CPB Netherlands Bureau for Economic
Policy Analysis

CPB Discussion Paper | 216

The identification of reporting accuracies from mirror data

Arie ten Cate

The identification of reporting accuracies from mirror data

Arie ten Cate[*]

July 17, 2012

Abstract

Mirror data are observations of bilateral variables such as trade from one country to another, reported by both countries. The efficient estimation of a bilateral variable from its mirror data, for example when compiling consistent international trade statistics, requires information about the accuracy of the reporters.

This paper discusses the simultaneous estimation of the accuracy of multiple reporters, from all mirror data. This requires a model with an identification restriction. Two models are presented, each with the same simple kind of identifying restriction. The inadequate treatment of this restriction in the literature might be an explanation for the limited presence of integrated international statistics.

Key words: international statistics; international trade; international migration; harmonization; generalized linear model; non-nested models

[*] CPB Netherlands Bureau for Economic Policy Analysis. Email: `a.ten.cate@cpb.nl`. The author thanks Tom Wansbeek (Groningen University) for his comments and suggestions.

Contents

1 Introduction

Mirror data are bilateral data where each quantity is reported twice. For instance with international trade data we may have for a particular trade flow the value reported by the exporter and the value reported by the importer. Other examples are international migration, direct foreign investment, and foreign debt. The UN does not produce consistent statistics of the trade among its member countries, nor do the OECD or the EU.

Efficient harmonization of mirror data requires the simultaneous estimation of the accuracies of all reporters. Such models require an identifying restriction; otherwise the model cannot distinguish between for instance all countries reporting exactly correct and all countries reporting 10% too much.

The following principles are proposed concerning the identifying restriction:

- If all discrepancies are zero, then the estimated reporting error parameters are zero.
- All countries are treated symmetrically, and also exporting and importing are treated symmetrically.
- One restriction is enough.

These principles are satisfied in the two reporting models presented below, in sections 3 and 4 respectively.

In the literature this identification problem has been dealt with in a variety of ways. None of the following papers satisfies all our principles.

Tsigas et al. (1992) try the impossible with a model with country-specific reporting biases: identify from the data a country with a zero export reporting bias and another country with a zero import reporting bias; see Ten Cate (2007), appendix E for a discussion. In Gaulier and Zignago (2010), equation (5), an identifying restriction for export reporting is assumed and also one for import reporting (without a constant term), instead of one restriction for both.

In Poulain and Dal (2008) and in De Beer et al. (2010), the immigration reported by Sweden is assumed to have no bias. The table 2.2 in Van Leeuwen and Lejour (2006) is based on a similar assumption about Belgium and Luxembourg combined.

The procedure of the GTAP organization does not use a simultaneous model: a country-specific accuracy is derived from the discrepancies of that country. In a second round this is recalculated without the discrepancies with its trading partner which was the least accurate in the first round. See Gehlhar (1996).

2 The setting

Without loss of generality, we will use the wording of international trade. For all pairs i, j with $i \neq j$, let Y_{ij} be the unknown true value of the trade flow from country i to country j. Let Y_{ij}^{exp} and Y_{ij}^{imp} be the Y_{ij} as reported by the exporter i and by the importer j, respectively. All Y_{ij}^{exp} and Y_{ij}^{imp} are stochastic, and distributed independently. Lowercase y indicates logarithms. Relative reporting discrepancies are defined as differences between logs:

$$\Delta y_{ij} \equiv y_{ij}^{\text{exp}} - y_{ij}^{\text{imp}} \tag{1}$$

3 Country-specific means

3.1 The model

Following the pioneering Tsigas et al. (1992), we assume country-specific systematic deviations from the true value, as follows. For all pairs i, j with $i \neq j$:

$$\mathrm{E}\left[y_{ij}^{\text{exp}}\right] = y_{ij} + \mu_i^{\text{exp}} \quad \text{and} \quad \mathrm{E}\left[y_{ij}^{\text{imp}}\right] = y_{ij} + \mu_j^{\text{imp}} \tag{2}$$

Hence, with definition (1) we have:

$$\mathrm{E}\left[\Delta y_{ij}\right] = \mu_i^{\text{exp}} - \mu_j^{\text{imp}} \tag{3}$$

Also, let all y_{ij}^{exp} and y_{ij}^{imp} have the same variance, say σ^2. Then this is an ANOVA-like homoscedastic linear regression model. In the case of international trade in goods, the right-hand side of (3) absorbs the cif/fob margin.

