in a flexible way and, as the results in Sections 3 and 4 will illustrate, it is particularly important to let \( F(x_{ij}) \) have a flexible degree of asymmetry so that the model can fit reasonably well both tails of the distribution. Although several models with this characteristic have been proposed (see, e.g., Ramalho et al., 2011), we suggest specifying

\[
F(x_{ij}) = 1 - \left(1 + \omega \exp(x_{ij})\right)^{-\omega},
\]  

(2)

where \( \omega > 0 \) is a shape parameter that allows the distribution to be symmetric (\( \omega = 1 \)), left-skewed (\( \omega < 1 \)), or right-skewed (\( \omega > 1 \)), as dictated by the data. This model is easy to estimate, is reasonably flexible, and has as special cases two well-known models: setting \( \omega = 1 \) we obtain the logit specification suggested by Papke and Wooldridge (1996) in a related context, and the complementary log–log model is obtained as a limiting case when \( \omega \to 0 \).

Putting Eqs. (1) and (2) together we get

\[
E(T_{ij}|x_{ij}) = S - S\left(1 + \omega \exp(x_{ij})\right)^{-\omega} + u_{ij},
\]  

(3)

Because Eq. (3) specifies a conditional expectation and \( S \) is a known constant, the model of interest can also be written as

\[
T_{ij}/S = 1 - \left(1 + \omega \exp(x_{ij})\right)^{-\omega} + u_{ij},
\]

where \( T_{ij}/S \) is bounded between 0 and 1, and \( u_{ij} \) is simply defined as \( u_{ij} = T_{ij}/S - E(T_{ij}|x_{ij}) \), which implies that \( E(u_{ij}|x_{ij}) = 0 \). Estimation of \( \beta \) and \( \omega \) can be performed in different ways. Because a detailed discussion of the different estimators is beyond the scope of this paper, here we simply follow Papke and Wooldridge (1996) and estimate the model by Bernoulli pseudo-maximum likelihood. This estimator is very easy to implement and it is consistent under very general conditions (see Courrèges et al., 1984). Specifically, as in Papke and Wooldridge (1996), we assume that the conditional variance of \( T_{ij}/S \) given \( x_{ij} \) is proportional to \( F(x_{ij})[1 - F(x_{ij})] \) and estimate \( \beta \) and \( \omega \) by maximizing an objective function with individual contributions of the form

\[
L(\beta, \omega) = \left(T_{ij}/S\right)\ln F(x_{ij})^\beta + \left(1 - T_{ij}/S\right)\ln \left(1 - F(x_{ij})^\beta\right),
\]  

(4)

where \( F(x_{ij})^\beta \) is given by Eq. (2). The first order conditions of Eq. (4) show that this estimator can be interpreted as a weighted non-linear least squares estimator of Eq. (3) that down-weights the observations

4 Strictly speaking, it is possible to avoid the specification of \( F(\cdot) \) by estimating it nonparametrically, for example using the estimators proposed by Ichimura (1993). However, for typical international trade problems, the implementation of this kind of estimator is too cumbersome to be routinely used.

5 To our knowledge, this specification was introduced by Santos Silva (2001) but not used since.

6 The model was obtained considering only the nature of \( T_{ij} \) and in particular the fact that it is bounded by 0 and \( S \). In Appendix 1 we show that, under suitable assumptions, a specification of this type can also be motivated by models such as those developed by Helpman et al. (2008), Chaney (2008), or Manova (2013).

7 Notice that the assumed heteroskedasticity pattern does not have to be correctly specified for the estimator to be consistent. Naturally, inference should be based on a “robust” estimator of the covariance matrix.
for which $F(x_i/\beta)$ is close to 0.5 because these are the observations that are likely to have larger variance.

One final point is worth emphasizing: given the non-linearity of $F(x_i/\beta)$ and the fact that we interpret it simply as an approximation to $E(T_{ij}|S_{x_i})$, or to the probability that a randomly drawn sector in country $j$ will export to destination $i$, the estimates of $\beta$ are not very informative. Therefore, inference should focus on the partial effects of the regressors of interest and not on the parameter estimates per se. In what follows we will focus on the average across the entire sample of the partial effect of the regressors on $E(T_{ij}/S_{x_i})$.

### 3. Simulation evidence

In this section we present the results of simulation experiments illustrating the performance of the proposed estimator and comparing it with that of other possible estimation approaches. The experiments were designed to be informative about the illustrative application to it with that of other possible estimation approaches. The experiments were performed with $S = 5000$, $\beta_0 = 0$, $\beta_1 = 0.25$, $\beta_2 = 1$, which again are chosen to mimic the estimates obtained in Section 4 with our preferred specification. The variables $T_{ij}$, $x_{1i}$, and $x_{2i}$ were newly generated for each of the 10,000 replications used in the experiments and all the simulations where performed in Stata (StataCorp., 2013).

We studied the performance of eight different combinations of specification and estimator, models for short.

The first model we consider was used by Flam and Nordström (2006) and specifies $E(T_{ij}|x_i) = x_i \beta$. The parameters are estimated by least squares and hence these results are labeled OLS.

