PhD Thesis No. 152

Microeconometric Applications in Development Economics

by

Mikkel Barslund

June 2007
Department of Economics
University of Copenhagen
Studiestraede 6, DK-1455 Copenhagen
www.econ.ku.dk

Microeconometric Applications in Development Economics.

By

Mikkel Barslund

Ph.D. Thesis
Submitted to
Department of Economics
University of Copenhagen
January 2007.

Contents

Acknowledgements

During my work on this Ph.D. thesis I have been lucky to receive inspiration from a number of gifted and talented people. Some of whom I have even had the pleasure to co-author papers with. The environment that all of you contributed to at Bispertorvet has always been both mind challenging and a good deal of fun to work in. To all my former and present colleagues I would like to say: Thanks a lot!

Professor Finn Tarp, my supervisor, provided opportunities for field work and activities I could not have gotten better anywhere else, and he was instrumental in getting me started with my work on this thesis – not least because of the three months we spent together in Vietnam some four years ago. Finn also has my gratitude for always being willing to read and comment on more or less worked through drafts of papers.

A special thanks to Thomas Barnebeck Andersen with whom I have spent innumerable hours discussing topics related to and – quite often – not related to development economics. Also warm thanks to Vibeke Kovsted and Leise Kjær for making all kinds of tasks a lot easier.

Thanks to you, Lisbeth and Sabi, for the countless dinners I have eaten at your place and for lending a friendly ear the (few) times I found Ph.D. live frustrating. Better friends are hard to come by! Your son, Magnus, helped me to get out of the office and clear my mind by always stressing that “Mikkel, det er altså længe siden vi har spist is og leget …”.

Dear family, although I am not always sure you think, that what I do makes sense, you have always supported me and without your backing I would never have made it to here! It is very much more appreciated than you think.

One other person was somewhat helpful in the process. Katleen, you encouraged me, read most of the content, corrected and came with suggestions and were always cheerful and positive. All this would of course count for very little if you were not around me everyday… Thanks for always being there for me!

Copenhagen
January 2007
Mikkel Barslund

Introduction and summary

This collection of essays constitutes my PhD thesis, submitted to the Faculty of Social Sciences at the University of Copenhagen in January 2007. It consists of five chapters. They are self-contained and can be read independently, but they have in common their use of microeconometric methods to address the research questions identified. A large part of my PhD has been dedicated to the study of development issues in two rapidly developing countries, Vietnam (chapter 1) and Mozambique (chapters 2, 3 and 4). The first four chapters rely on analysis of household survey data. The last chapter is methodological using simulated micro data. As such, this chapter is not specific to any country, but the research was inspired when writing chapter 4. A summary of the content of each chapter is given below.

Chapter 1 on rural credit in four Vietnamese provinces is co-authored with Finn Tarp. As background for the written output in this chapter, a significant amount of field work and data collection in Vietnam was involved. Here I was active in the process of training enumerators, piloting the survey and cleaning the data, and the formulation of the questionnaire was accomplished when I joined the research team.

It is widely recognised that credit is important to small scale farmers in developing countries, and with the Nobel Peace price recently awarded jointly to the Grameen Bank and its founder Muhammad Yunus the issue has received renewed attention. In the Vietnamese context not much is known about the rural credit market. This chapter attempts to expand this knowledge. The first part of the chapter is a descriptive analysis of credit conditions faced by rural households in Vietnam. The data set allows us to look at the development of the rural credit market from 1997 to 2002. This was a period of rapid economic growth in Vietnam and this is reflected in the credit market. First, loan volumes grew substantially as a result of an increase in the average loan size and the number of households obtaining a loan. Second, the composition of the rural credit market in terms of credit sources also changed during the period under consideration. On the one hand, the loan volume share obtained from informal (unregulated) sources (friends, private money lenders etc.) fell by 30 percent (from 24 to 17 percent of the total loan volume). On the other hand, the analysis also shows that around one third of all loans continue to originate from informal credit sources. This is likely to be a consequence of formal credit sources supplying loans almost entirely for production purposes, whereas informal loans are often used for health expenditures and consumption smoothing. Throughout the period under study, regional differences are apparent in all aspects of the rural credit market. In the South, most rural households participate in the credit market. In contrast, in the mountainous central part of the country little activity is going on in the credit market.

The second part of chapter 1 investigates the household level determinants of credit demand and credit rationing in 2002. We find that demand from formal versus informal sources is distinct, and that regional differences persist even when controlling for a range of household level covariates. With respect to credit rationing information on formal and informal segments were pooled due to the moderate sample size. The probability of being rationed in access to credit depends negatively on education and connections, positively on a bad credit history and, again, on the province in which the household is located. Encouragingly, no evidence of gender discrimination seems to be present in the data.

Chapter 2 is co-authored with John Rand, Finn Tarp and Jacinto Chiconela and contains an analysis of the risk of being victimized in Mozambique. A nationally representative household survey (known as the IAF2003) with a unique module including questions on individual victimization and the economic loss from being victimized allows us to analyse the effects of individual, household and enumeration area level determinants on the risk of being victimized. In addition, we investigate the relationship between relative monetary loss (loss over a consumption measure) and income. Combining the IAF2003 survey with additional information from the 1997 Mozambican census data further expanded the scope of the analysis. The key characteristic of the chapter is that we are able to control for two sets of explanatory variables. A first set of variables coming from the economic literature focuses on offender motives, while a second set from sociological studies focuses to a larger extent on the characteristics of the victim. This integrated approach turns out to be beneficial. Variables associated primarily with the sociological strand of the literature are significant. We find that the risk of being victimized is increasing in income but at a diminishing rate, and district level inequality increases the overall risk of being the victim of a crime. While poorer individuals are less at risk of being victimized they tend to suffer a bigger loss relative to their consumption level. An extensive robustness analysis of the results is provided in addition to a range of policy recommendations. A revised version of the paper is forthcoming in *World Development*.

Chapter 3 is concerned with one of the many unfortunate aspects of the AIDS pandemic in Africa – orphaned children. Together with co-authors Channing Arndt, Virgulino Nhate and Katleen Van den Broeck, I study whether children living in Mozambican households with a head, who is not their biological parent, are being discriminated. The background is a recent increase in the number of young orphans and projections based on HIV/AIDS prevalence rates. They show a continuous increase in the years to come. It is therefore of relevance to explore the existence of discriminatory effects of the current official policy of trying to integrate orphans in extended families. To analyse differences in intra household allocation of resources, both Deaton's indirect 'adult good' method and the direct Engel curve method studying purchases of children's clothing and educational expenditure are employed. Both methods have been used extensively in the literature to study boy/girl discrimination, but our study is – to the best of the authors' knowledge – the first time it has been applied to an orphan/non-orphan setting. While we find no evidence of discrimination in the full sample, the 'adult good' method shows evidence of discrimination against orphans in the sample of poor households. Similarly, purchases of children's clothing indicate discrimination in the rural sub-sample. When interpreting the results, it should be kept in mind that both the direct and the indirect method often fail to show significant (boy/girl) discrimination where it is known from individual data records to exist. Thus, any significant discrimination showing up should be cause for concern. In sum, we conclude that there is weak evidence of discrimination of orphans in Mozambique, at least for sub-samples of the national representative data set (IAF2003). A revised version of this paper has appeared in the *American Journal of Agricultural Economics* (Vol. 88, 2006).

In chapter 4, a detailed censored food demand system is estimated using Mozambican household survey data (IAF2003). Estimating censored demand systems poses two related challenges: formulation of an econometric model that takes censoring of the expenditure shares into account and estimation of the chosen model. I address the problem of censoring by formulating the model as a system of Tobit equations with correlated errors. This large system of equations is then estimated with a recently suggested Quasi Maximum Likelihood (QML) estimator, and an appendix documents the Stata command specifically written for this research activity. The deterministic demand functions are derived from a translog indirect utility function, which is augmented with demographic and household location variables, so as to facilitate demographic and location specific parameters and elasticities. The contribution of the paper is twofold. First, it provides the first set of estimates on Mozambican data of income and cross-price elasticities for the most important food groups. Second, a method is developed to test for significant differences in elasticities among household locations. Significant differences are found among the central, southern and northern parts of Mozambique for both income and own price elasticities. The findings serve as an input into the continuing process of constructing a regional CGE model for Mozambique.

Chapter 5 contains a Monte Carlo comparison of the performance of three estimators to estimate systems of Tobit models, which are often used in censored demand system contexts. The study was motivated by the application of this type of model to Mozambican data in chapter 4. The Quasi Maximum Likelihood (QML) estimator employed in chapter 4 is based on maximizing a sequence of bivariate Tobit models. This approach circumvents the need to evaluate higher dimensional integrals over the multivariate normal distribution necessary with the full information maximum likelihood estimator (FIML). Thus, the FIML estimator relies on methods to simulate the higher dimensional integrals. While the FIML estimator is more efficient, there are two main drawbacks in applying it. It is programming and computational intensive and it is difficult to achieve convergence from arbitrary starting values. Hence, it is of value to search for well performing simpler estimators. In addition to comparing these two estimators, I introduce another – simpler – QML (sQML) estimator based on the maximization of a sequence of univariate Tobit models. Both the QML and sQML estimators show good performance relative to the more cumbersome FIML estimator for empirical relevant error correlation structures. This is so although their performance deteriorates with large (in absolute value) error correlation coefficients. The study lends support to the use of both QML estimators in censored demand system applications, and the principle should be applicable to more general systems of censored equations.

Although different in scope each of the five chapters of this PhD dissertation were developed to contribute to the analysis of a set of policy relevant and methodological issues in the context of two interesting countries, Mozambique and Vietnam. It is my hope that the reader will agree that my Ph.D. thesis provides further insights into the covered areas of development economics.

Formal and Informal Rural Credit in Four Provinces of Vietnam*

Mikkel Barslund and Finn Tarp

University of Copenhagen, Development Economics Research Group (DERG)

May 2004
Revised: January 2007

Abstract

This paper uses a survey of 932 rural households to uncover how the rural credit market operates in four provinces of Vietnam. Households obtain credit through formal and informal lenders. Formal loans are almost entirely for production and asset accumulation, while informal loans are used for consumption smoothening. Interest rates fell from 1997 to 2002, reflecting increased market integration. Moreover, the determinants of formal and informal credit demand are distinct. While credit rationing depends on education and credit history, in particular, regional differences in the demand for credit are striking. A 'one size fits all' approach to credit policy in Vietnam would be inappropriate.

*JEL classification***:** O12, O16, O17, O18

*Keywords***:** rural credit, household survey, Vietnam

---

* Mikkel.Barslund@econ.ku.dk and Finn.Tarp@econ.ku.dk, Department of Economics, University of Copenhagen, Studiestræde 6, DK-1455 Copenhagen K, Denmark. We are grateful for productive and stimulating collaboration with the survey teams from the Vietnamese Institute of Labour Science and Social Affairs (ILSSA) coordinated by Nguyen Huu Dzung and Dao Quang Vinh. Thanks are also due to Mekong Economics for excellent interaction on the design and implementation of the survey underlying this study. We appreciate comments from participants in various seminars in Vietnam and would like to acknowledge financial support from Sida through the Stockholm School of Asian Studies (SSAS) at the Stockholm School of Economics and from Danida. Expert advice from two anonymous referees also helped improve the paper considerably. For this we are grateful. The usual caveats apply.

1. INTRODUCTION

Vietnam has come a long way since the *doi moi* reform process was initiated in 1986, and the past 15 years have witnessed one of the best performances in the world in terms of both economic growth and poverty reduction. People's living standards have improved significantly, and the country's socio-economic achievements are impressive. Wide-ranging institutional reforms have been introduced, including greater reliance on market forces in the allocation of resources and the determination of prices. A shift can also be noted from an economy dominated by the state and co-operative sectors to a situation where the private sector and foreign investment account for a relatively high proportion of GDP. Important strides have been made over a relatively short time span to further the transition from a centrally planned to a socialist market economy. Finally, while the ratio of credit to GDP is almost twice as high in Thailand and three times as high in China and Malaysia (see World Bank, 2005), the financial deepening of the Vietnamese economy that has taken place during the past decade is remarkable.

Nevertheless, Vietnam remains a poor country. Some 70 percent of the population continues to live in rural areas, and they depend on agriculture for their livelihood. How the country can transform itself and its agricultural sector to a more modern society remains a critical policy challenge. Access to credit for smallholders is as elsewhere a key ingredient in the promotion of agricultural production and transformation. It forms an essential element of any poverty oriented strategy for the future development of the financial system.[1] Access to credit affects as aptly demonstrated by Diagne, Zeller and Sharma (2000) household welfare through at least two key channels. First, it alleviates capital constraints on agricultural households. This can significantly improve the ability of poor households to procure needed agricultural inputs, and will also reduce the opportunity costs of capital-intensive assets, encouraging labour-saving technology and raising labour productivity. The second channel identified by Diagne et al. is that credit access increases the risk-bearing capacity of households, altering risk-coping strategies. Households with access to credit may be more willing to pursue promising but risky technologies, and will be better able to avoid adopting risk-reducing but inefficient livelihood strategies.

The above kinds of considerations have as elsewhere in the developing countries led the Vietnamese Government and its donor community to set up credit programmes aimed at expanding rural households' access to credit; and significant expansion is foreseen in the coming years (see World Bank, 2003). The reliance on informal credit continues, however, to be widespread. Formal and informal credit market segments are present in Vietnam much along the lines of the dual credit market described by for example Mohieldin and Wright (2000). They cite Hoff and Stiglitz (1993), and point out – with reference to Egypt – that there are two competing views as to why formal and informal credit markets co-exist. First, government may intervene, capping interest rates, and this remains the case in Vietnam. The alternative view that differences in the cost of screening, monitoring and contract enforcement across lenders lead to fragmentation appears, however, also to carry explanatory power. Similarly, the interaction between the formal and informal credit market segments is open to conflicting interpretations. This is evident in the theoretical papers by for example Gupta and Chaudhuri (1997) and Chaudhuri (2001), on the one hand, and the careful empirical work by Zeller (1994), Diagne (1999) and Diagne, Zeller and Sharma (2000) on Madagascar, Malawi and Bangladesh, on the other. Diagne and co-authors highlight that understanding how informal institutions serve the financial needs of households and interact with the formal credit institutions is important, especially for 'sustainable and market-oriented financial institutions that plan to expand and complement the services offered by the existing informal credit market rather than substitute for them'. Diagne, Zeller and Sharma also offer a concise methodological review, which together with papers by Kochar (1997) and Petrick (2005) provide general analytical background for the present work on Vietnam. Kochar points out in the context of India that the literature on rural credit has generally assumed that households are rationed in their access to subsidized 'formal' credit; but she adds that the validity of this assumption hinges on the level of effective demand for formal credit, which is in turn a function of the demand for credit and its availability from 'informal' sources. This implies that the extent of credit market rationing may be smaller than regularly assumed. We take these cautioning findings serious and rely on them in our attempt to get credit demand right in our study of formal and informal rural credit in Vietnam.