Of course, this model is not identified without a restriction on the parameters: adding the same number to all parameters has no effect on the right-hand side of (3).

3.2 The identifying restriction

In order to solve the identification problem using the principles stated above, we assume that the expected value of the average of the reported total export and the reported total import is equal to the true total trade:

$$\mathrm{E}\left[\frac{1}{2}\left(\sum_i\sum_j Y_{ij}^{\mathrm{exp}}+\sum_i\sum_j Y_{ij}^{\mathrm{imp}}\right)\right]=\sum_i\sum_j Y_{ij} \tag{4}$$

or

$$\mathrm{E}\left[\sum_i\sum_j\left(Y_{ij}^{\mathrm{exp}}-Y_{ij}\right)+\sum_i\sum_j\left(Y_{ij}^{\mathrm{imp}}-Y_{ij}\right)\right]=0 \tag{5}$$

Next, we use this approximation:

$$\mathrm{E}\left[\frac{Y_{ij}^{\mathrm{exp}}}{Y_{ij}}-1\right]\approx\mathrm{E}\left[\log\left(\frac{Y_{ij}^{\mathrm{exp}}}{Y_{ij}}\right)\right]=\mathrm{E}\left[y_{ij}^{\mathrm{exp}}-y_{ij}\right]=\mu_i^{\mathrm{exp}} \tag{6}$$

and similarly for the importers. Then equation (5) can be approximated as:

$$\sum_i\left(\mu_i^{\mathrm{exp}}\sum_j Y_{ij}\right)+\sum_j\left(\mu_j^{\mathrm{imp}}\sum_i Y_{ij}\right)=0 \tag{7}$$

Thus, the restriction equates the sum of the relative biases, weighted with the country's trade size, to zero. This satisfies the first of the principles above: if all discrepancies are zero then all estimated μ_i^{exp} and μ_j^{imp} are zero, with all regression residuals equal to zero. The other principles are clearly satisfied as well.

The weights must be approximated from the reported values; for example:

$$\sum_j Y_{ij}\approx\frac{1}{2}\sum_j\left(Y_{ij}^{\mathrm{exp}}+Y_{ij}^{\mathrm{imp}}\right) \tag{8}$$

3.3 Estimation

Estimation with an identifying restriction can be done easily as follows. First, fix arbitrarily one parameter and estimate the model. Then add a vector to the parameter vector such that the right-hand side of (3) does not change:

$$\widehat{\boldsymbol{\mu}}=\widehat{\boldsymbol{\mu}}_{\mathrm{F}}+c\mathbf{q} \tag{9}$$

where c is some number and $\mathbf{q}$ is a vector of ones, since adding the same number to all parameters does not change the right-hand side of (3). The restriction, say $\mathbf{g}'\widehat{\boldsymbol{\mu}}=0$, is met if and only if

$$c=-\frac{\mathbf{g}'\widehat{\boldsymbol{\mu}}_{\mathrm{F}}}{\mathbf{g}'\mathbf{q}} \tag{10}$$

The variance of $\widehat{\boldsymbol{\mu}}$ remains to be estimated, as follows. Substitution of (10) into (9) gives:

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_{\mathrm{F}} - \mathbf{q}\frac{\mathbf{g}'\widehat{\boldsymbol{\mu}}_{\mathrm{F}}}{\mathbf{g}'\mathbf{q}} = \left(\mathbf{I} - \frac{1}{\mathbf{g}'\mathbf{q}}\mathbf{q}\mathbf{g}'\right)\widehat{\boldsymbol{\mu}}_{\mathrm{F}} \equiv \mathbf{R}\widehat{\boldsymbol{\mu}}_{\mathrm{F}} \tag{11}$$

The estimated variance matrix of $\widehat{\boldsymbol{\mu}}$ is $\mathbf{R}\boldsymbol{\Omega}_{\mathbf{F}}\mathbf{R}'$ where $\boldsymbol{\Omega}_{\mathbf{F}}$ is the old estimated variance matrix with zeroes for the fixed parameter.

This approach can also be used with the more complicated models in section 4 and 5 below. For the present model there is a simple closed form of the restricted estimate. Writing the model as $\mathrm{E}[\Delta\mathbf{y}] = \mathbf{X}\boldsymbol{\mu}$ we have $\widehat{\boldsymbol{\mu}} = \mathbf{Z}\mathbf{X}'\Delta\mathbf{y}$ where $\mathbf{Z} \equiv (\mathbf{X}'\mathbf{X} + \mathbf{g}\mathbf{g}')^{-1}$. The variance matrix is $\sigma^2(\mathbf{Z} - \mathbf{Z}\mathbf{g}\mathbf{g}'\mathbf{Z})$. See for instance theorem 7 of chapter 13 in Magnus and Neudecker (1988) and later editions.