The second model is the one used by Eaton et al. (2004) and by Hillberry and Hummels (2008), and specifies $E(\ln(T_{ij}|x_i) = x_i \beta$. Estimation is performed by OLS and these results are labeled LogLin.

The third one specifies $E(T_{ij}|x_i) = \exp(x_i \beta)$. Estimation is performed by Poisson (pseudo) maximum likelihood as in Dennis and Shepherd (2007), Berthou and Fontagné (2008), and Persson (2013); these results are labeled Poisson.

The fourth approach uses the same exponential specification for $E(T_{ij}|x_i)$ but in this case estimation is performed by negative binomial (pseudo) maximum likelihood as done by Persson (2013); these results are labeled NegBin.

The fifth model uses a two-limit Tobit, an estimator that many practitioners use when the variate of interest is doubly bounded. The specification of $E(T_{ij}|x_i)$ implicitly assumed by this model is given, e.g., by Wooldridge (2010, page 704, Eq. 17.66) and the results obtained with it are labeled TL-Tobit.

The sixth model specifies $E(T_{ij}|x_i)$ as the limit of Eq. (3) when $\omega$ passes to 0. Estimation is performed by Bernoulli (pseudo) maximum likelihood as described in the previous section and the results are labeled CLL because of the relation of this model with the complementary log–log.

The seventh model also specifies $E(T_{ij}^{\omega}|x_i)$ as in Eq. (3) but now $\omega$ is set to 1. Estimation is again performed by Bernoulli (pseudo) maximum likelihood and, due to its similarity with the estimator proposed by

### Table 1

Simulation results for $\delta = 1$ and different values of $\omega$.

| Case 1: $\omega = 0.50$ | Flex | $\beta$ | S.E. | 0.00 | 0.51 | 0.01 | 0.05 |
| Case 2: $\omega = 1.00$ | PWW | $\beta$ | S.E. | 4.43 | 2.10 | 0.03 | 0.22 |
| Case 3: $\omega = 2.50$ | CLL | $\beta$ | S.E. | −7.50 | 3.50 | −1.55 | 0.38 |
| Case 4: $\omega = 6.00$ | PWW | $\beta$ | S.E. | 100.87 | 42.97 | 1.03 | 3.88 |
| Case 5: $\omega = 10.00$ | TL-Tobit | $\beta$ | S.E. | 324.27 | 92.08 | −4.42 | 2.31 |
| Case 6: $\omega = 15.00$ | LogLin | $\beta$ | S.E. | 14.59 | 166.21 | 3.49 | 194.03 |

| Case 7: | Flex | $\beta$ | S.E. | 493.59 | 104.56 | 1.52 | 1.98 |
| Case 8: | PWW | $\beta$ | S.E. | 603.18 | 128.35 | 1.88 | 2.58 |
| Case 9: | CLL | $\beta$ | S.E. | 1.67 | 166.21 | 349.64 | 1940.13 |
| Case 10: | OLS | $\beta$ | S.E. | 1.55 | 0.38 | 2.78 | 0.62 |

Note that Eq. (5) is more general than Eq. (3), which is obtained as a special case when $\omega = 1$.\(^6\)

As for the regressors, $x_{1i}$ was generated as independent draws from a Bernoulli distribution with Pr($x_{1i} = 1$) = 0.01 and $x_{2i}$ was generated as independent draws from a normal distribution with $\mu = -7 + 3x_{1i}$ and $\sigma = 3 + x_{1i}$. The distribution of $x_{1i}$ was chosen to mimic the distribution of the common currency dummy used in the illustration presented in the next section, and the distribution of $x_{2i}$ mimics the distribution of a linear combination of the remaining regressors used in the model. All experiments were performed with $\delta = 0.00$, $\beta = 0.56$, $\beta_2 = 0.05$, which again are chosen to mimic the estimates obtained in Section 4 with our preferred specification. The variables $T_{ij}$, $x_{1i}$, and $x_{2i}$ were newly generated for each of the 10,000 replications used in the experiments and all the simulations where performed in Stata (StataCorp., 2013).\(^7\)

\(^6\) In particular, notice that the interpretation of $\beta$ depends on the value of $\omega$.

\(^7\) Both $\omega \neq 1$ and $\delta \neq 1$ imply asymmetric tails, but each parameter allows for a different type of asymmetry. Because it combines both shape parameters, the functional form of Eq. (5) is very flexible.
Papke and Wooldridge (1996), the results for this model are labeled P&W.

Finally, in the eighth model the specification of \( E(T|x_i) \) is again as in Eq. (3), but in this case the value of \( \omega \) is not restricted. Estimation is performed by Bernoulli (pseudo) maximum likelihood. The estimates obtained with this more flexible approach are labeled Flex.

