

## Data Appendix

“Understanding European Real Exchange Rates,”

*American Economic Review*, 2005

by Mario J. Crucini, Christopher I. Telmer and Marios Zachariadis

This appendix provides further description of our data sources, data manipulations and statistical methods. It also tabulates the supplemental data used in our paper.

### Data

*National retail price data.* The retail price data were compiled and published by Eurostat, the statistical agency of the European Community, in cooperation with the national statistical agencies of the countries that participated. While all of the price data we utilize is published we are unaware of available electronic copies. All prices refer to cash prices paid by final consumers, including taxes, both VAT and any others paid by the purchaser. Sales points are selected in such a way that the sample selected is representative of the distribution in the capital city. Prices are collected at different locations so that the average price is representative of the distribution within the city. These data were not available electronically so we had a private firm key-punch the commodity codes, commodity descriptions and prices into an electronic format.

*Data Reconciliation.* In order to explain the price dispersion across goods that exists in our data we constructed measures of tradeability and the costs of non-traded inputs into production. The constructed variables for tradeability, and non-traded inputs from input-output tables are available at different levels of detail. For this reason, and in order to make the most of the information available for each of these factors, we matched the retail price data with each of the variables using two-digit, three-digit, and four-digit classifications depending on the level of detail available for each of the variables rather than attempting to match all variables using the same level of detail. The input-output data are also available at a three-digit level of detail that extends to four-digits for some industry groups. In order to reconcile the data as accurately as possible, we used the ISIC codes and descriptions available in the User Guide of the OECD International Sectoral Database. A description of each of our variables is as follows.

- *Tradeability.* We obtained data on imports, exports, and gross output for the period 1973 to 1990 from the 1994 edition of the OECD STAN Database. We use thirty-two non-overlapping subdivisions of manufacturing for which sufficient data are available and to the extent that they are relevant to the commodities in our price dataset. We also constructed additional tradeability indices for agriculture (sector 1) and electricity, gas, and water (sector 4) using the 1994 edition of the OECD Sectoral Database, which provides

value-added instead of gross output data. Unfortunately, we only have this data for six countries: Belgium, Denmark, France, Germany, Italy, and the United Kingdom. We assigned a zero trade share to the following industries: restaurants and hotels, transport, storage and communication, inland transport, maritime and air transport, communication, financing, insurance, real estate and business services, community, social and personal services. Otherwise, our measure of tradeability is calculated as:

$$\theta_k = \frac{\sum_{j=1}^{m_k} (X_{kj} + M_{kj})}{\sum_{j=1}^{m_k} Y_{kj}} ,$$

where for each sector  $k$  we sum over all countries  $j$  which have data for that sector.  $X_{kj}$  ( $M_{kj}$ ) stands for exports (imports) of sector  $k$  from country  $j$  and  $Y_{kj}$  stands for the gross output of sector  $k$  by country  $j$ .

- *Input–Output Data.* We use the input-output matrix for the United Kingdom in 1988. These data were compiled by Keith Maskus and Allan Webster (1995). We thank Tom Prusa for suggesting this data source, which is available at the National Bureau of Economic Research home page. Non-traded inputs are assumed to include: utilities, construction, distribution, hotels, catering, railways, road transport, sea transport, air transport, transport services, telecommunications, banking, finance, insurance, business services, education, health and other services. We compute the cost share of non-traded intermediate inputs computed as,

$$\Phi_k = \sum_{s=1}^S \phi_{ks} ,$$

where  $\phi_{ks}$  is the share of non-traded intermediate input  $s$  in the total cost of the output of sector  $k$ .

*Nominal exchange rates.* We obtained daily data on nominal exchange rates from the New York Federal Reserve Bank’s web page. The Eurostat survey provides a 2-3 month window during which the survey was conducted for each good (for each of our 4 cross-sections). Accordingly, the nominal exchange rate we use is an average of the daily values over the respective months for each good. Data and programs are available at <http://bertha.tepper.cmu.edu/eurostat>.