4 Country-specific variances

4.1 The model

The traditional method of optimally combining inconsistent reports uses the reciprocal error variances as weights; see Stone et al. (1942). In this section we present a model with country-specific error variances. Assume for all pairs i, j with $i \neq j$:

$$\mathrm{Var}\left[y_{ij}^{\mathrm{exp}}\right] = V_i^{\mathrm{exp}} \quad \text{and} \quad \mathrm{Var}\left[y_{ij}^{\mathrm{imp}}\right] = V_j^{\mathrm{imp}} \tag{12}$$

and

$$\mathrm{E}\left[\Delta y_{ij}\right] = \mu \tag{13}$$

Then we have:

$$\mathrm{E}\left[(\Delta y_{ij} - \mu)^2\right] = V_i^{\mathrm{exp}} + V_j^{\mathrm{imp}} \tag{14}$$

With given μ this is a regression model like (3) above, though with heteroscedastic disturbances:

$$\mathrm{Var}\left[(\Delta y_{ij} - \mu)^2\right] = (\gamma_2 + 2)\left(V_i^{\mathrm{exp}} + V_j^{\mathrm{imp}}\right)^2 \tag{15}$$

where γ_2 is the net kurtosis of the distribution of the Δy_{ij}. For a derivation, see appendix A. Note that equation (13), considered as a regression model, is also heteroscedastic, as shown in (14).

Similar to section 3, the parameters V_i^{exp} and V_j^{imp} are not estimable without an identifying restriction: adding the same number to all V_i^{exp} and subtracting this number from all V_j^{imp} has no effect on the right-hand side of (14).

In a way, the sections 3 and 4 are themselves mirror images of each other. This can be expressed by assuming that the discrepancies are normally distributed. Then the models can be written as $\Delta y_{ij} \sim \mathrm{N}(\boldsymbol{\beta}'\mathbf{x}_{ij}, \sigma^2)$ and $\Delta y_{ij} \sim \mathrm{N}(\mu, \boldsymbol{\beta}'\mathbf{x}_{ij})$, respectively. The first model is a standard regression model; the second model is not.

4.2 Ordering accuracies without the identification restriction

The estimation of the sums $V_i^{\text{exp}} + V_j^{\text{imp}}$ does not depend on the identifying restriction. Moreover all linear combinations of the $V_i^{\text{exp}}+V_j^{\text{imp}}$ are estimable without the restriction, such as

$$\left(V_i^{\text{exp}} + V_k^{\text{imp}}\right) - \left(V_j^{\text{exp}} + V_k^{\text{imp}}\right) = V_i^{\text{exp}} - V_j^{\text{exp}} \tag{16}$$

and similarly $V_i^{\text{imp}} - V_j^{\text{imp}}$. The differences (16) allow us to order the countries by export reporting accuracy without using the restriction

Such ordering is not possible with the means model in section 3: the sign of a difference between two numbers of unknown sign tells nothing about the difference between the two magnitudes.

The GTAP procedure, described by Gehlhar (1996) and mentioned in our Introduction above, ranks countries by reporting accuracy without using an identifying restriction. This is possible because they ignore the sign of the discrepancies, as we do in our model (14).

4.3 The identifying restriction

If there is no reason why export reporting would be systematically more, or less, accurate than import reporting, we choose the following restriction with $\rho = 1$:

$$\operatorname{Var}\left[\sum_i \sum_{j \neq i} Y_{ij}^{\text{exp}}\right] = \rho \operatorname{Var}\left[\sum_i \sum_{j \neq i} Y_{ij}^{\text{imp}}\right] \tag{17}$$

Otherwise we choose some other $\rho > 0$.