We performed two sets of experiments. In the first set data were generated with \( \delta = 1.00 \) and \( \omega \in [0.50,1.00,2.50] \). Therefore, in these experiments Eq. (3) is always correctly specified, and the model proposed by Papke and Wooldridge (1996) is also correctly specified when \( \omega = 1.00 \). The main purpose of these experiments is to evaluate the performance of the estimator based on Eq. (3) when it is correctly specified and to gauge the size of the bias resulting from using one of the other models. In the second set of experiments we used \( \omega = 2.50 \) and \( \delta \in [0.50,0.80,1.25,2.00] \). Now all the models are misspecified and the objective is to evaluate the resilience of Flex, the estimator based on Eq. (3), to different degrees of misspecification. Together, the two sets of experiments cover a wide variety of data generating processes, with the partial effects being estimated ranging from about 6 to more than 150.

Table 1 presents the main results obtained in the first set of experiments. Specifically, for each of the three values of \( \omega \) that were considered, the table reports the average across the entire sample of the partial effects of \( x_1 \) and \( x_2 \) (denoted PE\(_{x_1}\) and PE\(_{x_2}\), respectively), as well as the bias and the standard errors of the estimates of these partial effects obtained by each of the eight methods. For the continuous regressor \( x_2 \) the partial effects are just the derivatives of the estimate of \( E(T|x_i) \) with respect to \( x_2 \), while for the dummy variable \( x_1 \) the partial effect is the difference between the estimate of \( E(T|x_i) \) with the dummy equal to 1 and with the dummy equal to 0; the results reported for the Loglin model are the partial effects on the exponential of the fitted values of InT, averaged over all observations. Table 2 presents similar results for the second set of experiments.

The results in Table 1 show that, naturally, the results obtained with Flex are very good in all the three cases considered. For \( \omega = 1.00 \) P&W is also correctly specified and in this case its results are similar to those obtained with Flex. For other values of \( \omega \), however, P&W leads to excellent estimates of PE\(_{x_1}\), but to sizable biases in the estimate of PE\(_{x_2}\). This is particularly clear for the simulations with \( \omega = 2.50 \), the case that more closely resembles the one in the empirical illustration in Section 4. In this case the partial effect of the discrete regressor has a downward bias of more than 50%.

The performance of all the other models considered is very poor and all of them can lead to biases that can even exceed the partial effect being estimated, sometimes by orders of magnitude. It is also interesting to note that even closely related models, such as Poisson and NegBin, can lead to widely different results.

The results in Table 2 are particularly interesting in that they suggest that Flex can perform relatively well even in presence of some degree of misspecification. In particular, the results for Flex with moderate misspecification, i.e., for \( \delta \in (0.80,1.25) \), are excellent. Naturally, when more severe misspecification is present the results are less satisfactory: with \( \delta = 0.50 \) the bias of PE\(_{x_1}\) is still quite small, but the results for PE\(_{x_2}\) are not so good, while for \( \delta = 2.00 \) both partial effects are overestimated by about one third. Still, in these experiments Flex is not outperformed by any of its competitors and the results in Table 2 show that the estimator is reasonably resilient, being capable of delivering fairly accurate results even in the presence of some degree of misspecification.

The results of P&W are also worth mentioning. Indeed, except for \( \delta = 2.00 \), P&W provides fairly good estimates of PE\(_{x_1}\), comparable to those of Flex. However, for PE\(_{x_2}\) the biases of P&W are much larger. As before, the performance of all the other estimators is very poor, all of them leading to large biases.

These simulation results clearly show that the choice of estimator matters; indeed, it can matter a lot. All of the approaches previously
used to estimate the determinants of the number of exporting sectors can lead to highly biased results, and therefore it is important to make an effort to ensure that the model used in practice provides a good description of the data. Our results also show that it is perfectly possible for a misspecified model to lead to good estimates of the partial effect of one of the regressors, while completely failing in the estimation of the effect of another. This is what happens, for example, with P&W in the first set of experiments when $\omega = 2.50$. Finally, although the simulation results suggest that the proposed model is sufficiently flexible to produce accurate results in many situations, it is clear that its appropriateness should be checked in each application because it will also lead to biased results when the proposed functional form is not adequate.

4. Empirical application

We have argued for a different method to specify and estimate models for the extensive margin of trade; whether the use of this approach makes a material difference is an empirical question. To investigate this matter we estimated a model for the number of sectors exporting from a given country to a destination. The sectors are defined using the 1996 revision of the Harmonized Commodity Description and Coding System at the 6-digit level, which has 5132 categories, and the data were obtained from UN Comtrade for 2001; Table A1 in Appendix 2 lists the 218 countries and territories for which we were able to obtain data for this study.