*VAT data.* We obtained country-specific data on VAT rates for 23 different categories of goods and services for each of the years 1975, 1980, 1985 and 1990. We then categorized each of our goods into one of these categories and then divided each individual price by its associated country/year-specific VAT rate to arrive at before-VAT prices. All of our analysis in Tables 4-6 of our paper, and

Figure 3, use these before-VAT prices. The VAT rates across EU countries are taken primarily from the European Commission publication "VAT rates applied in the member states of the European Community" (2002), the OECD publication "Taxing Consumption", and the Ernst and Young publication "Vat and Sales Taxes Worldwide: A Guide to Practice and Procedures in 61 Countries" (1996). Secondary sources on VAT rates include the Tax Executives Institute publication "Value-Added Taxes: A Comparative Analysis" (1992), Alan Tait's "Value Added Tax: International Practice and Problems" (1988), and George Carlin's "Value-Added Tax: European Experience and Lessons for the United States" (1980). The data are available at <http://bertha.tepper.cmu.edu/eurostat>.

*Real GDP Per-Capita* was obtained from Penn World Tables 6.1 for each of the years corresponding to our cross-sections.

A statistical appendix and more extensive data appendix are available from the authors upon request.

## Regression Method

Given our measures of price dispersion, with  $N$  observations on  $y$ , our goal is to characterize the variation in  $y$  in terms of some vector of explanatory variables,  $x$ , on which we have  $N$  observations. Assuming that the regression of  $y$  on  $x$  is linear, we have,

$$y_i = \alpha + x_i \cdot \beta + u_i \quad , \quad (1)$$

where  $u_i$  is *i.i.d.*. The main problem we have — and one that any study of highly disaggregate data is likely to have — is that our observations on  $x$  are aggregated to a larger extent than those on  $y$ . Take for example, our measure of international tradeability. While we have data on price dispersion of, say, many different types of electronic goods, our measure of tradeability is limited to one aggregative value for electronic goods in general. In a nutshell, the variable we seek to characterize — good-specific price dispersion — is observable at a much 'finer' level than the variables we seek to characterize it with.

This type of aggregation has important consequences for statistical inference. In general, it will generate a heteroskedastic pattern in the variance of regression error terms (especially in finite samples) and, just as importantly, make goodness-of-fit measures difficult to interpret. In Appendix B of Crucini, Telmer and Zachariadis (2000) we formulate a statistical framework that allows us to obtain consistent, efficient estimates of  $\beta$ , it's standard errors, and meaningful goodness-of-fit measures. We briefly review some of the statistical issues in what follows. Readers familiar with our earlier paper may wish to proceed directly to the next sub-section.

We define our data as being partitioned into  $G$  distinct 'groups,'  $g \in \{1, 2, \dots, G\}$ , containing  $N_g$  sample observations. Examples of groups are textiles, automobiles

and personal care products. We modify the population regression, (1) as follows,

$$y_{ig} = \alpha + x_{ig} \cdot \beta + u_{ig} \quad , \quad (2)$$

where the subscript  $i, g$  denotes the  $i$ th observation on a good from group  $g$ . Define the within-group sample mean for  $x$  as  $\bar{x}_g$ ,

$$\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} x_{ig} \quad . \quad (3)$$

The regression equation, (2), can be written,

$$y_{ig} = \alpha + \bar{x}_g \cdot \beta + (x_{ig} - \bar{x}_g) \cdot \beta + u_{ig} \quad , \quad (4)$$

which is a statistical regression of  $y$  onto  $\bar{x}_g$ . Sampling variation in  $\bar{x}$ , however, generates heteroskedasticity in the error term,  $(x_{ig} - \bar{x}_g) \cdot \beta + u_{ig}$ , a feature which is particularly important in our dataset, where there is a great deal of variation in the within-group sample size. We take two approaches in estimating the parameters of (4), each of which turn out to yield qualitatively similar results. First, we estimate the regression (4) using generalized least squares (GLS), having characterized the exact form of the heteroskedastic covariance matrix for the errors.

