Instrumental Variable Interval Regression

Giulia Bettin[1] and Riccardo (Jack) Lucchetti[2]

[1] HWWI - Hamburg Institute for International Economics - Hamburg, Germany
bettin@hwwi.org
[2] Dipartimento di Economia - Universita Politecnica delle Marche - Ancona, Italy
r.lucchetti@univpm.it

Abstract. In this paper, we introduce a maximum-likelihood estimator for grouped data with endogenous regressors and briefly analyse its properties. An example application to migrants' remittances is included, which shows that endogeneity effects are substantial.

1 Introduction

The estimation of interval models by maximum likelihood, introduced by [16], is nowadays relatively straightforward and has been applied in a number of contexts, most notably in willingness-to-pay double bound models (for a recent example, see [15]). The data generating process is assumed to be

$$y_i^* = x_i'\beta + \epsilon_i \tag{1}$$

where y_i^* is unobservable *per se*; what is observed are the limits of an interval that contains it, that is

$$m_i \leq y_i^* \leq M_i$$

where the interval may be left- or right-unbounded. Once a distributional hypothesis for ϵ_i is made, estimation becomes a simple application of maximum likelihood techniques. Under normality, the log-likelihood for one observation is

$$\ell_i(\beta, \sigma) = \ln P(m_i < y_i^* \leq M_i) = \ln \left[\Phi\left(\frac{M_i - x_i'\beta}{\sigma}\right) - \Phi\left(\frac{m_i - x_i'\beta}{\sigma}\right) \right] \tag{2}$$

and the total log-likelihood can be maximised by standard numerical methods, which are, in most cases, very effective. The above procedure is implemented natively in several econometric packages, among which Gretl, Limdep, Stata and TSP.

However, the extension of this model to the case of endogenous regressors seems to be absent from the literature. To consider this case, equation (1) can be

generalised to

$$y_i^* = Y_i'\beta + X_i'\gamma + \epsilon_i \tag{3}$$

$$Y_i = X_i\Pi_1 + Z_i\Pi_2 + u_i = W_i\Pi + u_i \tag{4}$$

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2 & \theta' \\ \theta & \Sigma \end{bmatrix}\right) \tag{5}$$

where θ, the covariance between ϵ_i and u_i may be nonzero. In this case, the vector of m explanatory variables Y_i becomes endogenous and ordinary interval regression does not provide consistent estimates of β and γ.

2 Estimation methods

2.1 Limited-information Maximum Likelihood

The estimation problem can be tackled by maximum likelihood. If y_i^* were observable, the log-likelihood for one observation could be written as follows

$$\ell_i(\psi) = \ln f(\epsilon_i, u_i; \psi) = \ln\left[f(\epsilon_i|u_i; \psi)\right] + \ln f(u_i; \psi) = \ell_i^a + \ell_i^b \tag{6}$$

where ψ is a vector containing all the parameters, possibly transformed via an invertible mapping. Since y_i^* is imperfectly observed, however, the above has to be modified as

$$\ell_i^a(\psi) = \ln P(m_i < y_i^* \leq M_i|u_i) \tag{7}$$

Thanks to the assumed joint normality, the distribution of $(\epsilon_i|u_i)$ is

$$\epsilon_i|u_i \sim N\left(u_i'\lambda, \tilde{\sigma}^2\right)$$

where $\lambda = \Sigma^{-1}\theta$ and $\tilde{\sigma}^2 = \sigma^2 - \theta'\Sigma^{-1}\theta$. Hence,

$$\ell_i^a = \ln P(m_i < y_i^* \leq M_i|u_i) = \ln\left[\Phi\left(\frac{M_i - \hat{y}_i}{\tilde{\sigma}}\right) - \Phi\left(\frac{m_i - \hat{y}_i}{\tilde{\sigma}}\right)\right]$$

where $\hat{y}_i = Y_i'\beta + X_i'\gamma + u_i'\lambda$, and ℓ_i^b is just an ordinary normal log-likelihood:

$$\ell_i^b = \ln f(u_i; \psi) = -1/2\left[m\ln(2\pi) + \ln|\Sigma| + (Y_i - W_i\Pi)'\Sigma^{-1}(Y_i - W_i\Pi)\right] \tag{8}$$

Of course, we assume that the instruments Z_i satisfy the order and rank identification conditions.