Data for the regressors were obtained essentially from the CIA’s World Factbook and CEPII. In particular, the CEPII database was used to construct the following regressors: \textit{LOG DISTANCE}, defined as the natural logarithm of distance between capitals (in kilometers); \textit{BORDER}, a dummy that equals 1 when the two countries share a land border; \textit{COLONIAL TIE}, a dummy that equals 1 either if the importer has ever colonized or been a colony of the exporter or if the two countries were once part of the same country; \textit{COMMON LANGUAGE}, a dummy that equals 1 if either country has land borders; \textit{COMMON CURRENCY}, a dummy that equals 1 if both countries are at least in one common region; \textit{COMMON TIE}, a dummy that equals 1 if both countries use the same currency or if the exchange rates between their currencies is fixed. The CIA’s World Factbook was used to construct two additional dummies: \textit{BOTH ISLANDS}, which equals 1 if neither country has land borders; \textit{BOTH ISLANDS}, which equals 1 if both countries are landlocked. Finally, the variable \textit{RELIGION} was constructed as in Helpman et al. (2008); that is, the variable is the sum of the products of the shares of the population in each of the partners that are Catholic, Muslim, or Protestant. The information used to construct this variable is from multiple sources that include the CIA’s World Factbook, Wikipedia, and the work of Kettani (2010a,b,c,d,e). Finally, the model includes importer and exporter dummies, the multilateral resistance terms suggested by Anderson and van Wincoop (2003).

These data were used to estimate six of the eight models considered in the previous section; CLL and the TL-Tobit were not used here both because they have never been used in this context and because the simulation results show that their performance is generally poor.

Table 3 presents the estimates obtained with the different models and the respective $R^2$, defined as the square of the correlation between $T_{ij}$ and the corresponding estimate of $E(T_{ij}|x_{ij})$. Table 4 presents the average across the entire sample of the partial effects of each of the regressors on $E(T_{ij}|x_{ij})$; as usual, for the continuous variables (\textit{LOG DISTANCE} and \textit{RELIGION}) these are just the derivatives of the estimate of $E(T_{ij}|x_{ij})$ with respect to regressors (notice that the derivative is with respect to log distance, not distance itself), while for the dummy variables the partial effect is the difference between the estimate of $E(T_{ij}|x_{ij})$ with the dummy equal to 1 and with the dummy equal to 0. To provide a visual assessment of the goodness-of-fit of each of the six models considered, Fig. 1 displays the plots of $T_{ij}$ and of the

\begin{table}
\centering
\begin{tabular}{lcccccc}
\hline
 & OLS & LogLin & Poisson & NegBin & PBW & Flex \\
\hline
\textit{LOG DISTANCE} & -72.66 & -0.91 & -0.60 & -1.20 & -0.90 & -1.07 \\
 & (4.72) & (0.02) & (0.02) & (0.02) & (0.02) & (0.03) \\
BORDER & 444.89 & 0.49 & -0.14 & 0.96 & 0.42 & 0.59 \\
 & (55.21) & (0.09) & (0.08) & (0.12) & (0.08) & (0.09) \\
BOTH ISLANDS & -0.23 & 0.31 & 0.41 & 0.44 & 0.45 & 0.53 \\
 & (8.61) & (0.06) & (0.07) & (0.07) & (0.07) & (0.08) \\
BOTH LANDLOCKED & -2.14 & 0.25 & 0.26 & 0.30 & 0.04 & 0.16 \\
 & (12.15) & (0.06) & (0.11) & (0.08) & (0.10) & (0.09) \\
COLONIAL TIE & 291.39 & 0.70 & 0.49 & 1.03 & 0.76 & 0.97 \\
 & (59.15) & (0.08) & (0.07) & (0.09) & (0.07) & (0.08) \\
COMMON CURRENCY & 107.21 & -0.09 & -0.25 & 0.74 & 0.09 & 0.25 \\
 & (54.13) & (0.09) & (0.08) & (0.12) & (0.07) & (0.09) \\
RTA & 547.79 & 0.36 & 0.13 & 0.20 & 0.24 & 0.33 \\
 & (24.34) & (0.04) & (0.05) & (0.05) & (0.04) & (0.05) \\
COMMON LANGUAGE & 34.04 & 0.63 & 0.39 & 0.70 & 0.57 & 0.64 \\
 & (7.19) & (0.03) & (0.04) & (0.04) & (0.04) & (0.04) \\
BOTH WTO & 146.61 & 0.48 & 0.43 & 0.19 & 0.61 & 0.73 \\
 & (6.36) & (0.05) & (0.10) & (0.07) & (0.10) & (0.10) \\
RELIGION & 0.23 & 0.40 & 0.37 & 0.53 & 0.35 & 0.41 \\
 & (9.26) & (0.04) & (0.05) & (0.07) & (0.05) & (0.06) \\
Overdispersion parameter & - & - & - & - & - & - \\
$\omega$ & - & - & - & - & - & - \\
$\epsilon$ & - & - & - & - & - & - \\
$R^2$ & 0.56 & 0.18 & 0.76 & 0.07 & 0.92 & 0.92 \\
Sample size & 46,872 & 24,889 & 46,872 & 46,872 & 46,872 & 46,872 \\
\hline
\end{tabular}
\caption{Parameter estimates (and standard errors).}
\end{table}

Note: All models include importer and exporter dummies.

11 This variable has the obvious shortcoming of only accounting for three religions; for example, India and Nepal have a low value because the majority of the population in both countries is Hindu. However, we include this variable for consistency with Helpman et al. (2008). For more on the links between religion and economic activity, see Barro and McCleary (2003).