Since the reported values are independently distributed, the left-hand side of (17) can be rewritten as a sum of variances:

$$\begin{aligned}\sum_i\sum_j \mathrm{Var}\left[Y_{ij}^{\mathrm{exp}}\right] &= \sum_i\sum_j Y_{ij}^2 \mathrm{Var}\left[\frac{Y_{ij}^{\mathrm{exp}}}{Y_{ij}}\right] = \sum_i\sum_j Y_{ij}^2 \mathrm{Var}\left[\frac{Y_{ij}^{\mathrm{exp}}}{Y_{ij}} - 1\right]\\ &\approx \sum_i\sum_j Y_{ij}^2 \mathrm{Var}\left[y_{ij}^{\mathrm{exp}} - y_{ij}\right] = \sum_i\sum_j Y_{ij}^2 \mathrm{Var}\left[y_{ij}^{\mathrm{exp}}\right]\\ &= \sum_i\sum_j Y_{ij}^2 V_i^{\mathrm{exp}} = \sum_i \left(V_i^{\mathrm{exp}} \sum_j Y_{ij}^2\right) \qquad (18)\end{aligned}$$

At the $\approx$ sign, the approximation (6) was used. Rewriting similarly the right-hand side of (17) gives the following approximation of (17):

$$\sum_i \left(V_i^{\mathrm{exp}} \sum_j Y_{ij}^2\right) = \rho \sum_j \left(V_j^{\mathrm{imp}} \sum_i Y_{ij}^2\right) \qquad (19)$$

Thus, the variances are weighted with the squared trade. They must also be approximated using the reported values, as in (8).

The first principle in the Introduction above is satisfied in the sense that if all $\Delta y_{ij} - \mu$ are zero then all estimated V_i^{exp} and V_j^{imp} are zero, with all regression residuals of (14) equal to zero. The second principle is satisfied with $\rho = 1$.

4.4 Estimation

With (13), the average of the Δy_{ij} is an unbiased estimate of μ. This estimate can be improved later on by using estimated variances. Alternatively, in the case of international trade in goods the scalar μ might be set at minus some known average cif/fob margin.

Since we are interested in the variance parameters, we now focus on model (14). Given the μ, the variances in model (14) can be estimated unbiased by OLS. The identifying restriction can be applied in the same way as with the means model in section 3.3, by starting with one fixed parameter. In this case the vector $\mathbf{q}$ consists of elements equal to $+1$ or -1.

This estimate can be improved by taking into account the heteroscedasticity given in (15): the standard deviation of the dependent variable is proportional to its expectation. See Amemiya (1973). Unfortunately some of the OLS-estimated $V_i^{\mathrm{exp}} + V_j^{\mathrm{imp}}$ in (14) might be negative; see our appendix B.

4.5 Negative estimates of country variances

Even if all $V_i^{\text{exp}} + V_j^{\text{imp}}$ are positive, some of the estimated V_i^{exp} and V_j^{imp} might be negative after they are identified with the restriction. This might result in the adjustment of the parameters such that we may have for example a $V_i^{\text{exp}} = 0$. Then the weight of the exporting country in computing a harmonized value of the trade from i to j is as follows, using the reciprocal of the variances as weights:

$$\lim_{V_i^{\text{exp}} \to 0} \frac{1/V_i^{\text{exp}}}{1/V_i^{\text{exp}} + 1/V_j^{\text{imp}}} = 1 \tag{20}$$

Note that in this case the estimated V_j^{imp} cannot be zero because the estimated sum $V_i^{\text{exp}} + V_j^{\text{imp}}$ is positive.

5 Comparing models by likelihood

Any research project aimed at the publication of harmonized international statistics must start with the decision whether or not to use the sign of the discrepancies, such as with our means model in section 3 and our variances model in section 4, respectively[1].

In order to give this choice an empirical foundation, our two models might be compared empirically by their likelihood, assuming normally distributed discrepancies. Maximum likelihood was not considered above because there might be multiple local maxima in the variances model, with likelihood values near to each other but widely differing parameters; see the example in section 6.

The loglikelihood is $\sum_i \sum_{j \neq i} \ell_{ij}$ with

$$\ell_{ij} = -\frac{1}{2}\left(\frac{(\Delta y_{ij} - \mu_{ij})^2}{\sigma_{ij}^2} + \log \sigma_{ij}^2 + \log 2\pi\right) \tag{21}$$

where either the μ_{ij} and σ_{ij}^2 are conform the means model in section 3 (with all σ_{ij}^2 equal to an unknown σ^2), or conform the variances model in section 4 (with all μ_{ij} equal to a given μ). The μ of the variances model cannot be estimated with maximum likelihood; see appendix C. Given μ, this model

[1] These two possibilities are more or less evenly distributed among the papers discussed in the Introduction. Signed: Tsigas et al. (1992), Poulain and Dal (2008), De Beer et al. (2010). Unsigned: Gehlhar (1996), Gaulier and Zignago (2010). Van Leeuwen and Lejour (2006) do both in their table 2.2.

can be estimated with maximum likelihood using standard software when considering it as a GLM model; see appendix D.

