

Product differentiation, multi-product firms and estimating the impact of trade liberalization on productivity

Supplementary material: Data and Code

Jan De Loecker

Abstract

I provide more information on the three different data sources used in the empirical analysis. Appendix A of the paper already provides a detailed discussion of the main variables used in the analysis and how the datasets are structured. In this file I refer the reader to the specific data sources and data providers. In addition, I briefly provide the main resources for estimating the model described in section 4 of the paper. Where possible I refer to publicly available code using standard statistical software or code from other published work, which can be used in the estimation routine.

1 Data

I briefly discuss the three main data sources used in the empirical analysis.

1.1 Production Data

The plant-level production data is collected by the National Bank of Belgium and contains every unique establishment with a tax identifier. The database is commercialized by Bureau Van Dijk under the product BELFIRST. A preview of the data can be found at <https://belfirst.bvdep.com>, and full access requires purchasing the data either under a personal license or through an agreement with a research institution.¹ A reduced version of this data is available through AMADEUS, <http://www.bvdinfo.com>, another product of Bureau Van Dijk which is offered through the Wharton Research Data Services (WRDS).

The data contains standard information on firm-level production and similar data has been used throughout the literature.² In particular and as mentioned

¹I would like to thank Joep Konings at the K.U. Leuven for giving me access to the data through a site license at LICOS and the department of economics at K.U.L while I was a graduate student.

²see for example Olley and Pakes (1996) and Levinsohn and Petrin (2003).

in the paper under section 3.2.1, the data represents the population of producers of textile products over the period 1994-2002. The estimation of the production function, as described in section 4, requires information on plant-level revenues, value added, input use: labor as measured by full time equivalent production workers, raw materials and a measure of the capital stock. The latter is constructed from the balance sheet information on total fixed assets broken down into 1) Machinery and equipment, 2) Land and Buildings and 3) Furniture and Vehicles. Appropriate depreciation rates (based on actual depreciation rates) are used to construct a firm-level capital stock series using standard techniques. See for example the data appendix in Olley and Pakes (1996). In addition, the data reports investment and provides detailed information on firm entry and exit, where the latter has separate indicators for the type of exit. All the monetary variables are deflated using 4 digit deflators for producer and material prices, and investment is deflated using an investment price index (for Belgium). The deflators come from the National Institute of Statistics Belgium (<http://statbel.fgov.be/en/statistics/figures/index.jsp>).

1.2 Quota Data

I refer the reader to section 3.2.3 of the paper. The raw data on quota protection by product category is publicly available at <http://trade.ec.europa.eu/sigl/>. Appendix A.3 provides a detailed description of how the raw data is used to construct the producer-level protection variable (qr_{it}) as defined in section 3.2.3. Concordance tables are used to map the product categories used in the SIGL data to the firm-level production data (where multiple NACE Rev 1.1 industries codes are listed by firm).

1.3 Product Data

Product-level data has now become a standard feature of production datasets released by various census agencies across various countries. The number of products, product location in the various segments is presented in the Appendix A.1. The firm-level product information was provided to me by FEBELTEX in the summer of 2003. Currently, Fedustria the Federation of textile, wood and furniture industries absorbed FEBELTEX and the direct access to the micro data has been made confidential. Recent work by Goldberg et al (2010) for example, rely on plant-level production data with detailed product-level data.

2 Code

I provide the main resources for estimating the model described in section 4 of the paper. Where possible I refer to publicly available code from statistical software or other published work, which can be used in the estimation routine.

2.1 Static input: section 4.1.1

I rely on a modified estimation procedure of Levinsohn and Petrin (2003), LP, which provides publicly available code (*levpet.do*) in STATA. The Wooldridge version of LP can be implemented using *ivreg2* in STATA which runs the system GMM estimation routine as described in section 3 of 4.2. (Alternative approach). The original LP code is adjusted by including the additional demand variables capturing quota protection (qr_{it}), segment specific demand shifters (q_{st}) and the product dummies (D). All parameters are identified in one step (using a system GMM) in the estimation routine using standard GMM techniques (see section 4.1.1.). I would like to thank Amil Petrin for providing me with his code to estimate the LP-Wooldridge version of LP. A stylized version of STATA code is given below:

```
variables in logs:
revenue (r_p), inputs (l, m and k) and demand variables (qr_it,q_st,D)
plant id= Id
time = year

sort Id year
ivreg2 r_p $exoreg$ ($endoreg$=$instr$), gmm cluster(Id)
```

where the set of *exoreg* contains the capital stock, and lagged values of the variable inputs of production in addition to the lagged demand variables; the endogenous regressors *endogreg* are (current) variable inputs into production and the demand variables ($l_{it}, m_{it}, q_{st}, qr_{it}$). The set of instruments *instr* are the lagged variables (potentially including double lagged values) of the endogenous inputs.

2.2 Dynamic input: section 4.1.2

I adjust the original Olley and Pakes (1996) estimation routine and add two additional state variables (qr_{it}, D) and rely on either standard non linear least squares techniques or GMM to identify the parameters of interest. The non linear least squares procedure can be executed in different ways, for example a straightforward implementation of *nl* in STATA.

2.3 GMM approach

The various models presented in section 4 can be estimated using standard statistical software packages such as STATA, MATLAB, or others. In particular

in STATA's MATA language the use of *optimize* allows to estimate the full GMM model discussed under section 4.2.

A sketch of STATA code is provided in the file *sketch.txt*. The piece of code is offered as one of many alternatives to estimate the various models discussed in section 4. The specific details are omitted as this will vary with the actual structure of the panel dataset (on firms, products and time).

References

- [1] Goldberg, P.K., Khandelwal, A.K, Pavcnik, N. and Topalova, P. 2010, Imported Intermediate Inputs and Domestic Product Growth: Evidence from India, *Quarterly Journal of Economics*, 125 (4), 1727-1767.
- [2] Levinsohn, J. and Petrin, A. 2003. Estimating Production Functions Using Inputs to Control for Unobservables., *Review of Economic Studies*, Vol. 70, pp. 317-342.
- [3] Olley, S. and Pakes, A. 1996. The Dynamics of Productivity in the Telecommunications Equipment Industry, *Econometrica*, Vol 64 (6), 1263-98.