12 For comparability, in the LogLin model the $R^2$ is the square of the correlation (over the entire sample) between $T_{ij}$ and the exponential of the fitted values of $\ln T_{ij}$.

13 As before, the results reported for the LogLin model are the partial effects on the exponential of the fitted values of $\ln T_{ij}$ averaged over all observations.

14 It is important to keep in mind that because the results in Table 3 are averages of the partial effects across the entire sample, the actual partial effects for a given observation can be much smaller or much larger than the values reported here.
parametric fit of $E(T_{ij} | x_{ij})$ versus the estimated linear index, say $x' \beta$. To aid in the assessment of the fit, these plots also include non-parametric estimates of $E(T_{ij} | x_{ij})$, obtained by running a kernel regression of $T_{ij}$ on the values of $x' \beta$ obtained for each model.\footnote{Kernel regressions were performed in Stata (StataCorp, 2013) using the Gaussian kernel and the default bandwidth. For the LogLin model the nonparametric fit is the kernel regression of $T_{ij}$ on the exponential of the fitted values of $\ln T_{ij}$.}

In this example the OLS estimates generally have the expected sign but the magnitudes of some marginal effects appear to be clearly exaggerated. For example, the average increase in the number of sectors exporting from country $j$ resulting from being part of the same regional trade agreement is estimated to be almost 550, an increase that is more than 10% of the total number of sectors considered. The plot in the top-left corner of Fig. 1 clearly illustrates the inappropriateness of the linear model in this case. Indeed, we see that the fitted values of $E(T_{ij} | x_{ij})$ can be below zero and never get close to the upper bound of 5132. As a consequence, the parametric and non-parametric fits are far from each other. This implies that the partial effects are mismeasured for most observations and therefore it is not surprising that their average is sometimes quite unrealistic.

Results for the models that only take into account the lower bound of the observations with large values of $E(T_{ij} | x_{ij})$ can be far exaggerated. For example, the average increase in the number of sectors.

These plots also show the advantage of the proposed model over P&W. Indeed, the parametric and non-parametric fits for Flex are generally much closer to each other, especially for the upper part of the distribution. The reason for this difference in the ability to fit the upper tail of the distribution is easy to understand. The bulk of the observations are located in the lower tail; consequently these observations have a large influence in determining the shape of the estimated function. This means that in any model with a rigid functional form, the lower tail will tend to fit much better than the upper tail because a poor fit in the upper tail has relatively little impact on the value of the objective function. To be able to have a reasonable fit in both tails of the distribution it is necessary to allow the model to have a flexible degree of asymmetry, and that is what is achieved by the inclusion of the shape-parameter $\alpha$ in the Flex.

The advantage of the flexible specification is confirmed by noticing that P&W is rejected against the proposed model (the additional parameter $\alpha$ is significantly different from 1; see Table 3), and that this one is not rejected when tested against a more general specification.\footnote{Extreme examples of this are the Poisson and NegBin models that fit the lower tail of the distribution reasonably well but have a disastrous performance in the upper tail.}

The differences between the results of P&W and Flex are not restricted to their statistical properties. Indeed, although the average partial effects obtained with the two estimators are generally similar, for some of the regressors there are significant differences. In particular, the P&W model leads to an estimated average partial effect of COMMON CURRENCY equal to 8 sectors, much smaller than the estimate of 21 sectors obtained with the proposed model. Moreover, the coefficient of this regressor is not statistically significant in the P&W model, but it is significant in the more flexible alternative. These results parallel those in the simulations, where we found that often P&W led to good estimates of the partial effects of the continuous variable, but in many cases severely underestimated the partial effect of $x_1$, which was generated to mimic COMMON CURRENCY.

In short, this example illustrates that the choice of specification used can make a material difference for the results one obtains. In particular we find that even when using data at the 6-digit level it is vital to use models that specifically account for the upper-bound in the data;