In any case, a key motivation for our paper is that very little is actually known about the rural credit market in Vietnam, including both its degree of efficiency and the extent to which credit rationing impedes agricultural development. Appropriate development of market institutions based on well informed policies is a key prerequisite for success in Vietnam's ongoing transformation from a command-type to a more market based economy. Generating policy relevant insights into the characteristics and functioning of the rural credit market are on this background well justified. It is in this context helpful that the general academic literature on rural credit markets and their importance in developing countries (including the analysis of determinants of credit demand and the characteristics of credit constrained households) has seen a welcome expansion during the last 15 years. This has followed Japelli (1990) and Feder et al. (1990).[2] They relied on respectively household survey data from the US and China, and this methodological approach has subsequently been put to good use in most of the papers cited above. Our study is situated within this literature, and it relies on the methodological approach, which Diagne, Zeller and Sharma (2000) refer to as the 'direct method'. Accordingly, our household survey data allow us to establish whether households are credit constrained or not. They do not permit an analysis where the extent to which a household is credit constrained is in focus, even if we agree this would be desirable.

In sum, in this paper we provide a detailed review and an in-depth econometric analysis of how the rural credit market operates in four provinces of Vietnam, with a focus on basic characteristics and differences between the formal and informal credit markets.[3] We use a new survey of 932 households designed to elicit the full credit history of households during 1997 to 2002. These data are combined with information from the 2002 Vietnam Household Living Standard Survey (VHLSS) in the econometric analysis, where the determinants of credit demand and credit rationing are identified more rigorously. We are in this process able to account carefully for possible self selection.

The paper is structured as follows. After describing the data in Section 2, we provide in Section 3 a detailed descriptive overview of the characteristics of the rural credit market with a focus on the division between formal and informal credit. The data set has a time dimension, so trends during the 1997-2002 years can be spelled out, including developments in overall interest rates. In Section 4, we apply the econometric framework to identify the determinants of credit demand, and proceed to analyse in Section 5 household characteristics, which potentially influence the probability of being credit rationed. Some key policy measures to further the allocation of rural credit in Vietnam and develop the credit market overall are discussed in the concluding Section 6.

2. DATA

Key data used in this paper (including in particular information on the demand for credit) were generated in a comprehensive household survey of land, labour and credit markets in the provinces of Long An, Quang Nam, Ha Tay and Phu Tho. The survey, also known as the ILSSA Access to Resources Survey,[4] was carried out in the first quarter of 2003 in collaboration among the Institute of Labour Science and Social Affairs (ILSSA), Mekong Economics, the University of Copenhagen and the Stockholm School of Economics (see Mekong Economics, 2004). A total of 932 rural households were surveyed. These households are identical to the rural households previously interviewed in quarter 1 and 2 in the rural areas of the four provinces under study here as part of the nationally representative 2002 Vietnam Household Living Standard Survey (VHLSS).[5] In the VHLSS 2002, data were collected on income, expenditure and various other background variables. This largely pre-determined information is used in this paper in combination with our own data, collected about a year later to construct explanatory variables.[6]

The four provinces studied are located in four different regions of Vietnam as follows: (i) Long An in the fertile Mekong Delta, which is also the most densely populated of the four provinces; (ii) Quang Nam in the sparsely populated Central Highlands; (iii) Phu Tho in the North Western (Highlands), a mountainous region with a high share of ethnic minorities, and (iv) Ha Tay in the Red River Delta surrounding Hanoi, the Capital of Vietnam. The ILSSA survey is not nationally representative, but it is representative for rural households in the four provinces under study. They cover a lot of the variation in geographical and socio-economic conditions present in Vietnam, including regional differences between the north, centre and south of the country.

The ILSSA survey covered a large variety of topics related to land, labour and credit. In this paper, we rely on the credit component, including a number of illuminating questions on the source and use of loans, designed to elicit the full credit history of households during the recent past.[7] The general purpose of this part of the questionnaire was to help clarify the functioning of rural credit markets in Vietnam and to assess the extent to which credit rationing constrains agricultural development.[8] Questions covered issues such as (i) number of loans applied for and actually received, including information on amounts involved, interest, period and source of the credit, (ii) whether the household had at some point wanted to apply for a loan, but refrained from doing so, and (iii) various other relevant background such as the use of the loan, collateral requirements etc.

3. THE RURAL CREDIT MARKET

Due to the design of the questionnaire the credit history of each household in the sample can be followed. Table 1 shows the distribution of households by the number of loans obtained.

Table 1. ***Households distributed by number of loans obtained, 1997-2002***

Number of loans	Frequency	By province (percent)			
		Ha Tay	Phu Tho	Quang Nam	Long An
0	289	29	18	53	23
1	211	19	25	40	7
2	149	22	24	4	12
3	112	17	17	1	11
4	52	6	8	1	6
5	119	7	8	1	40
Total	932	100	100	100	100

Source: Authors' calculations based on ILSSA Access to Resources Survey 2003.

Over the period from the beginning of 1997 to 2002, a total of 289 households did not obtain any credit at all. However, 69 percent of the sample (643 households) obtained at least one loan, and around 46 percent (432) obtained more than one loan. Table 1 also reveals that there are differences among the four provinces. In Quang Nam less than 50 percent of the households obtained a loan, whereas 71 percent secured at least one loan in Ha Tay. In Phu Tho and Long An around 80 percent of the households participated in the credit market. If we focus on households with more than one loan, Ha Tay and Phu Tho are quite similar with more that 50 percent having more than one loan. In Quang Nam only 7 percent of the households obtained more than one loan in contrast to Long An where the corresponding share is more than two thirds.

Of the 289 households, who did not participate in the credit market during the period under study, only 12 got a loan application rejected, and another 65 reported having at some point refrained from applying even though they wanted credit. Thus, many of the 289 households can be seen as not effectively demanding credit. In sum, the overall picture emerging from Table 1 is that an active rural credit market exists in Vietnam and that regional differences are sizeable.

(a) *General trends*

The supply side of the rural credit market in Vietnam includes a number of formal and informal lending institutions. The Vietnam Bank for Agriculture and Rural Development (VBARD) is the biggest formal lender, and the much smaller Vietnam Bank for the Poor (VBP) is associated with VBARD.[9] VBP specialises in lending to poorer households. The credit market in many developing countries is characterised by segmentation in formal and informal sectors (see for example Zeller, 1994 and Yadav et al., 1992). Table 2 shows the distribution of loans by source of credit in terms of both percentages of all loans and percentages of all loans weighted with loan size. As revealed in Table 2, there is a sizeable informal credit sector in Vietnam. The informal sector consists of private money lenders, friends and relatives,[10] responsible for 35 percent of all loans in 2002.

In terms of loan amounts, the importance of the informal sector declined from 21 percent in 1999 to 17 percent in 2002, but measured by the actual number of loans the relative importance of the informal sector actually increased slightly. The figures in Table 2 compare well with previous work on credit markets in Vietnam. Duong and Izumida (2002), using data from a small household survey undertaken in 1995, found that the informal sector accounted for 17 percent of all loans.

Table 2. ***Distribution of loans by source (percent)***[a]

	1999		2002	
	Unweighted	Weighted by loan amount	Unweighted	Weighted by loan amount
VBP	11	5	5	2
VBARD	44	64	38	56
Private lenders	8	6	11	4
Relatives	23	15	24	13
Union	9	3	12	7
Others	5	7	10	18
Total	100	100	100	100

Source: Authors' calculations based on ILSSA Access to Resources Survey 2003.
Note: 'Unweighted' refers to the simple distribution of the number of loans. 'Weighted by loan amount' indicates the distribution of loans where each loan is weighted with loan size.
[a] VBP (Vietnam Bank for the Poor, now Vietnam Bank for Social Policies, VBSP), VBARD (Bank for Agriculture and Rural Development), Private Lenders (Private moneylenders and traders, and friends charging interest), Relatives (relatives and friends charging zero interest), Union (Farmers'/Veterans'/Women's Unions and People's Credit Funds), Other (Other institutions not mentioned above – see Appendix A)

'Others' include private banks, which have expanded rapidly in the south of Vietnam in recent years, and the sector composition of the rural credit market differs markedly among provinces. In Long An the formal sector provided 96 percent of the total loan amount in 2002 whereas only 64 percent came from the formal sector in Phu Tho, as further discussed in Section 3.3.

In what follows, we divide the rural credit market into three different segments, one formal and two informal. The formal segment includes all formal institutions,[11] while the informal sector consists of (i) private lending by unrelated individuals and friends charging interest, and (ii) lending from families, relatives and friends carrying zero interest. These two segments will be referred to as ‘private’ and ‘family’ in what follows. The distinction between friends, who lend and charge interest, and friends, who lend charging zero interest, may seem arbitrary. However, the data reveal a marked discontinuity at zero interest. Friends, who lend and charge interest, charge on average only slightly less than private money lenders (not characterised as friends).

To illustrate developments in the rural credit market in the late 1990s and early years of the new millennium, Table 3 shows the number of loans, the average loan size (in nominal terms) and the average monthly interest rate for the three different segments, year by year. To judge the magnitude of real interest rates the average monthly consumer price inflation for each year is also shown. The nominal overall volume of credit expanded rapidly by a factor of 2.6 in the years from 1999 to 2002. During this period Vietnam experienced an average annual consumer price inflation rate of around 1.5 percent, so the credit volume in real terms grew at about 6 percent less than the nominal growth.

Looking at the number of loans disbursed in the period, relatives and the informal sector increased their share from 29 to 36 percent, but in terms of loan amounts formal sector lending increased significantly. Formal credit accounted for 76 percent of total rural credit in 1997. By 2002 this share was 83 percent. The remaining 17 percent was divided almost equally between informal loans and loans from relatives.

The trend described above is mirrored in the development of loan sizes in the three segments. While loan size increased steadily in the formal sector, it remained almost constant for friends and relatives and decreased substantially in the interest bearing part of the informal sector.

Table 3. ***Rural credit, 1997-2002***

	1997	1998	1999	2000	2001	2002
Formal						
Loan size ('000 Dong) [a]	5,191	4,657	4,583	5,360	6,400	8,426
Interest (percent per month)	1.2	1.1	1.0	0.9	0.9	0.9
Number of loans	70	130	168	223	279	250
Informal – interest						
Loan size ('000 Dong)	3,222	7,686	3,196	3,206	2,468	3,904
Interest (percent per month)	3.8	3.8	3.6	3.0	3.0	1.8
Number of loans	9	18	24	31	47	55
Relative – zero interest						
Loan size ('000 Dong)	4,175	2,107	2,375	2,522	3,558	2,602
Interest (percent per month)	0	0	0	0	0	0
Number of loans	20	29	51	69	76	84
Total						
Loan size ('000 Dong)	4,807	4,548	3,983	4,547	5,403	6,529
Interest (percent per month)	1.2	1.2	1.0	0.9	1.0	0.8
Number of loans	99	177	243	323	402	389
Consumer price inflation						
Monthly consumer price inflation (pct.)	0.26	0.62	0.34	-0.13	-0.03	0.33
Distribution by source, unweighted (percent)						
Formal	71	73	69	69	69	64
Informal	9	10	10	10	12	14
Relative	20	16	21	21	19	22
Distribution by source, weighted by loan size (percent)						
Formal	76	75	80	81	82	83
Informal	6	17	8	7	5	8
Relative	18	8	13	12	12	9

Sources: Authors' calculations based on ILSSA Access to Resources Survey 2003 and IMF, World Economic Outlook Database, September 2006.
Note: 'unweighted' and 'weighted by loans size' as defined in Table 2.
[a] At the time of the survey in January 2003 the exchange rate was around 14,000 VND per USD.

Table 3 also allows us to investigate the development in loan terms. One striking feature is that overall interest rates have fallen – and more so for informal sector loans. The trend for real interest rates is less clear due to fluctuating inflation over the period. However, real interest rates in the formal sector for 2002 are in the low end for the period, and for the informal interest bearing segment there has been steady decline. The interest rate gap between the formal and informal sector was around 0.9 percentage point (per month) in 2002. The relatively large fall in the interest rate in the informal sector (for interest bearing loans) is clearly related to the general increase in rural incomes, which made borrowing less risky. This has tended to push interest rates down, and the same goes for the increase in formal credit possibilities during the period.

Another factor behind the interest rate fall is that monopoly rents obtained by private moneylenders are likely to have fallen in line with increased market integration. Increased access to collateral (in the form of red books, which are land tenure certificates issued by local authorities) have squeezed profit margins and the degree of risk associated with the portfolios of informal lenders.

Table 3 confirms that the combined informal sector is important in Vietnam with 36 per cent of the total number of loans in 2002. The interest bearing segment made up 14 percentage points hereof and about half in value terms. This suggests that poor rural households in Vietnam continue to rely on networks and relatives when they try to deal with shocks and face hard times. This is in line with what is generally found in the literature on rural households in developing countries, see Platteau (1997).

Looking at the changes in the structure of the credit market it is of interest to relate these to potential changes in the use of approved loans. Table 4 shows that such changes were limited in the sample.[12] It is highlighted that to increase the probability that the correct use of each loan was elicited, we asked both about the stated purpose in the loan application and about what the loan was actually used for.[13] Combining answers to these two questions suggests that loans were generally used as stated in the applications. In all years differences were identified in less than 5 percent of the loans, and these differences are not systematic in any way. However, even if loans are generally used for the purpose applied for, fungibility in the form of substitution and diversion – using the terminology of Von Pischke and Adams (1980) – can still be present. Substitution occurs when a household obtains a loan for a project or part of a project the household would still have undertaken in the absence of the loan. Diversion of a (small) part of the loan to other purposes can happen even if the main share of the loan is still used for the purpose stated in the application. Table 4 mainly indicates that changes in the structure of the credit market are not driven by changes in loan composition in terms of use of loans.

Table 4. ***Loan use (percent of total loans each year), 1997-2002***

Year	Production	Repayment of existing loan	Asset accumulation	Health	General consumption
1997	69	3	18	9	2
1998	70	2	11	3	15
1999	74	2	14	4	6
2000	73	3	11	4	9
2001	71	3	12	6	9
2002	68	4	12	6	11

Source: Authors' calculations based on ILSSA Access to Resources Survey 2003.

(b) *Land and credit market interaction*

Credit is obtained for many reasons, such as consumption smoothening and investment. Investment in land (including in particular land transactions) is critically important for the development of a market based economy and for the efficiency of the economy in general. It is therefore of interest to uncover any interactions between the credit and land markets. The credit and land sections of the ILSSA questionnaire were on this background designed to capture such relationships through a variety of questions; and it is apparent from the data that land (especially with a red book) is widely used as collateral in Vietnam.

In Long An province no less than 99 percent of the total number of loans involved collateral in the form of land with a red book. In Ha Tay, Phu Tho and Quang Nam the corresponding shares were 31, 77 and 63 percent. Thus, land plays not only a significant – but a fundamental – role in determining the operation of the credit market, including who gets access to credit. The opposite statement cannot be made. There is almost no credit-based land acquisition reflected in the data as would be the case in a more developed market economy. Only six loans were granted for buying land during the period studied. This appears credible, partly since there is no evidence in the data that the use of loans was misstated, and partly because of the still underdeveloped nature of land ownership and land transactions in Vietnam.