In order to guarantee that σ is positive during the numerical search, what is actually fed to the log-likelihood function is its logarithm. For similar reasons, the unconstrained parameters on which the log-likelihood function is based are

not the elements of Σ itself, but rather those of the Cholesky factorisation of its inverse. In practice, ℓ_i^b, the second component of the log-likelihood, is computed as

$$\ell_i^b = \text{const} + \ln |C| - \frac{\omega_i' \omega_i}{2}$$

where C is the Cholesky factorisation of Σ^{-1} and $\omega_i = C'(Y_i - \Pi' W_i)$. This produces faster and more accurate computation than evaluating (8) directly for two reasons: first, a matrix inversion is avoided; moreover, the determinant of C (which is by construction $|\Sigma|^{-1/2}$) is trivial to compute since C is triangular, via

$$-0.5 \ln |\Sigma| = \sum_{i=1}^{m} \ln C_{ii}.$$

The computational gain is negligible (arguably null) when $m = 1$, but may become substantial for $m > 1$; in fact, casual experimenting show non-negligible improvements even for $m = 2$.

A recent paper by [8] advocates the usage of the EM algorithm for dealing with numerical problems in a closely related case (the ordered probit model), but we found it unnecessary in our case.

The ML setup also enables us to build two hypothesis tests which are likely to be of interest: the first one is an exogeneity test, which is constructed by testing for $\lambda = 0$ by means of a Wald test. In addition, a LR test for overidentifying restrictions may be computed via the difference ℓ^a and that for an interval regression of (m_i, M_i) on W_i and $\hat{u}_i$, which would be the unrestricted log-likelihood.

2.2 Alternative estimators

In certain cases, it may be worthwhile to consider alternative estimators than ML. Two are briefly considered here, although no serious effort is made to analyse them in detail; both belong to the two-step category of estimators. As such, they may suffer from the typical shortcoming of two-step estimators: inefficiency and a cumbersome-to-compute covariance matrix[3]. Hence, we only sketch briefly the possibility for alternative estimators; proper analysis of their properties is left as a future project.

Both estimators depend on the availability of an ordinary interval regression routine. One possibility is:

1. perform first-stage OLS of Y_i on W_i and collect the residuals $\hat{u}_i$
2. perform an interval regression on Y_i, X_i and $\hat{u}_i$

[3] The obligatory reference here is [12], but see also [17], chapter 12.

This estimator ought to be consistent[4]. This estimator is very easy to compute if an interval regression routine is available: as a consequence, it was a natural choice for initialising our ML algorithm.

Another two-step estimator may be obtained by considering that, given an interval regression of the form (1), it is easy to form an unbiased estimator of y_i^* from

$$E(y_i^*|x_i, m_i, M_i) = x_i'\beta + E(\epsilon_i|x_i, m_i, M_i) = x_i'\beta + \sigma \frac{\varphi\left(\frac{m_i - x_i'\beta}{\sigma}\right) - \varphi\left(\frac{M_i - x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{M_i - x_i'\beta}{\sigma}\right) - \Phi\left(\frac{m_i - x_i'\beta}{\sigma}\right)}$$

and substituting unknown parameters with their estimates to get the estimate $\hat{y}_i$. The procedure is the following:

1. do an interval regression of (m_i, M_i) on W_i; that is, estimate the unrestricted reduced form of equation (3).
2. compute $\hat{y}_i$, an unbiased estimate if y_i^*; By construction,

   $$v_i \equiv y_i^* - \hat{y}_i$$

   will have the property $E(v_i|W_i) = 0$.
3. do TSLS using $\hat{y}_i$ as the dependent variable; this should be valid since (from equation (3)

   $$\hat{y}_i = Y_i'\beta + X_i'\gamma + (\epsilon_i - v_i)$$

   and the composite error term $(\epsilon_i - v_i)$ is incorrelated with the instruments W_i (although it will be heteroskedastic by construction).

Again, we conjecture that this estimator should also be consistent, but like the other one, it would be inefficient and the estimation of the parameters' covariance matrix would need a two-step adjustment.

3 Why bother?

From the viewpoint of an applied economist, the ML method outlined above may seem overkill. After all, how much inaccuracy do we introduce in the data by choosing the interval midpoint? In fact, a procedure that is commonly used is to approximate y_i^* by

$$\tilde{y}_i = \frac{M_i + m_i}{2}$$

[4] We do not have a formal proof, but it should follow from consistency of $\hat{\Pi}$ and the clear fulfillment of the identification condition stated in Wooldridge [17, p. 354].

and assume that $\tilde{y}_i$ can be used as a proxy for y_i^* more or less painlessly: an additional source of error in the model (most likely heteroskedastic), that could be accommodated via robust estimation of the parameters covariance matrix. Hence, running TSLS on $\tilde{y}_i$ may look as a simple and inexpensive procedure.