\begin{table}[h]
\centering
\begin{tabular}{lcccccc}
\hline
 & OLS & LogLin & Poisson & NegBin & P&W & Flex \\
\hline
LOG DISTANCE & $-72.66$ & $-263.53$ & $-87.44$ & $-2574.08$ & $-86.86$ & $-86.04$ \\
 & (0.000) & (1.000) & (1.000) & (1.000) & (0.000) & (0.000) \\
BORDER & 444.89 & 152.69 & $-19.72$ & 1908.71 & 44.76 & 53.82 \\
 & (0.000) & (1.000) & (1.000) & (0.000) & (0.000) & (0.000) \\
BOTH ISLANDS & $-0.23$ & 106.76 & 72.68 & 1192.84 & $-8.23$ & 13.53 \\
 & (0.097) & (1.000) & (1.000) & (0.000) & (1.000) & (0.091) \\
BOTH LANDLOCKED & $-2.14$ & 82.87 & $-8.23$ & 736.35 & 3.89 & 13.53 \\
 & (0.060) & (1.000) & (1.000) & (0.000) & (1.000) & (0.000) \\
COLONIAL TIE & 291.39 & 277.10 & 90.22 & 3558.86 & 86.35 & 95.64 \\
 & (0.000) & (1.000) & (1.000) & (0.000) & (0.000) & (0.000) \\
COMMON CURRENCY & $-26.27$ & 107.21 & $-32.85$ & 1680.78 & $-8.25$ & 20.43 \\
 & (0.048) & (1.000) & (1.000) & (0.244) & (0.007) & (0.000) \\
RTA & 547.79 & 98.72 & 19.19 & 402.52 & 23.66 & 28.00 \\
 & (0.000) & (1.000) & (1.000) & (0.000) & (0.000) & (0.000) \\
COMMON LANGUAGE & 34.04 & 209.91 & 66.37 & 1617.70 & 59.92 & 56.30 \\
 & (0.000) & (1.000) & (1.000) & (0.000) & (0.000) & (0.000) \\
BOTH WTO & 146.61 & 114.63 & 55.70 & 378.50 & 54.67 & 55.63 \\
 & (0.000) & (1.000) & (1.000) & (0.000) & (0.000) & (0.000) \\
RELIGION & 0.23 & 114.79 & 54.03 & 1137.58 & 33.14 & 33.21 \\
 & (0.980) & (1.000) & (1.000) & (0.000) & (0.000) & (0.000) \\
\hline
\end{tabular}
\caption{Average partial effects (and p-values).}
\end{table}

The differences between the results of P&W and Flex are not restricted to their statistical properties. Indeed, although the average partial effects obtained with the two estimators are generally similar, for some of the regressors there are significant differences. In particular, the P&W model leads to an estimated average partial effect of COMMON CURRENCY equal to 8 sectors, much smaller than the estimate of 21 sectors obtained with the proposed model. Moreover, the coefficient of this regressor is not statistically significant in the P&W model, but it is significant in the more flexible alternative. These results parallel those in the simulations, where we found that often P&W led to good estimates of the partial effects of the continuous variable, but in many cases severely underestimated the partial effect of $x_1$, which was generated to mimic COMMON CURRENCY.

In short, this example illustrates that the choice of specification used can make a material difference for the results one obtains. In particular we find that even when using data at the 6-digit level it is vital to use models that specifically account for the upper-bound in the data;
naturally this will be even more important if coarser classifications are used. In the example presented here the proposed flexible specification clearly outperforms its competitors. This is an encouraging result in that it suggests that the model is flexible enough to describe adequately the type of data we are considering. Although the choice of the appropriate specification to use is an issue that needs to be carefully studied in each application,18 our results suggest that the proposed specification can be a good starting point.

5. Conclusions

Understanding and quantifying the factors affecting the number of sectors exporting in a given country is potentially relevant for the assessment of the effects of different trade policies. This paper studies models for the number of sectors exporting from a country to a given destination. We argue that standard estimation methods previously used in the literature are not suitable due to the nature of the dependent variable, the number of sectors, which has both a lower and an upper bound (the latter being the number of classes in the classification system). The existence of these bounds implies that the partial effects of the explanatory variables on the conditional mean of the dependent variable cannot be constant and must approach zero when the dependent variable approaches its bounds. Ignoring the nature of the data and simply using OLS or count-data models that ignore the upper bound is likely to lead to erroneous conclusions due to the severe misspecification of the models used. Moreover, as our simulation results illustrate, just accounting for the lower and upper bounds is not enough to ensure reliable inference: it is important to use flexible specifications to ensure that the models fit the data reasonably well.

We propose a flexible approach that takes into account the doubly-bounded nature of the dependent variable and, both with simulations and with an empirical application using country-pair data, we compare its performance to that of alternative specifications previously used in the literature. The proposed approach clearly outperforms the traditional

---

18 Needless to say that this applies very generally. For example, when using binary models to study exporting decisions, as done for example by Berthou and Fontagné (2008), Baldwin and Di Nino (2006), and Helpman et al. (2008), researchers should not simply rely on off-the-shelf estimators such as the logit and probit, and should make sure the estimator used adequately describes the data.
Appendix 1

In this appendix we illustrate how the specification of Eq. (3) can be linked to the structural model for trade developed by Helpman et al. (2008), hereafter HMR. In their model, the operating profits for a firm of country j selling in country i are given by

$$\pi^{ij} = (1 - \alpha) \left( \frac{\tau_{ij} + c_{ij}}{\alpha P_i} \right)^{-1 - \epsilon} Y_i - c_{ij} f_{ij},$$

where $\alpha$ is the number of bundles of inputs needed for the firm to obtain one unit of product, $c_i$ is the cost of each bundle in country j, $P_i$ is the price index in country i, $Y_i$ is the income in country i, $f_{ij}$ is proportional to the fixed cost of exporting from j to i, $\tau_{ij}$ is the “melting iceberg” variable cost of exporting from j to i, and $\alpha \in (0,1)$ is a parameter such that $\alpha = 1/(1 - \alpha)$ is the elasticity of substitution across products. The firm exports to market i if $\pi^{ij}(a) > 0$ or, equivalently, if