(c) *Rural credit in 2002*

In this section we look in more detail at loans obtained in 2002. It is the most recent year from which data are available, and they provide the best up-to-date picture of the rural credit market in Vietnam. Table 5 illustrates some subtle differences between loans obtained in the different segments of the loan market. Arguably, the definition of the formal segment is broad (see the list of institutions in Appendix A). Nevertheless, the differences are illuminating.

Table 5. *Characteristics of loans obtained, 2002*

	Formal segment	Informal segment: Private lenders	Informal segment: Friends (zero interest)
Number of loans	250	55	84
Loan amount (Dong)	8,426	3,904	2,602
Duration (average number of months)	15 (N=248)	9 (N=24)	4 (N=11)
– Unspecified duration (percent)	1	56	87
Interest (percent per month)	0.87	1.78	0
Collateral (percent of loans)	71	0	0
Partial default [a] (percent)	8	11	1
Provinces:	*Pct.*	*Pct.*	*Pct.*
– Ha Tay (percent) (N=126)	52	14	35
– Phu Tho (percent) (N=106)	50	21	29
– Quang Nam (percent) (N=24)	77	8	15
– Long An (percent) (N=118)	88	10	2

Distribution of loans by source and province (weighted by loan size)

	VBP	VBARD	Private lenders	Relatives	Union	Others
– Ha Tay (percent) (N=126)	3	32	6	22	14	22
– Phu Tho (percent) (N=106)	4	42	10	27	7	12
– Quang Nam (percent) (N=24)	4	73	2	6	3	11
– Long An (percent) (N=118)	1	76	2	3	1	18

Source: Authors' calculations based on ILSSA Access to Resources Survey 2003.
[a] Partial default is the default rate measured as the percentage of loans where households have defaulted.

The differences in terms of volume and loan size were already evident from Table 3. Loans from the formal sector have an average duration of 15 months. The duration is shorter in the interest carrying informal sector, but with an average of nine months, it is clear that this segment of the loan market is not only used for short term purposes. Borrowing from friends and relatives at zero interest is either for a short period or no specific duration is agreed for the loan. A total of 87 percent of the loans among friends have no formal length specified, suggesting that this kind of loan typically involves lending among family members or close friends. Around half (56 percent) of the interest carrying informal loans from private lenders also have no duration specified. This suggests that some households may be at risk of not generating enough income to enter into specified agreements, including regularly scheduled payments. Studying this group in greater detail would be highly policy relevant from a vulnerability point of view, but is beyond the scope of this paper.

The default rate is the percentage of loans in each segment where households have defaulted, including non-payment of interest or repayment of the principal. The magnitude of the figures is hard to assess. One reason is that the principal is paid in full at the end of the loan term for most formal loans, so only interest payments are made regularly. Paying both interest and principal at the end of the agreed loan period is also quite common. Thus, an eight percent default rate within a period of one year (as shown in Table 5) is substantial if this involves non-payment of interest only. On the other hand, it is not clear from the data whether this payment came forward sometime later or whether the household simply stopped paying instalments on the loan.

Collateral is used for 70 percent of all formal loans whereas no collateral is needed in the informal sector. Land with red book is used as collateral in the majority of the loans. House and land without red book are also used, but to a lower degree, and there are as already alluded to significant regional differences in the use of collateral.

Table 5 confirms that Ha Tay and Phu Tho both have about 50 percent of the loans in the formal segment, whereas Long An and Quang Nam have much higher shares for this sector. In Long An almost 90 percent of the loans originate in the formal sector. This corresponds well with the perception that southern Vietnam (where Long An is situated) is relatively more ‘market-based’ than other regions of the country. Similarly, although households in Quang Nam obtain close to 80 percent of their loans in the formal sector, it is clear that very few households obtain any credit at all, reflecting the very underdeveloped nature of the economy of this province. The bottom of Table 5 provides information on the distribution of loans by different sources. The main difference is between Quang Nam and Long An, on the one hand, and Ha Tay and Phu Tho, on the other.

The above differences suggest that different segments in the loan market serve different needs. In Table 6 this is further explored by tabulating the use of loans in the three credit segments. The formal sector focuses almost entirely on demand for production loans and asset accumulation.[14] A higher share of loans from the informal sector is directed towards health expenditure and consumption. These loans are likely to be due to household shocks or unforeseen events. They carry a higher interest rate than those obtained in the formal sector, showing that households rely on loans from the informal sector to cope with shocks and unforeseen events due to lower transaction costs and more flexible terms of lending. It is also worth noting that more than 50 percent of the interest bearing loans from the informal sector is for production purposes, demonstrating the importance of this loan segment for the growth process of Vietnam.

Table 6. ***Use of loan by credit source (percent), 2002***

		Informal segment		
Use of loan:	Formal segment (N = 250)	Private lenders (N = 55)	Relatives (zero interest) (N = 84)	Total
Production	81	55	36	68
Repayment of other loans	4	9	1	4
Asset accumulation	9	5	23	12
Health expenditure	3	11	12	6
Consumption	3	20	29	11
Total	100	100	100	100

Source: Authors' calculations based on ILSSA Access to Resources Survey 2003.

4. DETERMINANTS OF CREDIT DEMAND

Basic characteristics and differences between the formal and informal credit markets were in focus above. In this section, an econometric framework is applied to identify more rigorously the level and determinants of credit demand at the household level. We restrict ourselves to credit demand in 2002 since this is the most recent year for which data are available and as such provide the most up-to-date picture of credit demand in Vietnam. Moreover, focusing on 2002 allows us to consider the explanatory variables relied on in this section as pre-determined as further discussed below.

In a setup where only matched (i.e. approved) loan applications are observable, the analyst cannot hope to identify correctly the characteristics affecting real credit demand at the household level. However, even with knowledge about rejected loan applications, identification of 'self constrained' households is normally complex and challenging. We are fortunate in the present paper that we have the information required to address these identification problems. Consequently, we are able to categorize households as demanding credit if they (i) obtained a loan, (ii) had a loan rejected or (iii) did not apply even if they wanted credit.

The underlying structural framework for analysing credit demand is a household production model with utility maximizing households, who demand credit (demand = 1) if a loan is expected to increase utility, and they do not demand credit (demand = 0) in the opposite case. If a household demands credit the size of the loan applied for is determined by variables related to the optimal investment if the loan is for investment purposes or the optimal consumption loan if the loan is for consumption. This framework leads to a hurdle model where demand for credit is first characterized by a probit model. Thus,

$$P(demand = 1) = \Phi(h\left(H_i, X_c, D_p\right)) \tag{1}$$

where h is a linear function of the vectors of explanatory variables: H_i is a vector of household characteristics, X_c captures village characteristics and D_p represents province dummies. The expected value of the amount of credit demanded given the household demands credit is described by a lognormal model such that:

$$\left\{\log(\textit{loan amount}) \middle| \textit{demand} = 1, g(H_i, X_c, D_p)\right\} \sim N\left(g(H_i, X_c, D_p), \sigma^2\right) \quad (2)$$

The function $g(H_i, X_c, D_p)$ is a linear form of the same explanatory variables as in the probit model for whether or not to demand credit. The parameters in this stage can be estimated by OLS[15]. From the demand equation (1) and the level equation (2), the expected level of credit demand conditional on explanatory variables is given by:

$$E\left(\textit{loan amount}) \middle| H_i, X_c, D_p\right) = \Phi\left[h\left(H_i, X_c, D_p\right)\right] \exp\left[g\left(H_i, X_c, D_p\right) + \sigma^2 / 2\right] \quad (3)$$

At the household level human capital controls include age and education of the household head, a proxy for the information level (a dummy capturing whether the newspaper ‘People’ is read or not), and productive assets. These are total land holdings, number of adults as a proxy for labour power, and feed expenditure as a proxy for the size of livestock holdings. We also control for the value of total household assets and the need for obtaining credit by including the number of dependents. Furthermore, a proxy is included to capture shocks at the household level in the form of a dummy showing whether a household member was hospitalized within the last 12 months. The gender of the household head is also included, and we control for ‘connectedness’ through the use of a dummy, indicating whether anyone in the household has acquaintances in the existing credit institutions. Credit history is controlled for through the variable ‘not paid’ capturing whether a household has defaulted, i.e. not made a repayment on a loan in full or in part on a loan obtained prior to 2001. Finally, we take account of the influence of security of land tenure by including the share of household land area for which a red book is in hand.

Village level information includes distance to the district centre where VBARD/VBP has an office, and four province dummies capture whether households live in Ha Tay, Phu Tho, Quang Nam or Long An.

In the present analysis data for the following explanatory variables originate from the VHLSS 2002: age, gender, education, adults, dependents, animal feed, total assets, distance, information, and hospitalization. These data were collected about one year before the ILSSA survey. They therefore precede our information on credit demand in 2002 by about one year. This allows us to treat these data as pre-determined. In addition to the provincial dummies, data for the remaining explanatory variables, i.e. total land, connections, credit history and share of land covered by a red book, come from the ILSSA survey. Since land ownership was collected with a time dimension we can use the amount of land owned in 2001, which is exogenous to credit demand in 2002. Connectedness is measured by a dummy variable constructed based on responses to whether anyone in the sampled households has close personal contacts in the existing credit institutions that go beyond a standard customer relationship.

Two sets of summary statistics are given in Table 7. The first two columns show for each variable the number of observations for which data is available in the total sample of 932 households used in Section 3. However, information is missing on distance and total assets for respectively 40 and 15 households (with no overlap). In addition, two households had no land in 2001. Accordingly, the last five columns provide summary statistics for the 875 households used in the empirical analysis, and they will be referred to as the full sample in what follows.[16]

It is clear from Table 7 that the reduction in sample size due to missing observations is not important. Means change very little. The age of the household head ranges from 22 to 93 years, and some 20 percent of households are female headed. In addition, the education variable confirms that household heads have on average more than six years of schooling. Other observations include that while the average land area is small (i.e. around two thirds of a hectare) there are indeed a few households with large landholdings and substantial assets in the form of livestock. Moreover, 19 percent of all households in the full sample had at least one member in hospital during 2002, and 21 percent of households read the newspaper ‘People’. Finally, some 8 percent of households have defaulted on a loan, and 79 percent of the total household land area was registered with a red book.

Table 7. ***Demand for credit: summary statistics, 2002***[a]

	N[b]	Mean	N[c]	Mean	Std. dev.	Min	Max
Demand for credit	932	0.34	875	0.367	0.48	0	1
Age	932	47.74	875	47.61	14.31	22	93
Total land (1,000 m^2)	932	6.33	875	6.49	15.44	0.02	177
Total land squared	932	265.5	875	280.2	1874.4	0.00	31,152
Gender (male=1)	932	0.80	875	0.81	0.40	0	1
Education	932	6.33	875	6.47	3.35	0	12
Adults	932	2.44	875	2.46	1.21	0	10
Dependents	932	1.93	875	1.96	1.18	0	6
Feed (mill. Dong)	932	1.38	875	1.44	4.91	0	80
Ha Tay	932	0.35	875	0.35	0.48	0	1
Phu Tho	932	0.21	875	0.22	0.42	0	1
Quang Nam	932	0.23	875	0.21	0.41	0	1
Long An	932	0.21	875	0.22	0.41	0	1
Total assets (mill. Dong)	917	12.86	875	13.02	20.91	0	370
Total assets squared	917	589.4	875	606.3	4938.0	0	137,122
Distance (km)	892	8.82	875	8.75	8.98	0	40
Information	932	0.22	875	0.21	0.41	0	1
Hospitalization	932	0.20	875	0.19	0.40	0	1
Connections	932	0.52	875	0.52	0.50	0	1
Red book	930	0.78	875	0.79	0.35	0	1
Not Paid	932	0.08	875	0.08	0.27	0	1

Source: Authors' calculations based on ILSSA Access to Resources Survey 2003.
[a] For complete definitions see Appendix B.
[b] Total number of observations available for each variable.
[c] Number of observations used in the empirical analysis. The full sample used contains 875 households due to missing data on distance and total assets for a total of 55 households, and two households had no land in 2001.

We hypothesize that productive capital (land holdings, number of adults and livestock holdings) will affect the propensity to demand credit and the level demanded positively. For example, the greater the landholdings the more likely a farmer is to demand credit to provide access to fertilizer and other inputs. The coefficient on the education of the household head is likewise expected to have a positive sign as greater ability and human capital should affect investment possibilities. Similarly, being better connected, informed and with secure land rights in the form of red books should have a positive impact on credit demand. Finally, many dependents and a person hospitalized in the last 12 months are proxies for a higher probability of the household being in need of credit. They are thus more likely to have a loan demand.

A priori expectations about the signs of the variables capturing the age and sex of the household head and credit history are less clear. A number of different arguments may hold, so these variables are included as controls without well defined priors. The same can be said for the total asset base, which could theoretically affect the probability of obtaining a loan both negatively and positively. A larger asset base would tend to make self financing of loans more viable. On the other hand, it may also improve the loan terms, which the household are offered, making it cheaper to obtain a loan.

It is expected that the distance (village) coefficient is negative. The further away the household lives from the district centre the more costly it is for the household to obtain the loan, due to for example travel costs. This argument will not necessarily hold if the household directs demand towards a local moneylender. Yet, in remote villages local moneylenders are likely to hold more monopoly lending power, demanding stricter repayment conditions (which we do not control for) and thus discourage demand for credit.

Table 8 reports results from estimation of equation (1) and (2) together with marginal effects. As explained previously the four regions where data is sampled from are diverse with respect to geography and economic development. To account for this and to investigate if coefficients differ between regions, variables of central interest were augmented with regional dummy interaction terms in the demand equation.[17] Specifically, land holdings, education, distance to village centre, gender and the share of landholdings with a red book were interacted with regional dummy variables. We estimated this large model on the loan demand equation (results not reported) and retained in all subsequent regressions the interaction terms which were either individually significant or where the joint test of insignificance failed when including that variable. The procedure suggested that land area be augmented with a dummy for Long An province, distance with Quang Nam and Phu Tho dummy variables and possession of red book also with a Quang Nam dummy. The augmented variables are listed in the tables under their 'main' counterpart labelled with the province name for which the variable is augmented.