Trivially, a first problem that arises with this method is that it does not provide an obvious indication on how to treat unbounded observations (that is, when $m_i = -\infty$ or $M_i = \infty$). A more serious problem, however, is that that the above procedure leads to substantial inference errors. The analytical explanation is obvious after rearranging equation (3) as

$$\tilde{y}_i = Y_i'\beta + X_i'\gamma + (\epsilon_i + \eta_i), \tag{9}$$

where η_i is defined as $\tilde{y}_i - y_i^*$. The intuition behind this reasoning is that if the interval (m_i, M_i) is "small", then σ_η^2 should be negligible compared to σ_ϵ^2. (It should be noted that, by construction, the support of η_i is a finite interval, whose length goes to 0 as $M_i - m_i \to 0$.)

However, even if the basic instrument validity condition $E(\epsilon_i|W_i) = 0$ holds, there is no reason why the midpoint rule should guarantee $E(\eta_i|W_i) = 0$. This can be proven by a simple extension to the IV case of the line of reasoning in [16]. As a consequence, the TSLS estimator converges in probability to a vector that differs from the true values of β and γ. It is worth noting that inconsistency is not a small sample issue, but a much more fundamental flaw.

Clearly, how serious the problem is depends on the relative magnitudes of σ_η^2 and σ_ϵ^2. To explore the consequences of the above, in a seemingly harmless case, we run a small Monte Carlo experiment. The Monte Carlo setup is:

$$\begin{gathered} y_i^* = \gamma_0 + Y_i\beta + X_i\gamma_1 + \epsilon_i \\ \gamma_0 = \beta = \gamma_1 = 1 \\ Y_i = 1 + X_i + Z_i + u_i \\ \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}\right) \end{gathered}$$

and the cutpoints are represented by the vector $[-2, 0, 1, 2, 5]$, so that, for example, if $y_i^* = 3$, then $m_i = 2$ and $M_i = 5$. The variables X_i and Z_i are independent $N(0, 1)$. A "naive" proxy for y_i^* was constructed via the midpoint rule, as

$$\tilde{y}_i = \begin{cases} -4 & \text{for } y_i^* < -2 \\ \frac{M_i + m_i}{2} & \text{for } -2 < y_i^* < 5 \\ 10 & \text{for } y_i^* > 5 \end{cases}$$

The above DGP was simulated with sample sizes of 100, 500 and 2500 observations. For each case, 4096 simulations were run.

Table 1. Monte Carlo experiment: sample size = 100

	γ_0	β	γ_1
TSLS (mean)	1.2448	1.2538	1.2480
TSLS (median)	1.2361	1.2502	1.2498
LIML (mean)	1.0009	1.0044	0.99968
LIML (median)	0.9976	1.0017	1.0025
mean estimated s.e. (TSLS)	0.2684	0.2694	0.1910
mean estimated s.e. (robust TSLS)	0.2420	0.2629	0.1872
mean estimated s.e. (LIML)	0.1736	0.1791	0.1294
Monte Carlo s.e. (TSLS)	0.2483	0.2699	0.1952
Monte Carlo s.e. (LIML)	0.1810	0.1894	0.1361
size of t-test at 5% (TSLS)	0.1133	0.1467	0.3008
size of t-test at 5% (robust TSLS)	0.1699	0.1606	0.3037
size of t-test at 5% (LIML)	0.0603	0.0635	0.0549

Table 2. Monte Carlo experiment: sample size = 500

	γ_0	β	γ_1
TSLS (mean)	1.2463	1.2505	1.2489
TSLS (median)	1.2429	1.2501	1.2500
LIML (mean)	1.0005	1.0007	0.9997
LIML (median)	1.0003	0.9995	0.9987
mean estimated s.e. (TSLS)	0.1169	0.1169	0.08273
mean estimated s.e. (robust TSLS)	0.1073	0.1170	0.08343
mean estimated s.e. (LIML)	0.0773	0.0798	0.05756
Monte Carlo s.e. (TSLS)	0.1068	0.1186	0.0836
Monte Carlo s.e. (LIML)	0.0781	0.0810	0.0585
size of t-test at 5% (TSLS)	0.5603	0.5737	0.8428
size of t-test at 5% (robust TSLS)	0.6389	0.5684	0.8369
size of t-test at 5% (LIML)	0.0527	0.0552	0.0515