$$\left( \frac{1 - \alpha}{c_{ij}} \right) \left( \frac{\tau_{ij} + c_{ij}}{\alpha P_i} \right)^{1 - \epsilon} Y_i > 1,$$

which, taking logs on both sides, leads to

\begin{align*}
0 &< \ln(1 - \alpha) - \ln c_{ij} - \ln f_{ij} + \ln Y_i + (1 - \epsilon) \left( \ln \tau_{ij} + \ln c_{ij} + \ln a - \ln \alpha - \ln P_i \right), \\
0 &< \theta + \psi_i + \psi_j - \ln f_{ij} + \frac{\alpha}{\alpha - 1} \ln \tau_{ij} + \frac{\alpha}{\alpha - 1} \ln a - \ln \alpha - \ln P_i, \\
\ln a &< \alpha \left( \theta + \psi_i + \psi_j - \ln f_{ij} \right) - \ln \tau_{ij},
\end{align*}

where $\theta = \ln(\alpha^{\epsilon - 1} - 1)$, $\psi_i = \ln(Y_iP_i^{\epsilon - 1} - 1)$, and $\psi_j = -\ln c_{ij}$. Notice that $c_{ij}$, $f_{ij}$, and $\tau_{ij}$ are assumed not to depend on the identity of the producer, but $a$ is a firm-specific random variable.

Suppose now that the firms in country j are partitioned into S sectors according to some classification of economic activities. Then, the condition for sector s $\in \{1,..,S\}$ of country j to export to i is that there is at least one firm in the sector for which $\pi^{ij}(a) > 0$. Therefore, the probability that sector s from country j exports to destination i is given by

$$\Pr \left( \ln a < \frac{1 - \alpha}{\alpha} \left( \theta + \psi_i + \psi_j - \ln f_{ij} \right) - \ln \tau_{ij} \right) = \int_{\ln a}^\infty f_{\ln a}(z|x_\beta) dz = F_{\ln a}(x_\beta),$$

where $a_s$ denotes the minimum value of $a$ for firms in sector s. $f_{\ln a}(\cdot)$ is the conditional density of $\ln a$, for sector s, $x_\beta = (1 - \alpha) \left( \theta + \psi_i + \psi_j - \ln f_{ij} \right)/\alpha - \ln \tau_{ij}$. $\beta$ denotes a vector of regressors including importer and exporter dummies and variables measuring the trade frictions between i and j, and $\beta$ is a conformable vector of parameters, and we let $F_{\ln a}(\cdot)$ vary with s because the distribution of $\ln a_s$ does not have to be the same for every sector.

Now let $T_{ij}$ be an indicator variable that is 1 when at least one firm from sector s in country j exports to country i, being 0 otherwise, and notice that $E(T_{ij}|x_i) = \Pr(T_{ij} = 1|x_i) = F_{\ln a}(x_\beta)$. Additionally, define $T_{ij} = \sum_{s=1}^S T_{ij}^s$ as the number of sectors exporting from j to i. Hence, conditioning on $x_i$, the expected value of the number of exporting sectors is

$$E(T_{ij}|x_i) = \sum_{s=1}^S F_s(x_\beta).$$

Notice that for $S = 1$ this model is very similar to the first step of the model considered by HMR in which $T_{ij}$ is just an indicator of whether country j exports to i (see Eq. (12) in HMR). However, we adopt a very different stochastic specification: here the unobservable $a_s$ is the source of randomness and we treat the other variables as given; in contrast HMR treat $a_s$ as given and the randomness of the exporting decision appears due to the unobservability of some elements of $f_{ij}$ and $\tau_{ij}$, which are viewed as random variables. In our model the possible presence of these unobserved costs only changes the form of $f_{\ln a}(\cdot)$.

If sectoral data are available, it may be possible to use binary models to estimate how trade frictions affect the conditional expectation of $T_{ij}$. This is done, for example by Baldwin and Di Nino (2006) and Hillberry and Hummels (2008). However, researchers often prefer to model $E(T_{ij}|x_i)$, which can be expressed as

$$E(T_{ij}|x_i) = SF(x_\beta),$$

where $F(x_\beta) = S^{-1} \sum_{s=1}^S F_s(x_\beta)$ is the probability that a randomly drawn sector in country j will export to destination i.

We proceed by specifying a functional form for $F(\cdot)$. The fact that $F(x_\beta)$ is the distribution of a minimum suggests that the complementary log-log model is a useful starting point. However, because restrictive distributional assumptions are unlikely to be valid in practice, we suggest specifying

$$F(x_\beta) = 1 - \left( 1 + \omega \exp(x_\beta) \right)^{-\theta},$$

where $\omega > 0$ is a shape parameter. This model is reasonably flexible and has the complementary log-log model as a limiting case when $\omega \to 0$. This choice of functional form corresponds to the assumption that the distribution of $a_s$ for a randomly picked sector is a generalized Pareto with location parameter equal to 0 and scale parameter equal to 1. The form of Eq. (6) suggests that $F(x_\beta)$ could also be specified as a mixture model. This approach, however, is computationally and statistically more demanding and therefore we do not pursue it here.