Table 8. *Determinants of credit demand, 2002*

Dependent variable according to column headings.	Full sample						Reduced sample[d]	
	Probit (demand=1) Marginal effects [a] (1)		OLS[b] Log(amount) if demand=1 (2)		Marginal effects[c] $\frac{\partial E(amount)}{\partial x}$ (3)		Probit (demand=1) Marginal effects [a] (1)	
Age	-0.41***	(0.12)	0.0071	(0.0059)	-12.4	(12.6)	-0.30**	(0.12)
Land	0.65**	(0.33)	-0.0004	(0.0030)	43.7	(46.7)	0.60**	(0.28)
- Long An	-0.66*	(0.34)	0.0140***	(0.0041)	-9.5	(54.6)	-0.63**	(0.30)
Gender (male=1)	-6.46	(4.40)	0.2550	(0.2003)	70.4	(408.3)	-2.74	(5.09)
Education	-0.07	(0.59)	0.0336	(0.0262)	43.6	(60.0)	-0.03	(0.57)
Adults	3.25**	(1.32)	-0.0610	(0.0590)	64.6	(148.1)	3.41***	(1.31)
Dependents	0.81	(1.35)	-0.0038	(0.0672)	34.5	(144.6)	0.56	(1.53)
Feed (mill. Dong)	0.63*	(0.34)	0.0374***	(0.0102)	107.5**	(53.7)	0.64*	(0.34)
Total assets (mill. Dong)	0.10	(0.08)	0.0077**	(0.0031)	20.8**	(9.3)	0.14*	(0.07)
Distance (km)	-0.45	(0.49)	-0.0212	(0.0154)	-65.7	(51.7)	-0.51	(0.54)
- Phu Tho	1.49**	(0.62)	0.0089	(0.0154)	97.8	(68.3)	1.65***	(0.63)
- Quang Nam	-1.48**	(0.72)	0.0184	(0.0474)	-61.9	(240.2)	-1.62	(1.09)
Information	-3.73	(4.43)	0.2715	(0.1903)	344.6	(425.1)	-3.76	(4.72)
Hospitalization	1.83	(4.66)	0.0941	(0.1873)	324.7	(426.3)	0.21	(4.55)
Connections	12.55***	(3.52)	-0.0188	(0.1386)	678.2*	(390.9)	11.75***	(3.56)
Red book	-0.90	(4.84)	0.2155	(0.2704)	397.9	(615.4)	1.04	(5.32)
- Quang Nam	27.47	(19.74)	0.5103	(0.6656)	6097.3	(10,776)	20.10	(19.60)
Not Paid	6.43	(6.01)	-0.1263	(0.2782)	112.0	(706.9)	4.22	(6.08)
Phu Tho	-14.38***	(4.86)	-0.4179**	(0.1974)	-1282.7**	(581.6)	-14.93***	(4.65)
Quang Nam	-31.98***	(12.34)	-0.2901	(0.7274)	-36215.4	(8.8E7)	-21.47	(15.09)
Long An	20.91***	(6.48)	0.6716**	(0.2679)	3187.2*	(1643.2)	27.67***	(7.27)
Constant	..		7.112***	(0.4409)			..	
Test: all coefficients are zero	Wald chi2(21) p-value = 0.0000		F(35,44) p-value = 0.0000		..		Wald chi2(21) p-value = 0.0000	
Goodness of fit	Mcfadden R^2 = 0.13		R^2 = 0.35		..		Mcfadden R^2 = 0.15	
Number of observations (clusters)	875 (46)		293 (45)		875		817 (46)	

Source: Samples from ILSSA Access to Resources Survey 2003 as described in the main text.
Note: Standard errors in parenthesis. Level of significance robust for clustering at the enumeration area throughout. *, **, ***significant at 10, 5 and 1 percent, respectively.
[a] Coefficients on continuous variables measure the marginal effect in percentage points on the probability of demanding credit, whereas they measure the effect of discrete changes for the dummy variables. All marginal effects are evaluated at sample means.
[b] Coefficients (semi-elasticities) from OLS regression on log(loan amount). Only received loans included.
[c] Marginal effects of coefficients on the unconditional expectation of loan amount evaluated at sample means. Robust standard errors obtained by non-parametric bootstrap with a 1000 replications (see Appendix C).
[d] The reduced sample excludes 58 households from the full sample, who obtained a zero interest loan from friends.

Apart from results on our full sample of 875 households, Table 8 shows the demand equation estimated on a sample which is reduced by removing 58 households, who obtained a zero interest loan from friends (column 4). The motives for demanding credit in this situation may differ from the framework set up above, and we wish to uncover whether our results are robust to removing these households.

It emerges from Table 8 that the probit regressions based on the full and the reduced sample are actually quite similar. Magnitude, significance levels and signs are (with one insignificant exception) the same for all variables. Therefore, we focus on the results from the full sample.

The results confirm as expected that land is a statistically significant determinant of credit demand. However, the nature of this impact differs between Long An and the three other provinces. Outside Long An the probability of demanding credit increases with land size but this is not the case for the size of the loan. For Long An province the opposite is true. While the size of land holdings have virtually no impact on the probability of demanding credit, the amount obtained depends significantly on landholdings. However, in economic terms the effects are not large. In Long An, an extra 1,000 m2 of land gives a 1.4 percentage point larger loan, whereas the probability of demanding credit goes up with 0.66 percentage points for an additional 1,000 m2 in the three other provinces. There are as already referred to above many reasons for expecting that land should be significant, and it is reassuring that this is reflected in the data. The connectedness variable is positive, large and strongly significant, which confirms that being connected has clear and positive impact on credit demand. As indicated above no province differences were found for the connectedness variable, and no impact is found on the loan size. This suggests that connectedness works through increased knowledge of opportunities rather than through preferential treatment. The number of adults affects credit demand strongly both in terms of statistical and economic significance. An extra adult in the household increases the probability of demanding credit with more than 3 percentage points, ceteris paribus. Apart from increased investment possibilities more adults also increase the scope for demand for consumption loans. Assets and the proxy for livestock holdings (feed) have small or no effect on the probability of demanding credit, but they affect the credit amount given a loan was obtained positively and significantly. The effects on loan size are small in economic terms. For instance, a doubling of livestock holdings (feed) from its mean level results in a 5 percent increase in loan size given the household obtains a loan. However, this result does confirm that when a household has productive assets (in this case livestock) the demand for credit goes up. The age of the household head is also significant, but the older the household head the less credit is demanded. This in all likelihood reflects that older people in the provinces studied are more settled and less likely to take new and capital demanding initiatives. Cultural values may play a role here as well.

Table 8 reveals very interesting differences in credit demand among the provinces under study. Recalling that Ha Tay is the base, there are large significant differences between Ha Tay and the three regional dummy variables for Phu Tho, Quang Nam and Long An. Controlling for other factors the demand for credit is lower in Phu Tho and Quang Nam than in Ha Tay and Long An (with a significant positive coefficient). Demand is lowest in Quang Nam, although not significantly lower than in Phu Tho, and highest in Long An. The differences have large economic significance as well. For otherwise similar households being located in Long An entails a 50 percent increase in the probability of demanding credit. This is further compounded when taking into account the differences in the amount of credit given a loan is obtained, and the marginal effects on the unconditional (on having a loan) expectation of household credit amount. These observations correspond well with the respective level of development of the provinces studied, and it confirms that credit issues are going to remain key challenges as the transformation of the Vietnamese economy proceeds. Apart from the effect of land holdings as discussed above, regional differences are also present with respect to distance from the village centre. Relative to Ha Tay province greater distance has a positive impact on the probability of demanding credit in Phu Tho. The opposite is the case in Quang Nam. While it is not obvious to see what is driving the result for Phu Tho, the finding for Quang Nam is in line with the prior expectation of this mountainous region. Finally, among the statistically significant variables, it is worth noting the coefficient on the variable 'Red book' – the share of land holding under the red book. For Quang Nam the coefficient is large and positive while for the base (i.e. the other three provinces) it is small and negative.

It is important to keep in mind that pooling demand for formal and informal credit risks blurring the picture of rural credit demand. It is likely that there are differences in the way in which the various households and other characteristics affect formal credit relative to informal credit demand. Distance to the district centre (office of a formal lender) may for example be negatively related to demand for formal credit and positively related to demand for informal credit. It is also sensible to expect that households with a problematic credit history are more likely to demand credit through the informal market. Finally, it is probably also correct that negative shocks like having a household member hospitalized is more directly correlated with informal credit demand. Households may well perceive it as difficult to obtain consumption loans from formal credit sources.

To explore this, Table 9 presents results of probit regressions where formal and informal credit demand is studied separately in a bivariate probit model where non-independence in the error term is allowed for. Thus, using *i* to indicate households,

$$\begin{aligned} z_{1i} &= 1 \quad if \ z_{1i}^{*} = \beta_1 q_{1i} + \varepsilon_{1i} > 0, 0 \ otherwise \qquad (demand\ for\ formal\ credit) \\ z_{2i} &= 1 \quad if \ z_{2i}^{*} = \beta_2 q_{2i} + \varepsilon_{2i} > 0, 0 \ otherwise \qquad (demand\ for\ informal\ credit) \end{aligned} \tag{4}$$

where ε_{1i} and ε_{2i} have mean zero and unit variance (for normalisation), such that formally $(\varepsilon_{1i}, \varepsilon_{2i}) \sim binorm(0,0,1,1,\rho_z)$, and ρ_z is the coefficient of correlation. q_j is a vector of explanatory variables with the first element being one, and β_j a conformable vector of coefficients to be estimated, $j = 1,2$. Our interest is whether factors determining credit demand differ between the formal and informal sectors, thus we ask whether $\beta_1 = \beta_2$. The explanatory variables used here are the same as those relied on in Table 8.

Table 9. ***Determinants of formal and informal credit demand***

Dependent variable according to column headings.	Demand Formal (Full sample) Probit (demand=1) Marginal effects [a] (1)		Demand Formal (Full sample) OLS[b] Log(amount) if demand=1 (2)		Demand Formal (Full sample) Marginal effects[c] $\frac{\partial E(amount)}{\partial x}$ (3)		Demand informal (Full sample) Probit (demand=1) Marginal effects [a] (4)		Demand informal (Full sample) OLS[b] Log(amount) if demand=1 (5)		Demand informal (Full sample) Marginal effects [c] $\frac{\partial E(amount)}{\partial x}$ (6)	
Age	-0.24**	(0.10)	0.014*	(0.008)	0.64	(9.3)	-0.23***	(0.08)	0.004	(0.009)	12.6***	(4.0)
Land	0.32**	(0.14)	0.010	(0.007)	35.7***	(10.2)	0.01	(0.09)	-0.016**	(0.008)	-4.6	(3.8)
- Long An	-0.25	(0.16)	-0.000	(0.007)	-18.7**	(9.3)	-0.03	(0.13)	0.109	(0.083)	14.0	(44.4)
Gender (male=1)	-3.38	(4.01)	0.196	(0.239)	-59.8	(315.0)	-3.11	(2.51)	0.581**	(0.269)	136.9	(143.1)
Education	0.48	(0.49)	0.052**	(0.025)	102.7***	(31.2)	-0.75**	(0.34)	-0.015	(0.040)	-31.5	(20.4)
Adults	2.82**	(1.09)	-0.058	(0.086)	106.0	(98.5)	0.67	(0.74)	-0.178*	(0.104)	-47.5	(54.7)
Dependents	0.09	(1.33)	0.022	(0.080)	31.4	(91.5)	1.54*	(0.79)	-0.042	(0.123)	7.7	(55.3)
Feed (mill. Dong)	0.53**	(0.27)	0.016***	(0.003)	51.5***	(8.7)	0.33**	(0.15)	0.058***	(0.005)	40.0***	(9.7)
Total assets (mill. Dong)	0.17***	(0.06)	0.005**	(0.002)	16.2***	(3.2)	-0.17***	(0.06)	0.026**	(0.012)	5.7	(6.2)
Distance (km)	-0.53	(0.39)	-0.017	(0.012)	-53.7***	(16.0)	0.03	(0.26)	0.006	(0.024)	16.4	(10.4)
- Phu Tho	1.06**	(0.49)	-0.008	(0.016)	52.8***	(20.0)	0.35	(0.33)	-0.016	(0.040)	-2.7	(18.2)
- Quang Nam	-1.03	(1.00)	-0.022	(0.064)	-77.5	(64.4)	-0.39	(0.43)	0.034	(0.033)	-13.6	(14.4)
Information	-1.63	(3.94)	-0.140	(0.250)	-237.8	(340.4)	-2.27	(2.15)	0.628**	(0.258)	139.8	(217.2)
Hospitalization	-0.90	(3.01)	0.250	(0.220)	267.1	(433.0)	4.44	(3.25)	0.192	(0.246)	396.3	(307.5)
Connections	6.96**	(2.70)	0.153	(0.132)	648.3**	(279.7)	6.74***	(2.33)	-0.278	(0.238)	261.7**	(130.7)
Red book	7.63	(4.84)	-0.302	(0.232)	207.8	(283.6)	-5.27*	(3.17)	0.417	(0.455)	39.0	(220.9)
- Quang Nam	11.17	(18.11)	1.102	(0.667)	1817.3**	(710.1)	7.83	(9.34)	-1.897	(3.970)	-573.6	(1878.6)
Not Paid	-0.79	(4.37)	-0.048	(0.392)	-144.9	(541.1)	7.38*	(4.35)	-0.280	(0.543)	-14.2	(250.1)
Phu Tho	-4.50	(4.50)	-0.432	(0.281)	-679.4**	(312.1)	-1.04	(4.21)	-0.529	(0.681)	-184.0	(292.6)
Quang Nam	-8.89	(16.75)	-0.384	(0.492)	-843.8	(981.1)	-15.4***	(4.62)	1.228	(3.355)	-330.4	(885.7)
Long An	29.91***	(8.51)	0.793***	(.275)	4458.4***	(1288.2)	-8.18***	(2.92)	-0.487	(0.730)	-420.5	(316.4)
Constant	..		7.33***	(.430)	..		..		7.12***	(0.659)		
Test: all coefficients are zero	Wald chi2(42) p-value = 0.0000		F(21,20) p-value = 0.0000		..		Wald chi2(38) p-value = 0.0000		F(21,13) p-value = 0.0000		..	
Goodness of fit	Wald test ρ=0, p-value 0.96		$R^2 = 0.37$				Wald test ρ=0, p-value 0.84		$R^2 = 0.34$			
Number of observations (clusters)	875 (46)		192 (41)		875		875 (46)		113 (34)		875	

Source: Samples from ILSSA Access to Resources Survey 2003 as described in the main text.
Note: Standard errors in parenthesis. Robust standard errors and adjustment for clustering at the enumeration area throughout. *, **, ***significant at 10, 5 and 1 percent, respectively
[a] Coefficients on continuous variables measure the marginal effect in percentage points on the probability of demanding credit, whereas they measure the effect of discrete changes for the dummy variables. All marginal effects are evaluated at sample means. Estimated jointly with bivariate normal error term. Estimate of correlation coefficient: ρ=0.03.
[b] Coefficients (semi-elasticities) from OLS regression on log(loan amount).
[c] Marginal effect of coefficients on the unconditional expectation of loan amount evaluated at sample means. Standard errors obtained by the delta method (see Appendix C).

The reported test for independence between the equations shows that the null hypothesis of independence cannot be rejected. Specifying an individual probit regression for each equation yields almost the exact same result (not reported) as the bivariate model.