This comparison requires the appraisal of a loglikelihood difference between non-nested model estimates. The absence of a null hypothesis makes it more difficult to answer the question “is this a large difference?” See Gourieroux and Monfort (1994) for a review of the literature.

Here we simply consider a loglikelihood difference of more than 2 as large, or “significant”. This value coincides with the widely used $\pm 2\sigma$ significance limits in the following situation: a statistic x is normally distributed with unknown mean μ and known variance σ^2.

To see this, note that the loglikelihood function of this μ is $-(\mu - x)^2/2\sigma^2$ + constant. Hence the loglikelihood at $\mu = x \pm 2\sigma$ is 2 lower than at $\mu = x$. See also Edwards (1972), p. 182.

6 Illustration

6.1 The data and the ML estimates

We illustrate the above models with the data in table 1, showing the reported trade in services between France, Germany, Italy and the UK, obtained from OECD statistics. The discrepancies are extremely large; it might be profitable to study individual transactions in order to find the cause of this, but these data may also serve us as an interesting example. All computations are based on the integer percentages in the last column of table 1.

Table 2 shows the results for our means model in section 3. The standard errors of the μ parameters are estimated using maximum likelihood, assuming normally distributed reporting and using the “observed information matrix”.

Table 3 shows three local likelihood maxima for our variances model in section 4. We used $\mu = 0$, since the mean of the $Y_{ij}^{\text{exp}} - Y_{ij}^{\text{imp}}$ is very small: -0.1 billion USD with a standard deviation of 2.3. Also, we used $\rho = 1$ in restriction (17). Compared with the benchmark loglikelihood difference of 2 from section 5, the differences between the loglikelihood values in the bottom line of table 3 are very small. The V are allowed to be negative here; this flatters the loglikelihood. Estimated standard errors of the $\sqrt{V}$ in the table are quite meaningless here, because of the variation between the three estimates. See appendix E for details. Further research must tell whether this occurs more generally.

Table 1: Reported trade in services, OCS (2002, billion USD)

reporting exporter i	reporting importer j	reported export Y_{ij}^{exp}	reported import Y_{ij}^{imp}	discrepancy (%) Δy_{ij}
France	Germany	1.3	4.8	−131
France	Italy	1.8	3.7	−72
France	UK	3.8	3.3	14
Germany	France	4.7	3.6	27
Germany	Italy	1.8	3.6	−69
Germany	UK	6.6	3.9	53
Italy	France	3.3	1.5	79
Italy	Germany	3.4	1.4	89
Italy	UK	3.6	1.2	110
UK	France	5.7	4.8	17
UK	Germany	7.5	9.1	−19
UK	Italy	2.9	7.0	−88
Total		46.4	47.9	

Source: OECD Statistics on International Trade in Services.
Note: The discrepancies are presented as a percentage by merely multiplying them with 100, without first transforming them with "exp(. . .) − 1".

Table 2: Estimate of the means model, in percentages

	μ^{exp}	st.err	μ^{imp}	st.err
France	−55	22	−19	21
Germany	−2	21	26	21
Italy	75	22	54	21
UK	−6	19	−52	22

Table 3: Three local likelihood maxima of the variances model

	$\sqrt{V^{\text{exp}}}$	$\sqrt{V^{\text{imp}}}$	$\sqrt{V^{\text{exp}}}$	$\sqrt{V^{\text{imp}}}$	$\sqrt{V^{\text{exp}}}$	$\sqrt{V^{\text{imp}}}$
France	29	$V < 0$	63	46	86	19
Germany	54	86	77	47	20	21
Italy	96	63	108	84	87	74
UK	50	$V < 0$	$V < 0$	$V < 0$	$V < 0$	44
loglikelihood	reference		−0.2		−0.3	

Notes. The $\sqrt{V}$ are percentages. The loglikelihood values give the difference with the first local maximum.

6.2 Empirical comparison of the two models

As discussed in section 5, the two maximum loglikelihoods are compared. The maximum loglikelihood of the means model in table 2 is 10 larger than the maximum loglikelihood of the variances model in table 3. This is a large difference, compared with the benchmark value of 2 from section 5. Also, it dwarfs the loglikelihood differences in table 3. See appendix F for details.

Although this is based on a very small number of countries, we try to explain this result briefly. Most sources of reporting error might have an effect on a mean, either positive or negative. One might think of omissions in the reporting, or incorrect identification of a partner country. (If A thinks its export go to B while it actually goes to C than B seems to under-report its imports and C seems to over-report.)