Second, we average our disaggregate data,  $y$ , within groups,  $g$ , dictated by our aggregate data,  $x$ . Specifically, we average across equation (4) and estimate the following,

$$\frac{1}{N_g} \sum_{i=1}^{N_g} y_{ig} = \alpha + \bar{x}_g \cdot \beta + \frac{1}{N_g} \sum_{i=1}^{N_g} u_{ig} \quad (5)$$

$$\Rightarrow \bar{y}_g = \alpha + \bar{x}_g \cdot \beta + \bar{u}_g \quad , \quad (6)$$

where the sample averages,  $\bar{y}_g$  and  $\bar{u}_g$  are defined in the obvious way and are understood to depend explicitly on the values  $N_g$ . Estimates based on equation (6) are also obtained using feasible GLS, given our knowledge of the covariance matrix of  $\bar{u}$ , which is a simple function of the values  $N_g$ . The main disadvantage associated with equation (6) is that it averages away potentially informative variation in  $y$ . The advantages are a simpler form of the covariance matrix and a more easily interpretable goodness-of-fit, due to the fact that we are not trying to explain variation in  $y$  with variation in  $x$ , after having removed a great deal of the latter (something inherent in equation (4)).

It is worth noting that GLS estimates of  $\beta$  based on equation (6) will be numerically identical to OLS estimates based on equation (4). Standard errors, however, may be quite different owing in large part to a much smaller number of observations associated with (6).

## More Details

We think of the overall commodity space as consisting of  $G$  distinct groups of goods, elements of each group having some economically meaningful attributes in common. Our categorization based on ISIC codes, for instance, contains groups such as textiles, automobiles, personal services, and so on. We denote  $y_{ig}$  as the measure of price dispersion (defined in the text) for some good,  $i$ , in group  $g \in \{1, 2, \dots, G\}$ . Similarly, we denote  $x_{ig}$  and  $z_{ig}$  as vectors of attributes associated with the  $i$ th good of group  $g$ . The distinction between  $x$  and  $z$  will involve aggregation within a group,  $g$ , for the former.

We assume that the joint distribution of  $y$ ,  $x$  and  $z$  is such that the regression of  $y$  onto  $x$  and  $z$  is linear, so that we can write,

$$y_{ig} = \alpha + x_{ig} \cdot \beta + z_{ig} \cdot \gamma + u_{ig} \quad (7)$$

where  $u_{ig}$  is *i.i.d.* with mean zero and variance  $\sigma^2$ , for all  $i$  and  $g$ . What distinguishes groups of goods is the conditional distribution. For  $x$  and  $z$  we assume variation across groups in the conditional mean but not the conditional variance. Denoting the conditional means,  $\mu_g = E(x_{ig} | g)$  and  $\delta_g = E(z_{ig} | g)$ , we assume that

$$x_{ig} \sim F(\mu_g, \Sigma) \quad , \quad z_{ig} \sim F(\delta_g, \Gamma) \quad ,$$

for some distribution function,  $F$ .

Turning to the basic issue —  $x$  being aggregated — suppose that every element of  $x$  were only observable up to the within-group mean. Further, suppose that this is not the case for  $z$ , where we do observe individual observations. Should we average both  $y$  and  $z$  in order to estimate  $\beta$  and  $\gamma$  and, just as importantly, their standard errors? If the answer is yes, this might be problematic, depending on the specifics of what  $z$  is. For example, suppose that  $z_{ig}$  is the cost of the  $i, g$ th good and that the costs uniformly distributed throughout each group,  $g$ . Then, by averaging within a group, we lose information on how within-group costs are related to price dispersion (another way to say this is that  $\gamma$  won't be identified under these conditions). We now turn to a discussion of the merits of each approach. We first discuss the merits of using the data we have, as-is, and then go on to discuss the advantages of averaging away the intra-group variation in both  $y$  and  $z$ .