Analogous to the results from the pooled formal and informal credit markets presented in Table 8, Table 9 shows determinants of logarithmic loan size and marginal effects conditioning on the households obtaining a loan in respectively the formal and informal sector (column 2, 3, 5 and 6).[18]

As regards the distinction between formal credit, on the one hand, and informal credit, on the other, it is clear why some of the insignificant statistical results were obtained in Table 8. Columns 1 and 4 of Table 9 show that counter veiling impacts between the formal and informal credit market segments are involved when it comes to education, dependents, assets, credit history and the red books. They tend to make the overall effect on credit demand in Table 8 insignificant. An additional year of education of the household head significantly reduces the probability of the household demanding credit from informal sources. Also, regarding the formal segment, although education is insignificantly positive as a determinant of credit demand, it increases the size of the loan obtained with around 5 percent given a loan is obtained. In both the formal and informal market a household’s asset base plays a significant role. For the formal market more assets increase the probability of demand credit; the opposite holds in the informal market. This is consistent with productive assets giving more opportunities for investments and therefore increased demand for credit from formal sources. On the other hand, a larger asset base makes borrowing less necessary in the case of negative shocks – hence, a lower probability of borrowing from the informal sector. If a loan is obtained, they tend to be larger in both segments. Arguably, this is due to easier access to collateral when the asset base is larger.

In addition to the observations outlined above two policy relevant differences are apparent between Table 8 and 9. The first relates to credit history (not paid). Recall that this dummy variable takes on the value one if the household has previously defaulted and zero otherwise. Pooling formal and informal credit demand yields a large positive marginal effect of a ‘bad’ credit history, although, it should be kept in mind that the effect is insignificant. Table 9 suggests an explanation for this result. A bad credit history significantly increases the probability of demanding credit from an informal source – and the effect is large in economic terms. For the formal sector the effect is negative, though insignificant. While caution is needed in interpreting this finding, it is consistent with ‘bad’ credit history households being unable to secure loans in the formal sector and therefore address their demand towards the informal sector. The second issue is that of red book coverage of land holdings. A larger share of land with a red book means more secure land rights. This in turn should induce investments in productivity enhancements due to better ability to put up collateral and more secure access to returns from investments (Besley, 1995). In the pooled sample no such effect is evident, except from – insignificantly – the province of Quang Nam. Splitting the formal and informal credit market gives a large positive effect on formal credit demand bordering significance. Demand for informal credit is significantly and negatively affected by red book status suggesting that the red book enables households to obtain loans on better terms in the formal sector than those available in the informal sector.

It is of interest to look further at households, who obtained a loan from both a formal and an informal credit source. In total 29 households received a loan from both segments of the credit market in 2002. Given the limited number of households it is not feasible to make a combined formal analysis (i.e. via a trivariate probit estimation) of demand characteristics. Instead some important statistics is presented in what follows. For the 29 households over half (16) of the loans from the informal segment was from relatives carrying zero interest. The loans from relatives do not differ in the average loan amount compared to loans obtained from relatives by households not having loans from both sources. However, loans taken out from moneylenders charging interest are on average of half the size of the loans taken by other households. This is similar with the formal loans, which are also around half the size compared with the rest of the sample. Regarding the duration, informal loans tend to have lower and formal loans longer duration for households involved in both segments. It is difficult to arrive at one simple explanation consistent with these observed patterns. There is nothing to suggest that 29 households were rationed from formal lending, and therefore had to turn to the informal segment of the credit market. Rather, it would seem that these households rely on formal lenders for longer term (i.e. longer than average) financing and on relatives and other private lenders for short term financing. However, a larger sample would be needed to unravel these explanations.

To sum up, the only variables in Tables 8 and 9 for which little systematic influence on credit demand can be uncovered one way or the other appear to be the information variable and hospitalization, which are admittedly rather crude proxies. Moreover, the data suggest as just alluded to that a key underlying distinction between formal and informal credit demand is that formal demand is particularly driven by factors such as total land and to a lesser extent by red book status. This reflects the need for credit for production and the management of assets whereas the effect of age does not differ. In contrast, informal credit is, in addition to being negatively associated with age and education positively dependent on the credit history (not paid) and on the number of dependents, reflecting household need to smooth consumption and address external shocks. When households have assets they are better able to manage these needs without relying on informal credit as reflected in the coefficient of total assets. Yet, being connected, for example, is statistically important throughout.

Finally, when it comes to provincial differences striking results stand out. In terms of the informal credit market Quang Nam and Long An have significantly less activity. For Long An this is more than compensated for by very high formal market participation relative to the base province of Ha Tay, whereas Quang Nam also has lower activity in the formal market (not significant). The province of Quang Nam is clearly a relatively underdeveloped province (as compared to Ha Tay) in terms of both formal and informal credit demand, whereas Long An stands out as the most developed province. All in all, the statistical results confirm that location specific circumstances (including the general level of development) are critical in understanding credit demand.

5. DETERMINANTS OF CREDIT RATIONING

(a) *Rationing by formal lenders, VBARD*

The Vietnam Bank for Agriculture and Rural Development (VBARD) is as shown in Table 2 by far the largest single lender to rural households in Vietnam, accounting for around one third of the total market in volume and more than half when loan size is accounted for.[19] It is therefore central to rural development that credit is disbursed efficiently by the VBARD. While a complete evaluation of the lending practices of VBARD is outside the scope of the present paper, our data make it possible to identify both the characteristics of households, who obtained credit from VBARD, and the characteristics of households, who had their application turned down. The sample size for those, who got their application rejected, becomes fairly small, so results should be interpreted as indicative only.

Table 10 displays the mean values of the variables examined in Section 4. Total land holdings and total assets are larger for households, who were approved for a loan than for rejected households. However, the difference between the two groups is only statistically significant for total land holdings, likewise for sex and the dummies for Phu Tho and Long An. Households residing in Phu Tho are 'overrepresented' among the rejected households whereas the opposite holds for Long An. If any gender discrimination is present it is a bias against men. Worth noting is also that education and family size are both larger (although not significant) in the rejected group; and loan default rates are clearly important in explaining rejection, at least for other formal lenders and informal lenders. In the province of Quang Nam few households apply for a loan and few households are rejected, in line with the results for credit demand analysed in Section 4.

Table 10. ***Household characteristics for approved and rejected loan applications by lenders***[a]

Variables	VBARD Approved	VBARD Rejected	Other Formal Lenders Approved	Other Formal Lenders Rejected	Informal Lenders Approved	Informal Lenders Rejected	Full Sample
Age	46.44	47.05	46.94	44.33	45.02	47.80	47.61
Total land (1,000 m^2)[b]	13.52	3.66	4.43	2.54	4.52	10.72	6.49
Gender (male=1)[b]	0.85	1.00	0.76	0.67	0.79	0.80	0.81
Education	6.74	7.05	7.07	6.89	6.65	5.77	6.47
Adults	2.79	3.16	2.62	2.44	2.51	2.66	2.46
Dependents	1.96	1.79	1.87	1.67	2.02	2.46	1.96
Feed (mill. Dong)	2.17	1.49	2.16	0.51	1.54	1.44	1.44
Ha Tay	0.25	0.16	0.42	0.67	0.51	0.60	0.35
Phu Tho[b]	0.19	0.58	0.38	0.11	0.35	0.09	0.22
Quang Nam	0.10	0.05	0.10	0.22	0.03	0.03	0.21
Long An[b]	0.45	0.21	0.10	0.00	0.10	0.29	0.22
Total assets (mill. Dong)	19.49	12.15	11.36	6.47	10.98	11.25	13.02
Distance (km)	9.75	12.05	7.52	11.94	9.41	7.09	8.75
Information	0.17	0.21	0.17	0.22	0.19	0.20	0.21
Hospitalization	0.22	0.16	0.18	0.22	0.24	0.26	0.19
Connections	0.60	0.68	0.59	0.44	0.61	0.57	0.52
Red book	0.85	0.83	0.80	0.85	0.74	0.69	0.79
Not Paid	0.09	0.05	0.06	0.33	0.06	0.17	0.08
Number of observations	209	19	124	9	186	35	875

Source: Samples from ILSSA Access to Resources Survey 2003 as described in the main text.
[a] Information for 2001 and 2002 is used, and variable mean values are indicated (see Appendix B for full variable definitions).
[b] Means are statistically (5 percent) different between the two first columns.

Given that VBARD specialises in production lending with relatively large loans compared to the other lending institutions (see Table 6) the findings in Table 10 are sensible. They once again spell out that the regional differences in the credit market are substantial and they illustrate that VBARD is focusing its lending on relatively large land and livestock holders.[20]

(b) *Characteristics of credit rationed households*

Earlier theoretical literature on rural credit markets in developing countries is based on the assumption that all households have a positive demand for credit (see Eswaran and Kotwal, 1989 and Braverman and Stiglitz, 1989). Thus, all households, who have not obtained credit within a given period, are considered credit rationed.[21] Several more recent papers have, however, documented that this assumption may be too restrictive in empirical analysis, see Kochar (1997).

In this section we pursue this theme and identify factors at the household level, which influence the probability that a household with given characteristics is credit constrained. It would have been interesting to study the formal and informal sectors separately, but the number of households, who had loan applications rejected, is as already mentioned quite low.[22] Nevertheless, the characteristics which influence credit rationing are likely to be at least similar in the formal and informal segments making it worthwhile to pursue the issue in the aggregated sample. Similarly, because of the sample composition, it is not feasible to augment variables with province level dummies. While this is a drawback, interesting results still emerge from the analysis.

Importantly, a household is defined as being credit rationed if it has *both* applied for a loan (in either the formal or the informal credit market) *and* had the application rejected.[23] In this setting the methodology differs from the one used in the section on credit demand. From household responses it can be established whether a household demands credit. However, for those households, who did not apply for credit, it is impossible to observe what the lender's decision would have been had those households actually applied. This sample selectivity issue is addressed by specifying a bivariate variant of Heckman's selection model (Wooldridge 2002) as follows:

$$\begin{aligned} y_{1i} &= 1 \quad if \; y_{1i}^* = \delta_1 x_{1i} + u_{1i} > 0, 0 \; otherwise \qquad (rationed) \\ y_{2i} &= 1 \quad if \; y_{2i}^* = \delta_2 x_{2i} + u_{2i} > 0, 0 \; otherwise \qquad (applied) \end{aligned} \tag{5}$$

Error terms are assumed to be bivariate normally distributed with zero mean, unit variance and correlation ρ_u. Thus $(u_{1i}, u_{2i}) \sim binorm(0,0,1,1,\rho_u)$ and y_{1i} (i.e. a loan is approved or rejected) is observed only when $y_{2i} > 0$. The vectors of explanatory variables, x_{1i}, x_{2i}, have one as their first element. The second equation is our selection equation determining characteristics, which influence the household decision to apply for a loan ($y_{2i} = 1$). Results from Section 4 are used in specifying this selection equation.[24]

Given that a household applies for credit ($y_{2i} = 1$), the outcome of the application process can be observed from the equation $1 - y_{1i} = 1$ if the household were awarded the loan and zero in the case of rejection. Characteristics at the household, commune and province level are aggregated together in respectively x_{1i} and x_{2i} to ease notation. This simultaneous approach allows us to try to identify determinants of credit rationing taking into account the possible selection bias in households applying for credit. Testing for independence between the two equations is equivalent to testing the hypothesis that ρ_u equals zero.

Table 11 displays the results from four different specifications of the equation determining the probability of a household being rationed. The first column (base applied) shows the coefficients (not marginal effects) from the selection equation, including all of the variables used in Section 4. The same selection equation is used for all four specifications. Only results from the selection equation for the first specification are reported. Due to the simultaneous nature it differs slightly across specification. Although not completely comparable – because of the difference in specification described above and since demanding and having applied for a loan differs in some circumstances, it is instructive to compare the results from the selection equation with those in Table 8. For the significant variables the results from the selection equation conform well with the demand equation in Table 8 – adding further robustness to the results. Thus, numbers of adults, livestock and being connected increase the probability of having applied for a loan. Also, residing in Quang Nam lowers the propensity to apply for a loan substantially whereas the opposite is true for Long An.

Table 11. *Credit rationing, 2002*

Variables	Base Applied[a]	1. Rationed Base [b]	2. Age, gender [b]	3. Distance, information [b]	4. Connections [b]
Age	-.0117*** (0.0030)		0.024 (0.044)		
Total land (1,000 m^2)	0.0030 (0.0039)	0.022 (0.024)	0.021 (0.025)	0.022 (0.025)	0.029 (0.023)
Gender (male=1)	-0.2081* (0.1076)		1.379* (.0754)		
Education	-0.0004 (0.0162)	-0.363*** (0.112)	-0.390** (0.171)	-0.375*** (0.112)	-0.331*** (0.095)
Adults	0.0951** (0.0387)				
Dependents	0.0012 (0.0367)				
Feed (mill. Dong)	0.0188** (0.0078)	0.054 (0.056)	0.050 (0.055)	0.057 (0.054)	0.052 (0.057)
Total assets (mill. Dong)	0.0023 (0.0026)	-0.019 (0.034)	-0.021 (0.032)	-0.020 (0.034)	-0.016 (0.032)
Distance (km)	0.0002 (0.0097)			-0.015 (0.082)	
Information	-0.1726 (0.1188)			0.513 (1.225)	
Hospitalization	0.1187 (0.1312)				
Connections	0.3686*** (0.0947)				-1.554* (0.892)
Red book	0.1341 (0.1360)	-1.607 (1.295)	-1.731 (1.223)	-1.598 (1.278)	-1.697 (1.263)
Not paid	0.1674 (0.1476)	6.206* (3.586)	6.571* (3.739)	6.401* (3.740)	6.767* (3.870)
Phu Tho	-0.0137 (0.1599)	-1.641* (0.945)	-1.648* (0.889)	-1.488 (1.015)	-1.585* (0.945)
Quang Nam	-0.7387*** (0.1807)	-2.967*** (0.806)	-2.987*** (0.775)	-2.959*** (0.828)	-2.940*** (0.782)
Long An	0.3043* (0.1650)	-1.472* (0.826)	-1.510** (0.771)	-1.366 (0.882)	-1.583** (0.791)
Constant	-0.2189 (0.2540)				
Test: all coefficients are zero		F(9,37), p-value = 0.006	F(11,35), p-value = 0.000	F(11,35), p-value = 0.017	F(10,36), p-value = 0.004
Test: Independence of equations		Wald test ρ=0, p-value 0.14	Wald test ρ=0, p-value 0.07	Wald test ρ=0, p-value 0.06	Wald test ρ=0, p-value 0.38
Number of observations / uncensored / clusters	875 / 311/ 46	875 / 311/ 46	875 / 311/ 46	875 / 311/ 46	875 / 311/ 46

Source: Samples from ILSSA Access to Resources Survey 2003 as described in the main text.
Note: Standard errors in parenthesis. Robust standard errors and adjustment for clustering at the enumeration area throughout. *, **, *** significant at 10, 5 and 1 percent, respectively.
[a] Coefficients from the selection equation estimated jointly with 'Rationed Base'. The selection results from the other specification differ only marginally due to the simultaneous structure and are not reported.
[b] Marginal effects in percent.