7 Conclusions

In this paper we have explored models and estimation methods for the analysis of discrepancies in mirror data. These models provide a foundation for harmonizing mirror data, such as the compilation of consistent international trade statistics. The models require an identifying restriction.

Starting with the country-specific biases of Tsigas et al. (1992), their treatment of the estimable functions has been improved and a proper identifying restriction is proposed.

A similar restriction has been applied to a new model, with country-specific reporting variances. This supplies the weights for the traditional method of combining mirror data, weighting with the reciprocal of the country-specific error variances.

A small data set on international trade is used as an illustration. The data fits by far the best to the means model, where "by far" is obtained from a somewhat unconventional criterion.

A Proof of equation (15)

The proof of equation (15) is given here. The following shorthand notations are used: Δ for the net discrepancy $\Delta y_{ij} - \mu$ and V for the variance $V_i^{\mathrm{exp}} + V_j^{\mathrm{imp}}$. Then:

$$\mathrm{E}[\Delta] = 0 \tag{22}$$

$$\mathrm{E}\left[\Delta^2\right] = \mathrm{Var}[\Delta] = V \tag{23}$$

Hence:

$$\begin{aligned}
\mathrm{Var}\left[\Delta^2\right] &= \mathrm{E}\left[\left(\Delta^2 - \mathrm{E}\left[\Delta^2\right]\right)^2\right] = \mathrm{E}\left[\left(\Delta^2 - V\right)^2\right] \\
&= \mathrm{E}\left[\Delta^4\right] + V^2 - 2V\mathrm{E}\left[\Delta^2\right] \\
&= \mu_4 + V^2 - 2V^2 = \mu_4 - V^2 = \left(\frac{\mu_4}{V^2} - 1\right) V^2 \\
&= (\gamma_2 + 2)\, V^2
\end{aligned} \tag{24}$$

where $\mu_4 \equiv \mathrm{E}\left[\Delta^4\right]$ and $\gamma_2 \equiv \mu_4/V^2 - 3$ is the excess kurtosis of the distribution of Δ.

With normally distributed discrepancies, (24) can be derived very simple as follows. The variances model can then be written as:

$$\Delta \sim \mathrm{N}\left(0, V\right) \tag{25}$$

and hence

$$\frac{\Delta^2}{V} \sim \chi_1^2 \tag{26}$$

Hence the variance of Δ^2/V is 2, which can be written as

$$\mathrm{Var}\left[\Delta^2\right] = 2V^2 \tag{27}$$

This is (24) with $\gamma_2 = 0$ due to the assumed normal distribution.

B Two remarks about Amemiya (1973)

Amemiya (1973) discusses the estimation of a model like our model (14) with (15). Equation (14) is estimated with OLS and the result is used to optimally weight (14) according to (15). Two remarks are made.

First: in many applications the dependent variable will be nonnegative. This is the case with our variances model and also with Amemiya's example, where the dependent variable is a time duration. Then it makes no sense to use the reciprocal of a squared estimated expectation as optimal weight, when this estimated expectation is *negative*: a low value of the expectation should give a large weight. This problem occurs with our example data in section 6 above. Also, when the last observation in the example in Amemiya (1973) is omitted then there are two (identical) observations with a negative estimated expected dependent variable.

This problem is not discussed in Amemiya (1973). One might solve it by subtracting the following expression from all elements of the vector of the estimated expected dependent variable, in a general notation:

$$a \min\left(0, \widehat{\mathrm{E}\,[y_1]}, \widehat{\mathrm{E}\,[y_2]}, \widehat{\mathrm{E}\,[y_3]}, \dots\right) \tag{28}$$

with $a > 1$. In our example in section 6 the result with $a = 2$ is: export $\sqrt{V}$ in percent = 77, 48, 82, 21; import = (negative V), 67, 46, 28. With $a = 4$ the result is nearly the same. Compare with table 3.