## Estimation Based on Raw Data

When intra-group variation in  $x$  is averaged away — a data restriction, not a choice — but the intra-group variation in  $y$  and  $z$  remains, the variation ends up in the

error term. To see this, note that the population regression (7) can be written as

$$y_{ig} = \alpha + \mu_g \cdot \beta + z_{ig} \cdot \gamma + (x_{ig} - \mu_g) \cdot \beta + u_{ig} \ . \quad (8)$$

The (population) error term from the regression of  $y$  onto  $\mu_g$  and  $z$  is therefore  $(x_{ig} - \mu_g) \cdot \beta + u_{ig}$ . This object is cross-sectionally uncorrelated (*i.e.*,  $\text{cov}(x_{ig} - \mu_g, x_{jg} - \mu_g) = 0$  for all  $i$  and  $g$ ) and, so long as  $x$  and  $z$  are orthogonal, is uncorrelated with the regression function,  $\alpha + \mu_g \cdot \beta + z_{ig} \cdot \gamma$ . Given that this is the case — the distributional assumptions above imply it — we can write the following variance decomposition:

$$\text{var}(y_{ig}) = \text{var}(\mu_g \cdot \beta + z_{ig} \cdot \gamma) + \text{var}((x_{ig} - \mu_g) \cdot \beta + u_{ig}).$$

The error term is homoskedastic as long as the conditional covariance matrix does not depend on  $g$ . That is,

$$\text{var}((x_{ig} - \mu_g) \cdot \beta + u_{ig}) = \beta^\top \Sigma \beta + \sigma^2 \ .$$

Estimates of  $\beta$  and  $\gamma$  based on (8) will therefore be consistent and efficient, conditional on the restriction that we don't get observations on  $x_{ig}$ . The fit of the regression, on the other hand, will understate the fit of the unrestricted regression, equation (7), and must be interpreted accordingly.

Finite sample considerations change matters in an important way. Suppose that we have  $N_g$  observations on  $y$ ,  $x$ , and  $z$ , from each group  $g$ . The sample analog of equation (7) is

$$y_{ig} = \alpha + \bar{x}_g \cdot \beta + z_{ig} \cdot \gamma + (x_{ig} - \bar{x}_g) \cdot \beta + u_{ig} \ , \quad (9)$$

where,

$$\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} x_{ig} \ .$$

Equation (9) is still a regression (*i.e.*,  $\text{cov}(\bar{x}_g, x_{ig} - \bar{x}_g) = 0$  for all  $i$  and  $g$ ), but, because of the sampling variance in  $\bar{x}_g$ , the covariance matrix of the residuals will have a particular, heteroskedastic structure. Given  $N = \sum_{g=1}^G N_g$  total observations, the covariance matrix is block-diagonal, with each block defined in terms of observations from a given group,  $g$ . Each block has off-diagonal terms equal to

$$\text{cov}((x_{ig} - \bar{x}_g) \cdot \beta + u_{ig}, (x_{jg} - \bar{x}_g) \cdot \beta + u_{jg}) = -\frac{1}{N_g} \beta^\top \Sigma \beta^\top \ ,$$

for a given  $g$  and  $i \neq j$ , and diagonal terms equal to

$$\text{var}((x_{ig} - \bar{x}_g) \cdot \beta + u_{ig}) = \frac{N_g - 1}{N_g} \beta^\top \Sigma \beta + \sigma^2 \ .$$

What's likely to be most important, therefore, are the off-diagonal terms in each block, where variation in  $N_g$  will have a much larger effect.

Finally, should we choose to correct the regression based on equation (9) for heteroskedasticity, we need estimates of  $\Sigma$  and  $\sigma$ . The latter can be obtained via the GLS regressions outlined below. The former is more problematic. In general, we cannot estimate  $\Sigma$  without observing individual observations of  $x_{ig}$ . That is, since,

$$\begin{aligned} \text{var}(x) &= E[\text{var}(x | g)] + \text{var}(E[x | g]) \\ &= \Sigma + \text{var}(\mu_g) , \end{aligned}$$

we can estimate  $\text{var}(\mu_g)$  but we can't estimate  $\Sigma$  without some information on  $x$  itself. So, efficient estimation based on equation (9), in which we do not average away any variation in either  $x$  or  $z$ , is not possible without further assumptions regarding the conditional covariance matrix,  $\Sigma$ . In what follows we experiment with a number of arbitrary, but sensible, values for  $\Sigma$  and examine the implications.