Table 3. Monte Carlo experiment: sample size = 2500

	γ_0	β	γ_1
TSLS (mean)	1.2469	1.2503	1.2485
TSLS (median)	1.2478	1.2503	1.2478
LIML (mean)	1.0004	1.0005	0.9995
LIML (median)	1.0009	1.0005	0.9990
mean estimated s.e. (TSLS)	0.05193	0.05193	0.03673
mean estimated s.e. (robust TSLS)	0.04777	0.05223	0.03722
mean estimated s.e. (LIML)	0.03449	0.03564	0.02570
Monte Carlo s.e. (TSLS)	0.04744	0.05329	0.03758
Monte Carlo s.e. (LIML)	0.03458	0.03608	0.02590
size of t-test at 5% (TSLS)	0.9985	0.9971	1.0000
size of t-test at 5% (robust TSLS)	0.9990	0.9971	1.0000
size of t-test at 5% (LIML)	0.0488	0.0552	0.0515

The results are summarised in tables 1–3, which are organised as follows: the first four lines report the Monte Carlo mean and median for the two estimators. The next three lines report the mean of the estimated standard errors: for TSLS both robust and non-robust versions are reported, while the robust "sandwich" estimator[5] is used for ML. The next two lines report the ex-post dispersion of the parameters, namely the standard error of the estimates across the 4096 replications. Note that estimated and Monte Carlo standard errors should roughly match, if inference is to be at all credible. The last group of three rows shows the frequency of rejection of the hypothesis that the corresponding parameter equals its true value at 95%.

The message should be rather clear: while the LIML estimator is consistent and remarkably reliable even at a moderate sample size, the application of TSLS to the naïve "midpoint" dependent variable proxy leads to seriously inconsistent estimates and substantial inference errors.

4 An Empirical Application: the Analysis of Immigrants' Remittance Behaviour

Remittance flows are certainly one of the most interesting aspects connected to international migration, drawing in the last decades the attention of economic literature.

Macroeconomic analyses of this phenomenon are usually built on aggregate data from countries' Balance of Payments and need to take care of the fact that these measure only official flows of remittances, while the huge amounts of money transferred through unofficial channels are not taken into account. This shortcoming is less affecting survey data, where generally information on remittances are collected regardless the channel used to send them in the country of origin.

At the microeconomic level, remittance behaviour of immigrants is usually analysed as a function of migrants' characteristics and of the household's welfare in the country of origin.

Since the pioneering work of [9] on Botswana, many attemps have been made to identify the motivations to remit: altruism, inheritance, self-insurance and so forth; for an exhaustive survey of the contributions on the topic, see [13]. Remittance behaviour, anyway, could hardly be expected to depend on a single driving force, since different motivations can coexist in the same individual. Moreover, discriminative tests are empirically difficult to build for the fact that surveys seldom account for characteristics of migrants together with

[5] See for instance Davidson and MacKinnon [4, chap. 10].

information on recipient households[6], that are both essential elements to infer explanations on the motivation to remit.

The most interesting and crucial aspect in our opinion is that the empirical literature dealing with the topic usually treats migrant's income as an exogenous determinant of remittance behaviour. Yet, the need of sending money back home can affect working, consumption and possibly also investment decisions. In order to remit more an immigrant could, for example, either decide to increase the number of hours worked per week, or invest a share of his savings and make profits out of it. The amount of money to remit (if any) is therefore determined jointly in the broader context of household's strategies. Hence, in our opinion the best way to address the problem would be estimating a remittance equation that detects the main determinants of remittance behaviour addressing endogeneity and reverse causality relationships between remittances, income, consumption and saving[7].

Another central point to be noted is that, as mentioned before, data for microeconomic analyses on remittance behaviour often are taken from household surveys and it is commonly the case that in surveys' questionnaires the amount of remittances is designed as a discrete ordered variable, with different intervals mutually exclusive. If the problem is then to analyse remittance behaviour dealing with a discrete ordered dipendent variable, and addressing reverse causality between remittances, income and consumption using IV techniques, the Gretl routine just illustrated is the instrument needed to carried out our estimations.