Second: repeating until convergence Amemiya's procedure of weighted least squares solves the first order condition for maximum likelihood with a gamma distributed dependent variable (and hence also for our variances model with normally distributed discrepancies). This can be seen as follows, starting with the first order condition for weighted least squares. In the notation of Amemiya (1973):

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}} \sum_t \frac{1}{\mathrm{Var}[y_t]} \left(y_t - \boldsymbol{\beta}'\mathbf{x}_t\right)^2 \\
&= \sum_t \frac{1}{\eta^2 \left(\boldsymbol{\beta}'\mathbf{x}_t\right)^2} \frac{\partial \left(y_t - \boldsymbol{\beta}'\mathbf{x}_t\right)^2}{\partial \boldsymbol{\beta}} \\
&= \frac{2}{\eta^2} \sum_t \frac{1}{\left(\boldsymbol{\beta}'\mathbf{x}_t\right)^2} \left(y_t - \boldsymbol{\beta}'\mathbf{x}_t\right) \mathbf{x}_t = 0
\end{aligned} \tag{29}$$

The last line shows Amemiya's equation (2.24): the first order condition for the maximum likelihood of the gamma model.

Amemiya's λ is known in our model to be one half. Amemiya finds that one-step optimally weighted least squares has the same asymptotic efficiency as ML with gamma distributed data.

C No ML estimation of the μ in the variances model

As noted above, it follows from (13) that the average of the Δy_{ij} is an unbiased estimator of μ. Can μ also be estimated with maximum likelihood, assuming normally distributed discrepancies? Following (21), the log of the density of discrepancy Δy_{ij} is:

$$\ell_{ij} = -\frac{1}{2}\left(\frac{(\Delta y_{ij} - \mu)^2}{V_i^{\text{exp}} + V_j^{\text{imp}}} + \log\left(V_i^{\text{exp}} + V_j^{\text{imp}}\right) + \log 2\pi\right) \tag{30}$$

Maximization with respect to μ alone is not possible, since the log density is not additively separable in μ and the other parameters.

Unfortunately it is also impossible to estimate μ simultaneously with the other parameters using maximum likelihood: the total loglikelihood $\sum_i \sum_j \ell_{ij}$ "can be made as large as desired" by taking μ equal to some Δy_{ij}, giving $\ell_{ij} = -(\log(V_i^{\text{exp}} + V_j^{\text{imp}}) + \log 2\pi)/2$. The $V_i^{\text{exp}} + V_j^{\text{imp}}$ can be made sufficiently close to zero by a proper choice of V_i^{exp} and V_j^{imp}. The quote is from Ferguson (1982), p. 831.

D ML estimation of the variances model (given μ) with GLM

Given a value of μ, the $(\Delta y_{ij} - \mu)^2$ are a set of sufficient statistics for the likelihood function of the V_i^{exp} and V_j^{imp} parameters. It follows from (26) that the $(\Delta y_{ij} - \mu)^2$ are distributed according to the gamma distribution. Hence (14) is a Generalized Linear Model (GLM) with gamma distributed data, as introduced by Nelder and Wedderburn (1972). The GLM allows us to estimate this with standard software.

The translation to the GLM is as follows. Define

$$s_{ij} \equiv (\Delta y_{ij} - \mu)^2 \tag{31}$$

The log of the density of s_{ij} is equal to the ℓ_{ij} in (30), apart from known terms, as follows, with the right-hand side being the standard GLM decomposition of a log likelihood:

$$\ell_{ij} - \log 2 - \frac{1}{2}\log s_{ij} = \frac{1}{\phi}\Big(s_{ij}\theta_{ij} - a\left(\theta_{ij}\right)\Big) + b\left(s_{ij}, \phi\right) \tag{32}$$

with $a(\theta_{ij}) = -\log(-\theta_{ij})$. The b function will be discussed later.

The left-hand side of (32) follows from the density of a square, with $s = x^2 \neq 0$:

$$f(s) = \frac{g(x)}{|\mathrm{d}s/\mathrm{d}x|} = \frac{g(x)}{|2x|} = \frac{g(x)}{2\sqrt{s}} = \exp\left(\log g - \log 2 - \frac{1}{2}\log s\right) \tag{33}$$

Following Wedderburn (1976) at the bottom of p. 28, it is assumed that all $\Delta y_{ij} - \mu$ are nonzero, which is almost surely the case.

In GLM vocabulary, the ϕ in (32) is the GLM scale (or "dispersion") parameter and θ_{ij} is the GLM canonical (or "natural") parameter, which is a function of $V_i^{\exp} + V_j^{\mathrm{imp}}$. In this case the GLM link function is the identity function, since $\mathrm{E}[s_{ij}] = V_i^{\exp} + V_j^{\mathrm{imp}}$. This link function is not the canonical link function for the gamma GLM. Unfortunately, with the identity link function the gamma GLM might have multiple local likelihood maxima; see Wedderburn (1976), the $Y = \mu$ line in table 1(a).