#### Estimation Based on Averaged Data

We now consider the merits of estimating  $\beta$  and  $\gamma$  by averaging away the within-group variation in both  $y$  and  $z$ . In this case, the distinction between  $x$  and  $z$  is not relevant so, for notational simplicity, we subsume  $z$  into  $x$ . Averaging equation (7) within groups, we have

$$\frac{1}{N_g} \sum_{i=1}^{N_g} y_{ig} = \alpha + \frac{1}{N_g} \sum_{i=1}^{N_g} x_{ig} \cdot \beta + \frac{1}{N_g} \sum_{i=1}^{N_g} u_{ig} \quad (10)$$

$$\Rightarrow \bar{x}_g = \alpha + \bar{x}_g \cdot \beta + \bar{u}_g , \quad (11)$$

where the sample averages,  $\bar{y}_g$ ,  $\bar{x}_g$  and  $\bar{u}_g$  are defined in the obvious way and are understood to depend explicitly on the values  $N_g$ .

Residuals based on equation (11) will also be heteroskedastic, but in a simpler way than those based on equation (9). The covariance matrix is diagonal with the  $g$ th diagonal element equal to  $\sigma^2/N_g$ . A consistent, efficient estimator of  $\beta$  is therefore the GLS estimator  $\tilde{\beta} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} Y$ , where  $Y$  and  $X$  denote observations on  $\bar{y}_g$  and  $\bar{x}_g$ , respectively, and  $\Omega$  is the covariance matrix of the (averaged) error terms. Note that, as is straightforward to show, if we omit  $z$  from equations (9) and (11), OLS estimates based on equation (9) are numerically identical to the GLS estimates,  $\tilde{\beta}$ . What turns out to differ in an important way are the standard errors and the measures of fit of the respective regressions.

The main advantage to estimation based on (11) is that we don't need an estimate of  $\Sigma$ , the conditional variance of the averaged regressors, in order to obtain efficient estimates. Goodness of fit measures based on (11) are also easier to interpret, since we aren't trying to explain unaveraged variation in  $x$  using averaged variation in

$y$ . The disadvantages are mainly related to averaging things we don't have to, in particular the variables in  $z$ .

Finally, there are a number of well known issues associated with computing goodness-of-fit measures based on a regression where estimates are obtained by GLS. The basic issue is whether one uses residuals based on (11) or residuals based on the standard 'transformed' GLS sample regression,

$$R\bar{y} = R(\alpha\iota + \bar{x}\beta) + R\bar{u} \text{ ,}$$

where  $R^\top R = \Omega^{-1}$ ,  $\bar{y}$  and  $\bar{x}$  are vectors of sample observations (where, again,  $z$  is subsumed into  $x$ ), and  $\iota$  is a vector of ones.

Our approach is simple. The quantity we are ultimately interested in is

$$\frac{\text{var}(x_{ig} \cdot \beta)}{\text{var}(y_{ig})} \text{ .}$$

A consistent estimator is of this is,

$$R^2 = \frac{\beta' \hat{\text{var}}(\bar{x}_g \sqrt{N_g}) \beta}{\hat{\text{var}}(\bar{y}_g \sqrt{N_g})} \text{ ,}$$

where  $\hat{\text{var}}(\cdot)$  denotes sample variance. While this quantity is not guaranteed to lie between zero and unity, our experience is that it gives sensible answers which incorporate the large amount of variation in  $N_g$  exhibited by our dataset.

## References

Maskus, Keith, and Allan Webster, 1995, Factor specialization in U.S. and U.K. trade: Simple departures from the factor-content theory, *Swiss Journal of Economics and Statistics* 1 (131).