In this appendix we have used the model developed by HMR to motivate the specification of Eqs. (8) and (3). Alternatively we could have used as starting points the models by Chaney (2008) or Manova (2013), which explicitly consider the existence of different sectors. However, because we consider only the case where no sectoral information is used, starting from the models by Chaney (2008) or Manova (2013) would have led exactly to the same result.

---

19 See the second equation on page 450 in HMR.


21 Indeed, $E(T_{ij}|x_i) = \sum_{s=1}^S \int_{\ln a}^\infty f_{\ln a}(z|x_\beta) dz = \int_{\ln a}^\infty \sum_{s=1}^S f_{\ln a}(z|x_\beta) dz$. The result follows by letting $\int_{\ln a}^\infty \sum_{s=1}^S f_{\ln a}(z|x_\beta) dz = F_{\ln a}(x_\beta)$, where $\sum_{s=1}^S 1^{-\theta} f_{\ln a}(\cdot)$ is the conditional density of $\ln a$, for a randomly picked sector.

22 The complementary log-log model would be valid under the assumptions that $\ln a_s$ follows the Gumbel (extreme value type I) distribution for a minimum and that $F_{\ln a}(x_\beta) = F_{\ln a}(x_\beta)$, $\forall i$. 
## Appendix 2

### Table A1

List of countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Central Asia</td>
</tr>
<tr>
<td>Albania</td>
<td>Europe</td>
</tr>
<tr>
<td>Algeria</td>
<td>Europe</td>
</tr>
<tr>
<td>Andorra</td>
<td>Europe</td>
</tr>
<tr>
<td>Angola</td>
<td>Europe</td>
</tr>
<tr>
<td>Anguilla</td>
<td>Caribbean</td>
</tr>
<tr>
<td>Antigua &amp; Barbuda</td>
<td>Caribbean</td>
</tr>
<tr>
<td>Argentina</td>
<td>South America</td>
</tr>
<tr>
<td>Armenia</td>
<td>Asia</td>
</tr>
<tr>
<td>Aruba</td>
<td>Asia</td>
</tr>
<tr>
<td>Australia</td>
<td>Oceania</td>
</tr>
<tr>
<td>Austria</td>
<td>Europe</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>Europe</td>
</tr>
<tr>
<td>Bahamas</td>
<td>Oceania</td>
</tr>
<tr>
<td>Bahrain</td>
<td>Asia</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>Asia</td>
</tr>
<tr>
<td>Barbados</td>
<td>Oceania</td>
</tr>
<tr>
<td>Belarus</td>
<td>Europe</td>
</tr>
<tr>
<td>Belgium</td>
<td>Europe</td>
</tr>
<tr>
<td>Benin</td>
<td>Africa</td>
</tr>
<tr>
<td>Bermuda</td>
<td>North America</td>
</tr>
<tr>
<td>Bhutan</td>
<td>Asia</td>
</tr>
<tr>
<td>Bolivia</td>
<td>South America</td>
</tr>
<tr>
<td>Bosnia Herzegovina</td>
<td>Europe</td>
</tr>
<tr>
<td>Botswana</td>
<td>South America</td>
</tr>
<tr>
<td>Br. Virgin Isds</td>
<td>Europe</td>
</tr>
<tr>
<td>Brazil</td>
<td>South America</td>
</tr>
<tr>
<td>Brunel Darussalam</td>
<td>Asia</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>Europe</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>Oceania</td>
</tr>
<tr>
<td>Burundi</td>
<td>Asia</td>
</tr>
<tr>
<td>Cambodia</td>
<td>Asia</td>
</tr>
<tr>
<td>Cameroon</td>
<td>Africa</td>
</tr>
<tr>
<td>Cape Verde</td>
<td>Europe</td>
</tr>
<tr>
<td>Cayman Isds</td>
<td>Oceania</td>
</tr>
<tr>
<td>Central African Rep.</td>
<td>Africa</td>
</tr>
<tr>
<td>Chad</td>
<td>Africa</td>
</tr>
<tr>
<td>Chile</td>
<td>Asia</td>
</tr>
<tr>
<td>China</td>
<td>Asia</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>Asia</td>
</tr>
<tr>
<td>Macao</td>
<td>Asia</td>
</tr>
<tr>
<td>Christmas Isds</td>
<td>Asia</td>
</tr>
<tr>
<td>Cocos Isds</td>
<td>Asia</td>
</tr>
<tr>
<td>Cook Isds</td>
<td>Asia</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>Asia</td>
</tr>
<tr>
<td>Croatia</td>
<td>Europe</td>
</tr>
<tr>
<td>Cuba</td>
<td>South America</td>
</tr>
<tr>
<td>Cyprus</td>
<td>Europe</td>
</tr>
<tr>
<td>Czech Rep.</td>
<td>Europe</td>
</tr>
</tbody>
</table>

### References


Ichimura, H., 1993. *Semiparametric least squares (SLS) and weighted SLS estimation of single-index models*. J. Econ. 58, 71–120.