The value of ϕ can be derived by equating the $s_{ij}\theta_{ij}/\phi$ in the right-hand side of (32) with the corresponding term in (30). This gives

$$\frac{\theta_{ij}}{\phi} = \frac{-1}{2(V_i^{\exp} + V_j^{\mathrm{imp}})} \tag{34}$$

With a gamma GLM we have $-1/\theta_{ij} = \mathrm{E}[s_{ij}]$. Substitution into (34) and using $\mathrm{E}[s_{ij}] = V_i^{\exp} + V_j^{\mathrm{imp}}$ gives $\phi = 2$. Hence the b function in (32) does not depend on unknown parameters and is therefore irrelevant.

E The computation of table 3

The three local maxima in table 3 on page 12 are obtained by repeatedly setting one parameter to zero, without any other restriction. This gives 8 solutions, with 3 unique loglikelihood values. These solutions have been transformed such that restriction (19) holds, by adding a suitable number to the export variances and subtracting the same number from the import variances. (This number might be negative, of course.)

The result has been verified by using these parameters as a starting point of the estimation under restriction (19). The negative V in table 3 are started at zero, thereby making sure that a reader can also verify table 3 by programming the estimation herself, using only tables 1 and 3.

F Computing non-nested loglikelihood differences

The computation of a loglikelihood difference between non-nested models with a different type of probability distribution requires some care.

In the first place, the densities might have a different dimension. In our case, we have the normal density of the discrepancies in section 3 and the gamma density of the squared discrepancies in section 4.

In the second place, the software may omit terms and positive factors from the loglikelihood which are independent of the parameters. For instance, the Stata procedure for GLM with the gamma distribution prints the value of $\sum\sum s_{ij}\theta_{ij} - a(\theta_{ij})$ in (32), instead of the value of (32) itself. The Stata procedure for GLM with the normal distribution omits nothing, printing the value of (21).

Both problems have been solved by computing the loglikelihood (21) for both models after the parameter estimation. Note that this "normalization" has no effect on the loglikelihood differences in table 3: differences between values of (32) for various parameters vectors are equal to the differences between values of (21) for the same parameter vectors.

As anecdotal evidence of this problem: unfortunately in table B.1 of Ten Cate (2007) the concentrated loglikelihood of the means model was computed from the value of $\widehat{\sigma}^2$ and the latter was based on the unbiased estimate of σ^2, using the degrees of freedom; contrary to what is stated in that paper. This reduces the loglikelihood with $(12/2)\log(12/(12-(8-1))) = 5.25$. Hence the loglikelihood difference of 5 in table B.1, instead of our 10.

References

Amemiya, T. (1973). Regression analysis when the variance of the dependent variable is proportional to the square of its expectation. *Journal of the American Statistical Association*, 68:928–934.

De Beer, J., Raymer, J., Van der Erf, R., and Van Wissen, L. (2010). Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe. *European Journal of Population 26*, 26:459–481.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge University Press, Cambridge, UK.

Ferguson, T. S. (1982). An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association*, 77:831–834.

Gaulier, G. and Zignago, S. (2010). BACI: International trade database at the product-level; 1994-2007 version. www.cepii.fr.

Gehlhar, M. J. (1996). Reconciling bilateral trade data for use in GTAP. Technical report, www.gtap.org.

Gourieroux, C. and Monfort, A. (1994). Testing non-nested hypotheses. In Engle., R. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 44, pages 2583–2637. North-Holland Elsevier, Amsterdam.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, New York.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society Ser. A*, 135:370–384.

Poulain, M. and Dal, L. (2008). Estimation of flows within the intra-EU migration matrix. Technical report, GéDAP-UCL, http://mimosa.gedap.be/.

Stone, R., Champernowne, D. G., and Meade, J. E. (1942). The precision of National Income estimates. *The Review of Economic Studies*, 9:111–135.

Ten Cate, A. (2007). Modelling the reporting discrepancies in bilateral data. CPB Memorandum 179, www.cpb.nl.

Tsigas, M. E., Hertel, T. W., and Binkley, J. K. (1992). Estimates of systematic reporting biases in trade statistics. *Economic Systems Research*, 4:297–310.

Van Leeuwen, N. and Lejour, A. (2006). Bilateral services trade data and the GTAP database. CPB Memorandum 160, www.cpb.nl.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63:27–32.

Publisher:

CPB Netherlands Bureau for Economic Policy Analysis
P.O. Box 80510 | 2508 GM The Hague
T (070) 3383 380

August 2012 | ISBN 978-90-5833-562-3