Our base specification of rationing is shown in column two (specification 1). We include only variables, which are believed to affect the borrower's ability to pay back the loan, and which are (at least in principle) observable to the lender, together with provincial dummies. Thus, we include land and assets, education,[25] feed expenditures as a proxy for livestock holdings, credit history and the share of land for which the borrower has a red book. This last variable is a proxy of the borrower's prospective for entering the land market to secure repayment of the loan. Arguably, the number of adults might also be a useful indicator of repayment ability. We do not include it in the rationed base since the lender is in effect unable to monitor the effort to repay. It might be possible for the lender to force sale of land in case of default, but not to force people to get an income-generating job. Including adults bring no qualitative changes to the result (not reported).

A bad credit history and education are significant with the expected signs. Also the coefficients for assets and the share of land with a red book have expected signs, although they fail conventional significance tests. The larger the share of land for which the household has a red book the lower the probability of being rejected credit. The sign of the coefficient on the land variable is contrary to prior expectations, but insignificant. The provincial dummies reveal significant differences in rationing given our controls among the four provinces studied here. Thus, the probability of having a loan application approved is, once we control for the propensity to apply for credit, statistically different among the various provinces considered here. Relative to Ha Tay, the households in the provinces of Phu Tho, Quang Nam and Long An all have lower probabilities of having a loan rejected. The differences within these three provinces are not significant. The statistically significant results also carry economic significance. For instance, since in the sample around 9 percent of loan applications are rejected, a 6 percent increase in the probability of being rationed as a result of 'bad' credit history constitutes a large relative increase in the risk of being rejected. To a lesser extent the same can be said about the differences between Ha Tay and the other provinces.

Apart from land holdings, the only variable, which does not conform to our prior, is the proxy for livestock. A lender should be more willing to lend if the borrower has livestock which can be sold in case of default. In contrast, the coefficient on feed is positive, suggesting a greater possibility of being rejected, but the coefficient is insignificant.

Finally, the hypothesis of all coefficients (excluding the constant) being equal to zero in the rationing equation is rejected at less than 1 percent, and it appears that the selection framework is in the present case not strictly necessary as the independence of equations cannot be rejected.

In specifications 2, 3 and 4, we augment the rationed base regression with other variables from Table 8, but which should not in theory affect lender decisions given the information contained in the variables from the base regression. In column three (specification 2), we include age and gender of the head of household. It is evidently of interest to uncover whether systematic biases against women are present in the process of reviewing credit applications. We find no such bias here. Keeping in mind that the gender variable has woman household head as its base, the data suggest that women who apply for credit are in fact more likely than men to be approved for a loan. Again, note that the size of the marginal effect is not trivial. This result is statistically significant at the 10 percent level, whereas the corresponding age parameter is clearly insignificant. The gender result must be interpreted with caution. The nature of this issue is complex, and we recall that we do not have individual level information on loan allocations, only at household level. So robustness and channels of influence is an issue for further study. However, our result does correspond with observations made in studies of the allocation of firm credit in Vietnam (see World Bank, 2005). With respect to the other base line variables, signs, magnitudes and significance levels are virtually unchanged for all variables. The test of independence of equations is rejected at the 10 percent level.

The third specification (in column 4) looks at the effect of distance to a district centre (distance) and a proxy for the household information level (information). We offer no prior expectations about the sign of the distance coefficient; but outreach is of particular concern, so insights on the importance of this variable is potentially important information in assessing how credit should be expanded in rural Vietnam. The rationed baseline variables remain virtually unchanged in terms of signs and magnitudes, except for the coefficients of provincial dummy variables Phu Tho and Long An which become marginally insignificant. In fact, specification 3 changes very little, and while distance has a negative and information a positive parameter, they are clearly statistically insignificant. Information has very little explanatory power with a t-value of 0.43, and the t-value of distance is not much higher at 0.18.

Finally, in the fourth specification we try to isolate the effect of being well connected (with respect to contacts in credit institutions). This is done by introducing a dummy variable (connections) equal to one if the household has contacts in any credit organisation. The estimated coefficient is negative and significant at 10 percent, which corresponds to stating that being well connected within credit institutions promotes the application process. Relative to the base specification, the coefficients are very robust to the inclusion of the connectedness variable. In this last specification the test of independence of equations cannot be rejected.[26]

Looking at the four sets of simultaneous regressions overall it is evident that the signs of the coefficients in the base regression are very robust. Households with older heads are less likely to apply for credit. All else equal, elder households are less likely to undertake risks (i.e. apply for loan where repayment is uncertain), but when they apply they neither gain particular preferential treatment nor are they rationed. There is some evidence that males and females are treated differently in the application process, but we interpret this result with caution as indicated above. It seems likely that better educated households are more likely to know when an application will be rejected and the data strongly suggest that once they have applied they are not being discriminated, quite the contrary. The better educated the household head, the better the probability of approval.

Feed, i.e. the measure of assets in the form of livestock, has the expected positive sign in the selection, but plays no role in rationing. This is slightly surprising since if a household decides to apply for a loan then – everything else equal – the ability to repay measured in terms of assets, which can be transferred to the lender should be negatively related to the probability of being rationed. Yet, we also note that the relevant parameters in the rationing equations are statistically insignificant.

Furthermore, as one would expect, the indicator for a bad credit history (not paid), which indicates that a household have defaulted on a loan instalment previously, is positively related to being credit rationed. Yet, it does not appear to deter household from applying in any statistically significant manner, although as noted in the previous section, demand for credit shifts from formal to informal lenders. While clearly important to rural credit, overall, the possession of a red book is not significant when it comes to the decision to apply, but there is some indication that those households who have a red book are less likely to be rationed. The variable controlling for connections has the expected sign in both the selection and rationing equations, but it is only significant at 10 percent level in the rationing equation. The household information level might be said to have the 'wrong' signs in both selection and rationing. We offer no convincing story for this result but note that this is statistically insignificant. The same goes for the distance parameter, though it should be kept in mind that the regressions in this section are pooled over formal and informal lending institutions. Turning to the province dummies, it is clear that provincial differences play a role as all three dummy variables are statistically significant in the majority of specifications. However, in case of rationing it seems that only Ha Tay differs significantly from the three other provinces.

In general the sign of most coefficients as analysed in this section are in line with what we expected a priori. We acknowledge that there are a few exceptions and that several variables lack statistical significance. However, we believe this is more likely to be a feature of the data not having enough variability in central variables. Given the regional differences pointed out above it is also likely that the dummy variables capture a bit too much of the differences in the data. Ideally and with unlimited data, interacting the dummies (as done in Section 4) with core variables to detect province specific effects would be desirable. This is left for future research when better data become available.

6. CONCLUSIONS

Little is known about the characteristics and the operation of the rural credit market in Vietnam. This paper was written with the aim of helping to fill this gap based on a new data set covering 932 households in four provinces (Ha Tay, Phu Tho, Quang Nam and Long An) surveyed in early 2003. In the formal analysis this data was complemented with information available in the 2002 Vietnam Household Living Standard Survey (VHLSS). A number of general observations emerge, which deserve close attention in efforts to further develop the existing credit system.

An active and growing rural credit market exists in Vietnam, and formal credit is clearly expanding its share of total credit. This is in line with the general rapid development of the economy, and overall interest rates have also fallen suggesting that market integration is in fact progressing. In parallel, a sizeable informal sector remains in existence, accounting for about one-third of all loans, and reflecting that poor rural households continue to rely on informal networks and relatives. Different segments in the loan market serve different needs, and we note that the formal sector focuses almost entirely on production loans and asset accumulation. In contrast, both the descriptive statistics and the formal analysis in this paper demonstrate that households actually demand loans for other purposes, such as consumption smoothening and health expenditures. Such loans are often obtained in response to temporary shocks (i.e. having a person hospitalized) and thus work as a consumption smoothing device.

Because of the limited formal lending for consumption smoothening, households direct this demand for credit at private money lenders. This may well be welfare enhancing if the money lenders offer alternatives preferred by the borrower. Yet, to the extent that the borrower can provide collateral (i.e. in the form of land) it should in theory make no difference to formal lenders whether a loan is used for production purposes or for temporary consumption smoothing. If the formal sector entered the market for non-production loans (on financially sustainable terms) this would provide borrowers with an alternative to private money lenders. This could well be welfare increasing, especially for marginalized low-income households. They have limited connections, and this characteristic is as shown in Sections 4 and 5 a strongly constraining factor for credit demand in both the formal and the informal sector. In the informal sector it is moreover typical that older and better educated households have less credit demand. In contrast, a larger number of dependents and a bad credit history tend to increase a household's informal credit demand. This does not necessarily reflect market failure, but it does suggest there is need and space for careful, well designed public action in expanding credit facilities. The social returns of such action may well be high.

Another key characteristic of the rural credit market in Vietnam is the one-way interaction with the land market. Land (with a red book) is widely used as collateral and plays a fundamental role in the operation of the credit market. Land is a statistically significant determinant of overall credit demand. This result is as shown in Section 4 driven by formal credit demand geared towards production purposes and asset management. This further reinforces the above conclusion about the need for carefully metered public action; but it is in parallel striking that there is almost no credit-based land acquisition in rural areas. This highlights the very considerable challenges, which remain to be addressed in establishing the necessary market based institutional framework for a more efficient functioning of the economy.

It comes as no surprise that land is widely used as collateral. Land is immobile and its quality cannot be changed at short notice. Yet, an active land market depends critically on a well functioning credit system for land transactions. The lack of such a market is due to both supply (i.e. lending institutions do not generally finance land transactions) and to the demand side. Accordingly, the land section of the present household survey reveals that the land market among non-relatives is very thin indeed. However, productivity increases in rural agriculture will depend crucially on land consolidation and development in the years to come. The demand for loans to finance land transactions may appear small at present, but formal lending institutions should actively prepare for a more active role in this market. This will as well require that complementary institutions are put in place with the capacity to value land, and also an effective legal system to solve potential land disputes will be required.

The most striking and cross-cutting general insight emerging from this paper is the extent of regional differences in almost all aspects of the credit market. Some broad national generalizations are as already discussed possible. At the same time, it is in designing public policy indispensable to be very careful about the region, the household group and the market segment in question. The formal sector accounts for around 50 percent of loans in Ha Tay and Phu Tho. Long An and Quang Nam have much higher shares, but this characteristic is a reflection of very different levels of development in these two provinces. Few households in Quang Nam obtain credit, and credit demand in this province is clearly limited compared to the other provinces in our sample. This is so both overall and in the various market segments. Pooling demand for formal and informal credit risks blurring the picture of rural credit demand. Counter veiling effects are at work between the formal and informal credit segments when it comes to education, distance, credit history and also the provincial dummy effects differ.

In sum, the econometric analysis confirms that specificity and the general level of development are fundamental in understanding credit demand in Vietnam. A ‘one size fits all’ approach to expanding credit is not going to be the most effective. This dimension therefore needs to be kept in mind in the planned expansion of rural credit through the Vietnam Bank for Social Policies. The VBSP aims at operating a large number of new branches throughout Vietnam (World Bank, 2003). An additional observation in this regard is that expansion needs to be carefully metered to take account of the need for credit in areas where access is presently low – such as in Quang Nam. In Ha Tay and Phu Tho the informal sector is sizeable and as such compensate for an insufficiently developed formal sector, whereas the formal market is already much better developed in Long An. It is in this context also to be noted as shown in Section 5 that VBARD is focusing its lending on relatively large land and livestock holders. We stress that regional differences in credit rationing seems to be limited, although there are small differences showing up once selection is accounted for. In Quang Nam few households apply for credit and few are rejected. On the other hand, the analysis in Section 5 reveals that households with a bad credit history are more likely to get rationed. This merits attention as these households in all likelihood include those households, who are subject to shocks and who find it difficult to manage their lives. It would, given the regional differences pointed out above, be desirable to interact the provincial dummy variables with a larger number of core variables to detect province specific effects. Yet, this is left for future research when better data become available, and the same goes for the challenge of establishing the degree of credit rationing which households experience.

NOTES

[1] See for example Kovsted et al. (2004).
[2] Diagne, Zeller and Sharma (2000) provide a series of other references.
[3] See Duong and Izumida (2002) and McCarty (2001) for earlier work on rural credit and microfinance issues in Vietnam.
[4] For documentation and the questionnaire used see Barslund et al. (2004).
[5] Some 28 households interviewed during the VHLSS could not be interviewed and had to be excluded in the ILSSA survey.
[6] The following website http://www.worldbank.org.vn/data/household_survey.htm provides access to a complete description of the 2002 VHLSS and the questionnaire.
[7] Retrospective questions always entail a risk of imprecise or erroneous answers. However, obtaining a loan is not an 'every year' event and as such is more likely to be remembered correctly than more recurring events. Furthermore, taking out a loan often coincides with 'big' events such as major shocks or purchases, which are likely to be recalled correctly.
[8] The credit market section of the ILSSA survey is Module B (questions 168-224), with three sections: B1 for loans actually received, B2 for loans not received, and B3 on general questions.
[9] The VBP has recently been renamed the Vietnam Bank for Social Policies (VBSP). BARD and VBP are associated in the sense that they often share office facilities. See World Bank (2003) and Kovsted et al. (2004) for a more detailed description of the institutional set up.
[10] Private Trader was also a category in the questionnaire. It turned out that this group does not play an important role in the credit market in the four provinces studied.
[11] See appendix A for the full list of institutions included in the questionnaire.
[12] If loans for primary consumption are only obtainable from informal sources and there is a general increase in incomes, which makes consumption loans less needed, a change in the composition of loans may be expected. Similarly if it is easier to obtain loans for specific purposes such as production, rather than for consumption smoothening or health purposes.
[13] The questions were, respectively: "What was the stated purpose of the loan (select one for each loan)?" and "What did your household mainly use the loan for (select one for each loan)?"
[14] This includes buying/building a house, the few instances of buying land and re-lending and buying other assets.
[15] Only households obtaining a loan were used in this stage, since loan amounts are not available for rejected and self-constrained households.
[16] All regressions were also carried out on a sample excluding outliers, defined as observations, situated outside an interval of three standard deviations from the mean. All qualitative results remained unchanged. Full tables are available on request.
[17] We are not able to specify a fully unconstrained model (i.e. with regional interaction terms on all variables). Our data are sampled in clusters (46 different clusters/enumeration areas) and, thus, have less degrees of freedom in our estimation procedures than with an 'unclustered' approach. The advantage is that the significance of our statistical results are robust to observations being independent between but not within clusters. Assuming independence of all observations strengthens our results considerably.
[18] Self-rationed households did not indicate in which sector they would have applied if they had applied for a loan. Thus, self-rationed households were treated as not demanding credit in the sector specific analysis.
[19] The second most important state bank, Vietnam Bank for the Poor (VBP) has recently been reorganised and is now operating under the umbrella of the Vietnam Bank for Social Policies (VBSP), which is scheduled for a large expansion in the years to come (World Bank, 2003).
[20] While a general characteristic, this effect does to some extent reflect higher BARD lending activity in Long An, which also tends to have larger farms.
[21] In what follows, the terms credit constrained and credit rationed are used interchangeably.
[22] In 2002, 25 households in the sample of the 875 had their loan application rejected by a lending institution (formal and informal). For the sample of 932 households the number was 29 households.
[23] In fact a household may be approved for a loan smaller than it applied for. These households are also to some extent credit rationed. We asked questions about amount obtained, amount wanted and amount applied for to identify households rationed in the loan amount. In our sample 21 households reported (credibly) that they were rationed in the amount they obtained in 2002. For simplicity these households

are considered not rationed in the present study. The qualitative results hold if we include the 21 household (except three households which were rationed in large loan amounts) as rationed.