4.1 Data and estimation issues

The dataset used in this empirical application is the Longitudinal Survey of Immigrants to Australia (LSIA), a longitudinal study of recently arrived visaed immigrants undertaken by the Research Section of the Commonwealth Department of Immigration and Multicultural and Indigenous Affairs.

We consider the first cohort of the LSIA (LSIA1), that was selected from visaed immigrants aged 15 years and over, who arrived in Australia in the two year period between September 1993 and August 1995. The sampling unit is the Primary Applicant (PA), the person upon whom the approval to immigrate was

[6] An exception is represented by the paper by [11], where migrants are considered together with their respective origin-families. Such a complete information, on the other hand, come together with a very limited number of observation, 61 pairs.

[7] [5] propose a simple theoretical model where the optimal level of remittances and savings are jointly determined. However, being mainly interested in how temporary migration affect remittance behaviour, in the empirical part they only address the possible endogeneity of the decision to come back permanently to the home country.

based. The population for the survey consisted of around 75000 PAs and was stratified by the major visa groups and by individual countries of birth.

Individuals were interviewed three times: the first time five or six months after arrival, the second time one year later and the third a further two years later[8]. Questionnaires were divided into sections and each of them is related to a different topic: migrant's family in Australia and relatives left in the country of origin, the immigration process, the initial settlement in Australia, financial assets and transfers (remittances), working status, income, consumption expenditures, education and English knowledge, health, citizenship and return visits to the former country. All these information together give an incomparable socio-economic picture of immigrants, that is essential to understand their remittance behaviour.

The sample includes 5192 individuals, but only 3752 were interviewed in all the three waves. What is relevant here is that data do not concern only a specific ethnic group, but people from more than 130 different countries[9] (both developed and developing countries). As we will highlight later, the exploitation of this cross-country dimension is an important element of the present analysis. The remittance equation that we estimate can be written as:

$$r_i = \alpha_1^* y_i + \alpha_2^* c_i + \alpha_3^* X_i + u_i$$

where r_i represents the amount of money sent home every year, y_i the yearly income of the migrants' household and c_i the total yearly consumption expenditures. r_i, y_i and c_i are all expressed in natural logarithms.

Both income y_i and consumption c_i, our endogenous variables[10], are regressed on X_i together with another set of exogenous variable, Z_i, defined as instruments:

$$y_i = \beta_1^* X_i + \beta_2^* Z_i + \epsilon_i$$
$$c_i = \gamma_1^* X_i + \gamma_2^* Z_i + \upsilon_i$$

When immigrants were asked about the amount of money sent home, they had to choose between six different intervals: 1-1000 AUS $, 1001-5000, 5001-10000, 10001-20000, 20001-50000, more than 50000 AUS $.[11] Since observa-

[8] Unfortunately, the time between interviews may vary substantially between households; this problem, together with considerable sample attrition, led us to ignore the "panel" aspect of our dataset and use all data as pooled data.

[9] As a matter of fact, the vast majority of the contributes investigate remittance behaviour of a specific nationality of migrants. Exceptions are the studies carried on using data from the German Socio-Economic Panel [10, 5, 14, 7] and the work by [3].

[10] Strictly speaking, y_i and c_i are not observed continuously either, but expressed in intervals just like the remittance variable. Not to introduce further difficulties, we take the midpoints of the intervals.

[11] There is an explicit question - section T in wave 1, section F in wave 2 and 3 - where immigrants have to answer about the amount of money sent overseas. Moreover, immigrants were

tions concentrate mainly in the first two intervals, 1-1000 AUS $ and 1001-5000 AUS $, the upper four are reduced to a single one that goes from 5001 AUS $ upwards. The final outcome is therefore a variable r_i with three possible different outcomes:

$$r_i = \begin{cases} 1 \text{ for } 1 < R_i < 1000 \\ 2 \text{ for } 1001 < R_i < 5000 \\ 3 \text{ for } R_i > 5001 \end{cases}$$

where R_i represents the real amount of money remitted. Table 4 shows the frequency distribution for the remittance variable used in the estimations.