**Table A1**  
**Non-Traded Input Cost Shares**

Category	Input share	Category	Input Share
Agriculture	0.144	Glass	0.195
Forestry and fishing	0.318	Clay refractories	0.201
Milk and milk products	0.080	Cement, concrete	0.259
Meat, fruit, veg, fish processing	0.096	Iron and steel; plant	0.111
Oils and fats, grain products	0.124	Non-ferrous metals	0.092
Bread, biscuits etc	0.104	Metal castings etc	0.119
Sugar, confectionery	0.139	Office machinery, computers	0.103
Foods, nes	0.174	Other machinery	0.116
Alcoholic drink	0.097	Telecomms equipment	0.120
Soft drink	0.160	Domestic electric appliances	0.135
Tobacco	0.046	Electronic consumer goods	0.113
Woven textiles	0.095	Electronic components	0.147
Hosiery, other knitted goods	0.097	Electric lighting equipment	0.106
Carpets etc	0.112	Shipbuilding and repairing	0.134
Clothing and furs	0.096	Motor Vehicles and parts	0.090
Leather and leather goods	0.073	Other vehicles	0.110
Footwear	0.086	Instrument engineering	0.149
Wood furniture	0.107	Other manufacturing	0.122
Paper and board products	0.135	Utilities	0.157
Printing and publishing	0.211	Hotels, catering etc	0.144
Synthetic resins, man-made fibers	0.134	Railways	0.272
Paints, dyes etc	0.154	Road transport etc	0.151
Soap and toiletries	0.226	Air transport	0.253
Chemicals nes	0.133	Transport services	0.204
Mineral oil processing	0.054	Telecomms and postal	0.124
Rubber and plastic products	0.127	Business services etc	0.134
		Other services	0.304

Values are based on the author's calculations, deriving from the 1988 input-output matrix for the U.K., compiled by Maskus and Webster (1995) and available at <http://www.nber.org>. Further details are available in the data appendix.

**Table A2**  
**Trade Shares**

Industry	Trade Share			
	1975	1980	1985	1990
Agriculture, hunting, forestry and fishing	0.26	0.41	0.38	0.56
Food	0.27	0.26	0.29	0.31
Beverages	0.20	0.25	0.28	0.33
Tobacco	0.06	0.17	0.21	0.21
Textiles	0.44	0.53	0.64	0.70
Wearing apparel	0.48	0.47	0.58	0.71
Leather products	0.47	0.63	0.77	0.82
Footwear	0.42	0.58	0.73	0.77
Furniture and fixtures	0.17	0.23	0.29	0.32
Paper and paper products	0.40	0.45	0.54	0.61
Printing and publishing	0.15	0.15	0.16	0.15
Industrial chemicals	0.58	0.75	0.90	0.98
Other chemicals	0.29	0.44	0.54	0.57
Chemical products, n.e.c.	0.25	0.46	0.59	0.65
Misc. products of petroleum and coal	0.62	0.46	0.42	0.36
Rubber products	0.40	0.50	0.58	0.62
Plastic products, n.e.c.	0.21	0.27	0.30	0.28
Pottery, china etc.	0.23	0.28	0.26	0.25
Non-metal products, n.e.c.	0.19	0.19	0.21	0.22
Iron and steel	0.37	0.43	0.48	0.52
Non-ferrous metals	0.69	0.66	0.71	0.73
Fabricated metal products, except machinery and equipment	0.28	0.34	0.39	0.40
Office and computing machinery	1.08	1.24	1.35	1.40
Machinery and equipment, n.e.c.	0.56	0.56	0.60	0.62
Electrical machinery	0.38	0.44	0.52	0.61
Radio, television and communication equipment	0.50	0.49	0.62	0.79
Electrical apparatus, n.e.c.	0.39	0.47	0.49	0.54
Shipbuilding and repairing	0.53	0.40	0.42	0.44
Motor vehicles	0.51	0.58	0.66	0.72
Motorcycles and bicycles	0.55	0.50	0.62	1.00
Professional goods	1.00	1.13	1.52	1.49
Other manufacturing n.e.c.	0.94	1.20	1.39	1.47
Electricity, gas and steam and water	0.63	1.16	0.84	0.68

Source: OECD Sectoral Database and OECD STAN Database. The following industries (not shown) have been assigned a zero trade share: restaurants and hotels, transport, storage and communication, inland transport, maritime and air transport, communication, financing, insurance, real estate and business services, community, social and personal services.