[24] It is recalled that the definitions for households demanding credit and applying for credit differ as described above.

[25] See Nga Nguyet Nguyen (2004) who reports significantly increasing returns to schooling in recent years.

[26] A fifth specification with the remaining three variables from Table 8, i.e. including Adults, Dependents, and Hospitalization was also carried out. This changed none of the key results discussed, and provided no further insights except that the number of dependents is potentially important. This specification is therefore left out here, but results are available on request.

REFERENCES

Barslund, M., Tarp, F., Nguyen Huu Dzung, and Dao Quang Vinh, (2004). Documentation of the Vietnam Access to Resources Household Survey (VARHS) 2002. Available at http://www.econ.ku.dk/ftarp/ and on request from the authors.

Besley, T., (1995). Property Rights and Investment Incentives: Theory and Evidence from Ghana. *Journal of Political Economy*, Vol. 103, pp. 903-937.

Braverman, A., Stiglitz, J.E., (1989). Credit Rationing, Tenancy, Productivity, and the Dynamics of Inequality. In Bardhan, P. (ed.), *The Economic Theory of Agrarian Institutions*. Oxford University Press., New York.

Cameron, A.C. & Trivedi, P.K., (2005). *Microeconometrics – Methods and Applications.* Cambridge University Press, New York, USA.

Chaudhuri, S., (2001). Interaction of Formal and Informal Credit Markets in Backward Agriculture: A Theoretical Analysis. *Indian Economic Review*, Vol. 36(2), pp. 411-428.

Diagne, A., (1999). Determinants of Household Access to and participation in Formal and Informal Credit Markets in Malawi. Food Consumption and Nutrition Division (FCND) Discussion Paper 67, IFPRI, Washington.

Diagne, Al., Zeller, M. and Sharma, M., (2000) Empirical Measurement of Households' Access to Credit and Credit Constraints in Developing Countries: Methodological Issues and Evidence. Food Consumption and Nutrition Division (FCND) Discussion Paper 90, IFPRI, Washington.

Duong, P. B. and Izumida, Y., (2002). Rural Development Finance in Vietnam: A Microeconometric Analysis of Household Surveys. *World Development*, Vol. 30(2), pp. 319-335.

Eswaran, M. and Kotwal, A., (1989). Credit and Agrarian Class Structure. In Bardhan, P. (ed.), *The Economic Theory of Agrarian Institutions*. Oxford University Press, New York.

Gupta, M. R. and Chaudhuri, S., (1997). Formal Credit, Corruption and the Informal Credit Market in Agriculture: A Theoretical Analysis. *Economica*, Vol. 64, pp. 331-343.

Hoff, K. and Stiglitz, J. (1993). Imperfect Information and Rural Credit Markets: Puzzles and Policy Perspectives, pp. 33-52 in K. Hoff, A. Braverman and J. Stiglitz (eds.) *The Economics of Rural Organization: Theory, Practice and Policy*, Oxford University Press, Oxford.

Kochar, A., (1997). An Empirical Investigation of Rationing Constraints in Rural Credit Markets in India. *Journal of Development Economics*, Vol. 53, pp. 339-371.

Kovsted, J., Rand, J. and Tarp, F., (2004). *From Monobank to Commercial Banking: Financial Sector Reforms in Vietnam.* NIAS Press, Copenhagen.

McCarty, A., (2001). Microfinance in Vietnam: A Survey of Schemes and Issues, Survey commissioned by the British Department of International Development (DFID) and the State Bank of Vietnam (SBV). Available from: http://ideas.repec.org/p/wpa/wuwpfi/0110001.html

Mekong Economics, (2004). Access to Resources. A study available at http://www.mekongeconomics.com/publications/Access%20to%20Resources%20.doc

Mohielsin, M. S. and Wright, P., (2000). Formal and Informal Credit Markets in Egypt. Economic Development and Cultural Change, Vol. 48(3), pp. 657-670.

Nguyen N.N., (2004). Trends in the Education Sector. Chapter 12 (pp. 425-465) in Glewwe, P., Agrawal, N. and Dollar**,** D. (eds) *Economic Growth, Poverty, and Household Welfare in Vietnam*. World Bank, Washington.

Petrick, M., (2005). Empirical Measurement of Credit Rationing in Agriculture: A Methodological Survey. *Agricultural Economics*, Vol. 33(2), pp. 191-203.

Platteau, J-P., (1997). Mutual Insurance as an Elusive Concept in Traditional Rural Communities. *Journal of Development Studies*, Vol. 33(6), pp. 764-796.

Yadav, S., Otsuka, K., and David, C.C., (1992). Segmentation in Rural Financial Markets: The Case of Nepal. *World Development* Vol 20(3), pp. 423-436.

Zeller, M., (1994). Determinants of Credit Rationing: A Study of Informal Lenders and Formal Credit Groups in Madagascar. *World Development*, Vol. 22(12), pp. 1895-1907.

Von Pischke, J.D. and Adams, D.W., (1980). Fungibility and the Design and Evaluation of Agricultural Credit Projects. American Journal of Agricultural Economics, Vol. 62(4), pp. 719-726.

World Bank (2003). Vietnam Development Report 2004. Hanoi.

World Bank (2005). Vietnam Development Report 2006. Hanoi.

Wooldridge, J.M., (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT press, Massachusetts, US.

APPENDIX A

Lending institutions in the questionnaire:

Bank for the Poor (includes National Poverty Alleviation Program)
Bank for Agriculture and Rural Development
Other State-Owned Bank
National Employment Generation Program
Other National Government Program
Other (non National) Poverty Alleviation Program
Private Bank
Farmers' Union
Veterans' Union
Women's Union
People's Credit Funds
Other Credit Associations
Private Trader
State Owned Enterprise (SOE)

International Organisation
Private Money Lender
Friends/Relatives
Other (please specify)

APPENDIX B

List of variables

Name in tables	Definition	Source
Demand for credit	Dummy variable equal to 1 if household demanded credit in 2002	ILSSA 2002
Age	Age of household head in years	VHLSS 2002
Total land	Total landholdings in 1,000 m^2	ILSSA 2002
Gender	Dummy variable equal to 1 if the household is male and equal to 0 if household head is female	VHLSS 2002
Education	Education of household head, number of years of schooling	VHLSS 2002
Adults	Number of adults defined as household members between 15 and 65 years of age and not studying	VHLSS 2002
Dependents	Number of dependents are full time students and household members aged less than 15 or above 65 years	VHLSS 2002
Feed	Expenditures on livestock feed during last 12 months in mill. Dong	VHLSS 2002
Province dummies	Ha Tay, Phu Tho, Quang Nam, Long An.	ILSSA 2002
Total assets	Total value of assets in mill. Dong	VHLSS 2002
Distance	Distance to district centre in km	VHLSS 2002
Hospitalization	Dummy variable equal to 1 if at least one household member hospitalized within the last 12 months and equal to 0 is no member hospitalized	VHLSS 2002
Connections	Dummy variable equal to 1 if anyone in the household has contacts in the existing credit institutions	ILSSA 2002
Red book	The share of household land area for which a red book is in hand	ILSSA 2002
Information	Dummy variable equal to 1 if the household reads the newspaper People	VHLSS 2002
Alternative information	Index where having a radio counts 0.5 and a television 1	VHLSS 2002
Got help	Dummy equal one if the household at some point prior to 2001 got help from the authorities to apply for a loan	ILSSA 2002
Not paid	Dummy equal one if the household did at some point prior to 2001 not pay a loan instalment in full	ILSSA 2002

APPENDIX C

The hurdle model

Consider the hurdle model presented in the main text:

$$P(demand = 1 \mid x) = \Phi\left(x\gamma\right)$$
$$\log(loan\ size) \mid (x, demand = 1) \sim N\left(x\beta, \sigma^2\right)$$

where *demand* is a binary variable indicating if the household demanded credit (or applied for credit depending on the context) and zero otherwise. The variable *Loan_size* is assumed to follow a log-normal distribution with the expectation conditional on *demand=1* given by:

$$E\left(loan\ size \mid x, demand = 1\right) = \exp\left(x\beta + \sigma^2/2\right)$$

The unconditional (on demand) expectation of loan size given the set of explanatory variables, x, is given by:

$$\begin{aligned} E\left(loan\ size \mid x\right) &= P(demand = 1 \mid x) \cdot E\left(loan\ size \mid x, demand = 1\right) + \left(demand = 0 \mid x\right) \cdot 0 \\ &= \Phi\left(x\gamma\right) \cdot \exp\left(x\beta + \sigma^2/2\right) \end{aligned}$$

Marginal effects of continous and discrete multi value variables are obtained by differentiation

$$\begin{aligned} \frac{\partial E\left(loan\ size \mid x\right)}{\partial x_i} &= \phi\left(x\gamma\right)\gamma_i \cdot \exp\left(x\beta + \sigma^2/2\right) + \Phi\left(x\gamma\right) \cdot \exp\left(x\beta + \sigma^2/2\right)\beta_i \\ &= \exp\left(x\beta + \sigma^2/2\right) \cdot \left(\phi\left(x\gamma\right)\gamma_i + \Phi\left(x\gamma\right)\beta_i\right) \end{aligned}$$

where explanatory variables are evaluated at sample means. For discrete (0/1) dummy variables the marginal effects are evaluated using the expression

$$\frac{\partial E\left(loan\ size \mid x\right)}{\partial x_j} = E\left(loan\ size \mid x^{-j}, x_j = 1\right) - E\left(loan\ size \mid x^{-j}, x_j = 0\right)$$

where x^{-j} denotes the set of mean values of explanatory variables excluding x_j.

Standard errors

The delta method

To save on notation, let the full parameter vector $\left\{\gamma, \beta, \sigma^2\right\}$ be represented by the $1 \times r$ vector θ and denote by $m_i\left(\theta, x\right)$ the marginal effect of x_i evaluated at the mean values of x. Thus, $m_i\left(\theta, x\right) = \frac{\partial E(loan\ size|x)}{\partial x_i}$. Let V_θ be the estimated $r \times r$ asymptotic variance matrix of θ from respectively the probit and OLS regressions. The asymptotic variance matrix of $m_i(\theta, x)$, $V_{m_i(\theta,x)}$, is then given by (Wooldrigde, 2002)

$$V_{m_i(\theta,x)} = g_i V_\theta g_i^{'} \text{ where } g_i = \left(\frac{\partial m_i(\theta, x)}{\partial \theta_1}, \frac{\partial m_i(\theta, x)}{\partial \theta_2} ..., \frac{\partial m_i(\theta, x)}{\partial \theta_r} \right)$$

The $1 \times r$ vector g_i is composed of derivatives of the elements of θ.

In terms of the hurdle model the derivatives of the marginal effects of the continuous variables are derived as

$$\begin{aligned}
\frac{\partial m_i^c(\theta, x)}{\partial \beta_j} &= x_j \exp\left(x\beta + \sigma^2/2\right) \left[\phi(x\gamma)\gamma_i + \beta_i \Phi(x\gamma)\right] + \exp\left(x\beta + \sigma^2/2\right) \Phi(x\gamma) I[i = j] \\
\frac{\partial m_i^c(\theta, x)}{\partial \gamma_j} &= \exp\left(x\beta + \sigma^2/2\right) \left[x_j \beta_i \phi(x\gamma) - x_j \gamma_i \phi(x\gamma) + \phi(x\gamma) I[i = j]\right] \\
\frac{\partial m_i^c(\theta, x)}{\partial \sigma^2} &= 1/2 \cdot \exp\left(x\beta + \sigma^2/2\right) \left[\phi(x\gamma)\gamma_i + \beta_i \Phi(x\gamma)\right]
\end{aligned}$$

$I[.]$ is an indicator function evaluating to one if the expression inside square brackets is true and zero otherwise. The superscript c indicates that the derivatives are valid for continous variables. Similar for discrete variables

$$\begin{aligned}
\frac{\partial m_i^d(\theta, x)}{\partial \beta_j} &= \frac{\partial E\left(loan\ size \mid x^{-j}, x_j = 1\right)}{\partial \beta_j} - \frac{\partial E\left(loan\ size \mid x^{-j}, x_j = 0\right)}{\partial \beta_j} \\
\frac{\partial m_i^d(\theta, x)}{\partial \gamma_j} &= \frac{\partial E\left(loan\ size \mid x^{-j}, x_j = 1\right)}{\partial \gamma_j} - \frac{\partial E\left(loan\ size \mid x^{-j}, x_j = 0\right)}{\partial \gamma_j} \\
\frac{\partial m_i^d(\theta, x)}{\partial \sigma^2} &= \frac{\partial E\left(loan\ size \mid x^{-j}, x_j = 1\right)}{\partial \sigma^2} - \frac{\partial E\left(loan\ size \mid x^{-j}, x_j = 0\right)}{\partial \sigma^2}
\end{aligned}$$

Bootstrap method

An alternative to the delta method for obtaining standard errors for the marginal effects is the bootstrap method (see Cameron & Trivedi, 2005). With this approach the sample of *N* households surveyed is considered the population (referred to as the population in the following). Using a non-parametric boot strap, syntetic samples of size *N* are obtained by drawing with replacement *N* times from the population. For each sample drawn, the hurdle model is estimated and marginal effects are calculated and stored. Repeating the procedure a large number of times, say 1,000, produces a distribution of marginal effects. The empirical variance estimator of the sample of marginal effects is then obtained by the usual variance formula. This bootstrap estimate of the variance is used for statistical inference. This approach is used in Table 8. The code is available from the author.

Understanding Victimization: The Case of Mozambique

Mikkel Barslund, John Rand and Finn Tarp *

University of Copenhagen, Denmark
Development Economics Research Group (DERG)

Jacinto Chiconela
Ministry of Planning and Development (MPD), Mozambique

November 2004

Revised: October 2006

Abstract

This paper analyzes how economic and non-economic characteristics at the individual, household and community level affect the risk of victimization in Mozambique. We use a nation wide representative household survey from Mozambique with unique individual level information and show that the probability of being victimized is increasing in income, but at a diminishing rate. The effect of income is dependent on the type of crime, and poorer households are vulnerable. While less at risk of victimization, they suffer relatively greater losses when such shocks occur. Lower inequality and increased community level employment emerge as effective avenues to less crime.

Keywords – Victimization, Crime, Mozambique

JEL classification: K40, K42, O55

---

* Address correspondence to: mikkel.barslund@econ.ku.dk. We are grateful to Channing Arndt, Kenneth Simler and Katleen Van den Broeck for advice and excellent comments. The same goes for participants at the 10th Annual Conference of the African Econometric Society (AES) held in Nairobi in July 2005 and the March 2006 CSAE conference held in Oxford. Effective research assistance by Christopher Sonne-Schmidt is appreciated. Most constructive reviews by three anonymous referees helped clarify a variety of thorny issues and sharpen our focus. For this we are grateful. The usual caveats apply.