Table 4. Remittance behaviour of immigrants in LSIA1: frequency distribution

Amount remitted	Absolute Freq.	Cumul. Freq.	%	Cumul. %
1-1000 AUS $	1584	1584	64%	64%
1001-5000 AUS $	728	2312	29.4%	93.6%
more than 5000 AUS $	164	2476	6.6%	100%

X_i includes two different sets of control variables that can influence the remittance behaviour. The first one refers to immigrants' individual characteristics: age, a dummy for the gender, a dummy for the presence of close relatives (partner, children, parents, brothers) in the country of origin, another dummy for the intention to return to the home country and the time passed since the arrival in Australia. Moreover, the level of education attained is added as a further control. Migrants may actually send money at home to repay a loan used to finance their investment in human capital. If this was the case, the higher the level of education achieved, the higher should be the amount of money sent back to the family at home. Educational attainment is divided into five levels, the first corresponding to upper tertiary education and the last to primary education.

The second set of control variables includes macroeconomic characteristics of the countries of origin. If on one hand we cannot help but recognise that the biggest shortcoming of this dataset is the complete absence of any concrete information about remittances' recipients, necessary to deal exhaustively with the motivations to remit, on the other hand the wide set of countries of origin allows

asked also about the value of assets transferred from Australia to relatives or friends overseas, in the form of personal effects, capital equipment or funds. All these transfers should be considered as remittances, especially funds, but we are not able to put them together and hence have a broader measure of remittances for the different codification of the answers fixed in the questionnaire. Hence the analysis here refers to the specific question about the money sent overseas and does not consider other transfers.

us to consider if and how remittance behaviour is influenced by macroenonomic aggregates. What we use here is first of all the level of per capita GDP, as a general measure of the level of development and wealth[12]. Secondly, the level of financial development proxied by a measure of demand, time and saving deposits in deposit money banks as a share of GDP, taken from the widely employed dataset on financial structure built by [1] for the World Bank. The rationale behind this choice is that the decision to send money back home could be influenced by the trust in the domestic financial system, especially in case of return migration when immigrants could be inclined to invest or simply save money in their home country. Finally, the distance between Australia and the country of origin is considered to proxy somehow for the costs connected to money transfers that are likely to increase the farther the homecountry is[13].

As explained before, to address problems of endogeneity both income and consumption are instrumented using a set Z_i of five instruments. The first two instrument refer to the migrant's knowledge of the English language: we use two dummy variables, one for English being the language the immigrant speaks best and another one which equals 1 if the immigrant declared a good knowledge of English; we assume that language skills should influence income prospects but not remittance behaviour. The third instrument is a dummy variable stating if the immigrant lives in a urban or a rural environment. The idea behind this choice is that the level of consumption may different between urban and rural population, but this should not affect the amount of money sent home. A dummy for child presence in migrant's household is also employed, supposing that its incidence on remittances is limited to the effects of having children on consumtpion levels. Finally, the last instrument is represented by the number of migrant's household members, expressed in natural logarithm.

4.2 Dealing with the selection problem

While estimating the remittance equation just illustrated above, we do not consider selection problems that arise when the sample is made up of people who remit and people who do not, and the variable is therefore truncated below a zero threshold. Remittances actually could be equal to zero either because immigrants are not interested in remitting to anybody, or simply because they do not earn enough to send a share of their income overseas.

[12] Data are from the World Development Indicators database.

[13] Even if we are not entering the debate on motivations to remit, geographical distance could represent in a sense also a measure of the strength of family relationship with those left behind. The source of the data employed here is CEPII (Centre d'Etudes Prospectives et d'Informations Internationales) dataset on bilateral distances.

A solution widely used in the empirical literature on the topic is the estimation approach outlined by [6], following which the decision to remit is modeled as a two-stage sequential process. In our case, an extension of sample selection models à la Heckman that involves interval estimation as a second step instead of OLS does not exist in the literature, to the best of our knowledge, and is certainly not straightforward to conceive and implement. All the more so, when instrumental variable interval estimations are needed.

We have made a rough attempt to redesign the remittance variable adding one initial class that includes people who send zero AUS$ in the country of origin. These observations are singled out looking if at the yes/no question whether they remit, immigrants gave a negative answers. If this is the case, the observation is considered as zero.

The main idea behind this strategy is that, if the population were homogeneous, the problem coul be dealt with simply by considering the non-senders as having $-\infty < y_i^* \leq 0$ (in the language of equation (3)). In this scenario, the only possible reason for not sending money abroad is the budget contraint. If, on the other hand, there are people who would not send remittances whatever their income, then the two-stage decision process should be modelled separately.

Running again estimations with the zero-augmented dependent variable we get results that are quite different from the original ones, both in terms of significance and in terms of magnitude. This has to be read as a clear sign that sample selection is a problem we absolutely have to deal with, but at the same time it also shows that a suitable tool is needed. Considering a class of zeros *de facto* does not address correctly the selection mechanism, because we use a common model for the two different group of individuals and the estimation of a remittance equation actually does not make much sense for an immigrant who is not interested in sending money back home.

We prefer therefore not to introduce a strong source of heterogeneity by joining two samples (remitters and non remitters) that are most likely to be structurally different; we are aware that our results should be considered as conditional on the fact that the individual is a remitter.

The next step in our research then will be to include in the model a selection equation to control properly for the selection mechanism, but meanwhile the main focus of the work is on endogeneity treatment and the adoption of IV technique in interval estimations.

4.3 Results

Results are reported in Table 5. The first three columns show results obtained with simple interval estimations, while from column 4 onward we introduce IV techniques.

Table 5. Estimates for the Australian remittances data

	Non-IV			IV		
	[1]	[2]	[3]	[4]	[5]	[6]
const	-1.50	2.04	1.53	**13.03**	**17.35**	**17.48**
male	**0.26**	**0.31**	**0.31**	*0.24*	*0.29*	*0.32*
age	0.00	0.00	0.00	0.00	0.00	0.00
time in AUS	**0.50**	**0.49**	**0.48**	**0.51**	**0.51**	**0.51**
back home	*0.37*	*0.40*	*0.42*	*0.40*	*0.43*	*0.44*
relatives overseas	0.11	0.04	-0.09	0.03	-0.04	-0.08
qualifications_2	0.24	0.17	*0.33*	0.21	0.13	0.18
qualifications_3	-0.13	-0.16	-0.03	-0.19	-0.24	-0.16
qualifications_4	*-0.38*	*-0.43*	-0.24	-0.14	-0.24	-0.14
qualifications_5	**-0.60**	**-0.66**	**-0.54**	**-0.62**	**-0.71**	**-0.61**
per capita GDP	*0.12*	**0.19**	0.08	**0.22**	**0.29**	0.16
deposit			*0.14*			*0.16*
distance		**-0.41**	*-0.29*		**-0.43**	*-0.27*
income	**0.24**	*0.22*	*0.20*	**1.13**	**1.00**	**1.15**
consumption	**0.36**	**0.36**	**0.43**	**-2.16**	**-2.12**	**-2.31**
N	1136	1135	983	1132	1131	979
σ	1.20	1.19	1.19	1.46	1.43	1.48
Wald test				15.22	17.00	15.06
Wald test p-value				0.00	0.00	0.00
Over-id. test				9.75	7.31	3.98
Over-id. test p-value				0.02	0.06	0.26

Note: coefficients in **boldface** are significant at 1%; coefficients in *italics* are significant at 5%; coefficients in normal fonts are significant at 10%; coefficients in small fonts are not significant.

When not instrumented, both income and consumption are statistically significant with a positive sign. The result is expected for income, and in line with the previous findings of the empirical literature[14], but quite puzzling for consumption. If we consider remittances as a sort of savings, the natural prediction would be that they diminish as the level of consumption expenditure of the immigrants' household increases. This clearly shows how results are biased when we do not take reverse causality into account.

Moving to IV interval estimations (column 4-6), income and consumption are statistically significant at 1%, the former with a positive sign and the latter negatively. Elasticity of remittances to income is around 1-1.15, while elasticity to consumption is slightly bigger than 2. Remittances seem therefore much more responsive to change in consumption expenditure compared to change in the level of income.

Among individual characteristics, the age of immigrants seems not to influence their remittance behaviour, while gender differences result statistically significant: other things being equal, male migrants remit on average 30% more than female. The desire to return living in the country of origin predictably affects the amount remitted in a significant way, with potential returnees remitting around 40% more. Time elapsed from the arrival in Australia has also a positive and significant effect.

The presence of a close relative still living overseas does not play such a significant role in determining remittances. This somehow surprisingly result could be due to the fact that in the sample considered here almost everybody who sends money overseas has at least one close relative (spouse, children, parents, brothers/sisters) still living in the country of origin.

As far as the immigrants' education is concerned, just one out of four dummies is significant across all the specifications, with a negative sign, and is the one associated to the lowest level of education (primary school). What emerges is hence that, even after controlling for the level of income, more educated migrants are likely to remit higher amounts than the less educated.

The Wald test rejects firmly the exogeneity hypothesis for income and consumption. Endogeneity effects are therefore highly significant, so specification 1-3, which do not take this into account, must be regarded as incorrect. If covariances between residuals from the first steps and residuals from the remittance equation are considered, it is clear that the result of the Wald test is driven mainly by consumption that is strongly endogenous, while income is understandably less affected from problems of reverse causality. Moreover, the result from the LR test of over-identifying restrictions confirms the validity of the set

[14] Among others, see [2] and [3].

of instruments we have chosen to address endogeneity of income and consumption, especially in the complete specification.

As explained before, the cross-country dimension is taken into account adding to individual characteristics macroeconomic variables concerning the home countries. Surprisingly, per capita GDP of immigrants' country of origin turns out to be significant with a positive sign. Immigrants coming from richer countries seem to remit more. This result is confirmed when we consider also the distance between Australia and immigrants' home country (column 5).

The most interesting aspect is that per capita GDP loses all its explanatory power when the level of financial development is added as a further explanatory variable (column 6), while financial development is significant at 5% and plays an important role in immigrants' household decision. Per capita GDP hence seems to act, when significant, as a sort of proxy for the level of financial development of immigrants' country of origin, but this is indeed the macroeconomic feature that matters when immigrants consider how much money to remit.

5 Conclusions

We argue that estimation of models in which the dependent variable is observed by intervals and explanatory variables may be endogenous ought to be conducted via maximum likelihood, all the alternative possibilities being inefficient at best and plain wrong at worst.

An example with Australian remittances data shows that our procedure is effective. Endogeneity of income and consumption in the context of immigrants' remittance behaviour does matter. Consumption is strongly endogenous, while income is less affected from problems of reverse causality; anyway, endogeneity effects are altogether highly significant and need to be addressed in empirical models. Failing to account for them will lead to incorrect estimates.

Bibliography

[1] BECK, T., A. DEMIRGUÇ-KÜNT, AND R. LEVINE (2000): "A New Database on Financial Development and Structure," *World Bank Economic Review*, 14, 597–605.

[2] BROWN, R. P. (1997): "Estimating Remittance Functions for Pacific Island Migrants," *World Development*, 25(4), 613–626.

[3] CLARK, K., AND S. DRINKWATER (2007): "An Investigation of Household Remittance Behaviour; Evidence from the United Kingdom," *The Manchester School*, 75(6), 717–741.

[4] DAVIDSON, R., AND J. G. MACKINNON (1999): *Econometric Theory and Methods*. Oxford University Press, Oxford.

[5] DUSTMANN, C., AND J. MESTRES (2008): "Remittances and Temporary Migration," mimeo.

[6] HECKMAN, J. J. (1979): "Sample Selection Bias As a Specification Error," *Econometrica*, 47(1), 153–161.

[7] HOLST, E., A. SCHAEFER, AND M. SCHROOTEN (2008): "Gender, Migration, Remittances : Evidence from Germany," SOEPpapers 111, DIW Berlin, The German Socio-Economic Panel (SOEP).

[8] KAWAKATSU, H., AND A. G. LARGEY (2009): "EM Algorithms for Ordered Probit Models with Endogenous Regressors," *Econometrics Journal*, forthcoming.

[9] LUCAS, R. E., AND O. STARK (1985): "Motivations to remit: evidence from Botswana," *Journal of Political Economy*, 93, 901–918.

[10] MERKLE, L., AND K. F. ZIMMERMANN (1992): "Savings, Remittances and Return Migration," *Economics Letters*, 38, 77–81.

[11] OSILI, U. O. (2007): "Remittances and Savings from International Migration: Theory and Evidence using a Matched Sample," *Journal of Development Economics*, 83, 446–465.

[12] PAGAN, A. (1986): "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, 53(4), 517–538.

[13] RAPOPORT, H., AND F. DOCQUIER (2005): "The Economics of Migrants' Remittances," IZA Discussion Papers 1531, Institute for the Study of Labor (IZA).

[14] SINNING, M. (2007): "Determinants of Savings and Remittances: Empirical Evidence from Immigrants to Germany," IZA Discussion Papers 2966, Institute for the Study of Labor (IZA).

[15] SOLIÑO, M., M. X. VÁZQUEZ, AND A. PRADA (2009): "Social demand for electricity from forest biomass in Spain: Does payment periodicity affect the willingness to pay?," *Energy Policy*, 37(2), 531–540.

[16] STEWART, M. B. (1983): "On least squares estimation when the dependent variable is grouped," *Review of Economic Studies*, 50(3), 737–753.

[17] WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Mass.