1. INTRODUCTION

This paper is a contribution to the literature on violence and the socio-economic determinants of individual and household level victimization. Crime is rightly perceived as a critical constraint on economic development, and the many ways in which crime and insecurity may affect human welfare in developing countries are aptly summarized by Fafchamps and Moser (2003). Consequently, efforts geared towards a better understanding of victimization and the formulation of comprehensive and effective policies to combat criminal behavior and minimize victimization are areas with potentially very high social and economic returns. This is certainly so in Mozambique. This war- and drought-stricken southern African country provides the case context for the present paper, and in spite of recent economic growth, Mozambique remains one of the poorest countries in the world. Mozambique also continues to suffer from a particularly violent economic and political history with roots going back to an extreme case of colonial domination and exploitation. Misguided economic policies in the immediate post-independence years added further complications, and the apartheid regime in South Africa fiercely undermined the development efforts of the 1980s.

By way of background and to put crime in Mozambique in perspective, Table 1 provides comparable statistics for homicide rates around the world. According to UN (2005) intentional homicide is reasonably accurately measured in the official police statistics, especially due to relatively high rates of reporting. Among 33 countries in the African region, only South Africa, Lesotho and Swaziland report higher homicide rates than Mozambique, and based on world wide information from WHO (2002) it appears that only Columbia and Brazil in other regions experience a higher rate of homicides per 100,000 people. Moreover, data in Dgedge *et al*. (2001) and Nizamo *et al*. (2006) suggest that the homicide rate in Mozambique increased significantly between 1994 and 2000. In 1994, 188 homicides were recorded in the capital city of Maputo as compared to 225 in 2000. This is equivalent to homicide rates of 17.0 and 22.1, respectively. Finally, official crime data from the Mozambican Ministry of the Interior (2004) suggest that 39,061 and 40,630 criminal offences of different types took place in respectively 2002 and 2003 in the whole of Mozambique. With a population of around 18 million this indicates that no less than 220 per 100,000 people are victimized every year.[1]

Table 1: Homicide rates in perspective

	Homicides per 100,000 people
Mozambique	**22.1**
Angola	9.4
Botswana	12.9
Lesotho	50.6
Namibia	16.3
Sao Tome and Principe	6.2
South Africa	59.7
Swaziland	50.9
Tanzania	7.8
Uganda	20.9
Zambia	8.1
Zimbabwe	7.2
The Americas	19.3
South East Asia	5.8
Europe	8.4
Eastern Mediterranean	7.1
Western Pacific	3.4
All countries	8.8

Note: Mozambique is based on data for Maputo only. This may give an upward bias, since homicide rates are generally higher in urban areas. The figures reported for Lesotho, Namibia, South Africa, Swaziland, Tanzania and Uganda are based on simple averages of estimates from two or three different sources. The table does not present statistics for African countries with a homicide rate below five. These countries include: Algeria (1.0), Benin (4.5), Burkina Faso (0.4), Cameroon (0.4), Cote d'Ivoire (3.3 – based on an average of two estimates), Djibouti (3.5), Egypt (0.4), Eritrea (2.0), Ghana (2.1), Libya (2.3), Madagascar (4.7 – based on an average of two estimates), Mali (0.7), Mauritania (0.8), Mauritius (2.5 – based on an average of three different estimates), Morocco (0.5), Niger (0.9), Nigeria (1.5), Rwanda (4.5), Senegal (1.2), Seychelles (4.4 –based on an average of two different estimates), Sudan (0.3) and Tunesia (1.2).
Source: Mozambique: Nizamo et al. (2006), Demombynes and Özler (2005), Fafchamps and Minten (2006), Fajnzylber et al. (2002), UN (2005) and WHO (2002).

While officially reported criminal statistics are available in Mozambique, they are as is the case elsewhere in the developing world not representative of the 'true' criminal situation. This is so first of all due to underreporting of crimes to the police. Alternative sources of information such as national surveys where information is available at the level of individuals and households are therefore a potentially rewarding avenue to improve our understanding of victimization in developing countries. Here we are fortunate that a nationally representative household survey was conducted in Mozambique during 2002 and 2003. It included a novel and informative section on individual level victimization on which the empirical section of this paper is built.

The first and main objective of our paper is to provide a ‘map’ for identifying individuals with the highest risk of being subjected to crime of different kinds in Mozambique and to identify appropriate policy measures, or at least broad areas, where policy action appears to be effective in curtailing crime. In this part of our analysis, we look at the risk of being victimized due to property crimes and physical assaults of various kinds and identify the correlates of victimization. We subsequently look more narrowly at how income and the economic loss from property crime are related at the household level. Some 90% of those victimized in our data suffered a property loss, and we ask whether relatively well off households are victims of minor property crimes and poor households are victims of economically more disruptive crimes.

When it comes to the question of which determinants of crime and victimization the analyst should include in empirical work, there is no clear cut, commonly agreed methodology. Economic theory calls attention to offender motivation and behavior (Gaviria and Pàges, 2002); and Becker (1968) and Ehrlich (1973) suggest that criminal acts can be viewed as being directly linked to rational economic decision making by the offender. He/she carries out an *a priori* cost-benefit type analysis, where perceived economic costs and benefits are weighed against each other. Subsequent action is decided on this basis. This suggests that the probability of being a victim can be expected to be a positive function of indicators related to income, education, and employment status as well as to the severity and effectiveness of the preventive and punitive actions taken by society. Thus, in economic attempts to explain crime, focus has generally been on how the offender perceives the optimal balance between gains from criminal activity, on the one hand, and associated constraints and risks (i.e. the risk of getting caught and the punishment involved), on the other. In addition, economic studies have generally focused on the application of household and community level data and information. In contrast, sociological studies have in their theories of criminal victimization put emphasis on victim behavior and individual (personal) level characteristics, which are at least to some extent non-economic in nature. Limited attention is paid to factors associated with offender motivation and perception. The most prominent perspectives in this group are the ‘lifestyle exposure’ and ‘routine activity’ theories associated with Hindelang et al. (1978) and Cohen and Felson (1979).

Existing literature reveals with some notable exceptions a lack of attempts at trying to integrate economic and sociological perspectives.[2] We believe this is unfortunate. Economic and sociological variables (and data at household, community as well as individual level) should be considered on par with each other to identify which individuals are most likely to be victims of criminal acts. A second main objective of the present paper is therefore to develop an integrated analysis where both economic and sociological factors are 'allowed to play their part'. More specifically, we aim at helping to clarify the role of economic and sociological factors and to come better to grips with the value added of an integrated framework as a basis for future research. This is so both with the regard to the general analytical approach and in the search for appropriate policy measures.

The paper is organized as follows. Section 2 digs into existing economic and sociological literature to identify in an organized manner a pertinent set of specific determinants of victimization to be included in the empirical analysis. This is followed in Section 3 by an overview of our empirical methodology as well as the data used, together with some descriptive statistics. Section 4 presents results, while Section 5 looks into robustness. Section 6 concludes and outlines policy implications.

2. DETERMINANTS OF VICTIMIZATION

In this section, we aim at identifying the set of economic and sociological variables, which should enter our unified empirical framework. To structure the process we follow the sociological categorization of Cohen et al. (1981), and focus on the role played by: Exposure, guardianship, proximity to potential offenders, and attractiveness of potential targets. Definitions of these four factors are listed in the Appendix, and Table 2 provides a summary of the empirical variables at individual, household and community level associated with each category. The expected sign of each variable is also indicated.

Table 2: List of empirical variables by sociological category

Group	Variables	Exposure	Guardianship	Proximity	Attractiveness
Individual	Individual income				x (?)
	Gender (Female = 1)	x (-)			
	Age	x (-)			
	Employment status (Employed)	x (+)			
	Marital status (Single)	x (+)			
	Education			x (?)	
Household	Household consumption				x (?)
	Possessions: Durable goods				x (+)
	HH gender (Female = 1)	x (-)			
	HH age	x (-)			
	HH employment status	x (+)			
	HH education			x (?)	
	Distance to police station		x (-)		
	Members in household		x (-)		
	Family composition		x (-)		
Community	Population density			x (+)	
	Unemployment rates			x (+)	
	Distribution of Income: Inequality			x (+)	
	Average level of educational attainment			x (-)	
	Integration			x (+)	

Note: All regressions in the econometric analysis also include regional and urban/rural dummies as community variables. In parenthesis are the expected sign of the variable on victimization. Income is listed with a '?' since income involves different predictions depending on the type of crime. We expect (i) larceny to be positively linear in income, (ii) burglary to be increasing in income, but at a diminishing rate, and (iii) assault to be indeterminate.

Within sociological theories of victimization the most cited contributions are as already noted the 'lifestyle-exposure perspective' and the 'routine activity theory'. Meier and Miethe (1993) argue that the key difference between these frameworks is that the routine activity perspective was created to account for changes in crime rates over time, while the lifestyle-exposure approach was developed with a view to capturing differences in victimization risks across social groups. Common to the lifestyle and routine activity theories is their emphasis on 'exposure'. When this factor is included in the analysis it amounts to saying that social and economic interactions increase the risk of victimization, and it is indeed a dimension that has been mostly ignored in economic approaches to crime and victimization. Accordingly, differences in the likelihood of being a victim can be explained by differences in the lifestyles of the potential victims. People, who are more active in the public domain and engage more in non-household activity, use less time within the family and are more frequently in contact with individuals with criminal tendencies. We believe this is captured by individual level variables such as gender and age as well and employment and marital status, in combination with selected characteristics of the household head. The sociological 'exposure' argument is supported by empirical observation. Fajnzylber *et al*. (2000) offer one of the few studies of victimization in the economic literature using individual level characteristics. It shows in line with the overview of sociological studies by Meier and Miethe (1993) that young, single, employed males have higher probability of being victimized than their demographic counterparts.

Considering the guardianship factor next, the economic literature often refers to the nature of security measures (including distance to police station). They are expected to deter crime by increasing the offenders' expected costs through a higher risk of being caught. As such this is a typical example of the economic incentives approach. Yet, in developing countries where public services of police and justice are often of questionable quality and in limited supply (or distributed unequally), private deterrence may well become more important. This may call into play a set of more sociological individual level factors related to members of the household and family composition, not captured in the typical economic analysis. The prediction in this kind of individual level analysis is that the risk of victimization should decrease with the number of members in a household and the share of males because the household is considered a social network of protection.[3] Turning to existing studies, Fajnzylber *et al*. (2000) find no significant evidence of members of the household influencing the risk of victimization, and the review in Smith and Jarjoura (1989) concludes that the empirical evidence regarding the effect of the household size and family composition is mixed. To pursue this topic through the household and family composition variables listed in Table 2 is interesting.

In line with economic thinking, we also include distance to local police stations as a measure of guardianship. This measure (and other measures of the density of police personnel available in the nearby neighborhood) has to our knowledge mostly been included in more aggregate analysis of crime (see for example Fajnzylber *et al*., 2002a and 2002b). As in Zenou (2003) our prediction is to find a negative relationship between the police guardianship variable and the probability of being victimized. The underlying assumption is that the longer the distance to a local police station is the higher probability of being victimized. However, conclusions based on this variable should be handled with caution. Causality could potentially run from crime/victimization rates to distance to police as further elaborated on in Section 4.

Turning to the 'proximity' category, most explanatory variables considered here are at the community level; household head and individual education being the exceptions. The effect of education has been studied at different levels of aggregation and great care should be taken when trying to compare these results. In the following we limit our comments to the literature considering victimization studies. First, at the cross-country level Soares (2004) finds that education reduces crime/the probability of being victimized. This is consistent with the result obtained in Fajnzylber *et al*. (2000) showing that the average educational attainment level in a society is positively associated with lower levels of victimization (even more so for assaults). At a more disaggregated (household) level, Gaviria and Pàges (2002) show that education of the household head increases the risk of being victimized; and at individual level Fajnzylber *et al*. (2000) find no significant effects of individual years of schooling on the risk of victimization. These results suggest that societies with higher levels of education tend to have lower rates of victimization, but whether an individual is more or less educated within a given society does not necessarily affect the probability of victimization. We therefore have no clear priors on the sign of the education variables.

Considering the community level proximity characteristics captured here through five indicators it is to be expected that the closer people reside to relatively large groups of motivated offenders, the greater is the risk of victimization. This is supported by the empirical evidence of Meier and Miethe (1993). People living in larger urban areas are in this perspective more exposed to crime, a finding which has however been challenged by Fafchamps and Moser (2003) in the case of Madagascar. Here crime and insecurity are associated with isolation, not urbanization. Moreover, it is often argued that individuals living in areas with high unemployment rates are at a greater risk of becoming a victim. This corresponds with the results obtained by Cohen *et al.* (1981), who note that the risk of being victimized increases in poorer neighborhoods. This observation is also in line with Bourguignon *et al.* (2003), Demombynes and Özler (2005), Lederman *et al*. (2002) and Soares (2004). They consider the effects of inequality on crime, and find that income inequality affects crime rates positively. The share of foreigners in total district population is included here as well as a proxy for the social tension that potentially exists in a former colonized country. We predict that larger social tension at the community level increases the probability of being victimized. This issue has to our knowledge not been directly addressed in the victimization literature before.

Finally, if crime is motivated by instrumental (economic) ends it is generally expected that the greater the attractiveness of a target (income level and ownership of expensive and portable consumer goods as indicated in Table 2), the greater the risk of victimization. Yet, the effect of income on victimization risk is probably highly dependent on the nature of the crime as noted by Cohen *et al.* (1981). This has in our view not been adequately acknowledged in the economic literature and is an area where the current sociological and economic literatures differ substantially. For example Cohen *et al.* (1981) argue that in terms of assault the proximity, exposure and guardianship effects seem to dominate increased attractiveness caused by higher incomes. This may, *ceteris paribus*, lead to a negative relationship between income and the risk of assault, in contrast to the positive relationship emerging from economic thinking. Income also has two opposing effects on burglary victimization risk, and according to Cohen *et al.* (1981) it is not clear whether proximity, exposure and guardianship will dominate the influence from increased attractiveness in this case. According to the routine activity approach and consistent with the economic approach to victimization the risk of being a victim of larceny will always be increasing in income. Overall, it is not entirely clear *a priori* whether more economic resources allow individuals to avoid risky and vulnerable situations or whether this attracts more criminals (and thus increases the probability of becoming a victim). The attractiveness of individuals and their associated properties play an important role in increasing the risk of victimization, but higher levels of self-protection or guardianship will decrease individual risk. This can only be settled through empirical analysis.

In sum, the sociological and economic literatures offer somewhat different underlying perspectives on the causal links behind victimization. It is equally clear that conclusions about the effects of specific economic and sociological variables (such as income) on the probability of becoming a victim vary. We therefore feel motivated to go on to analyze how different economic and sociological characteristics at the individual, household and community level affect the risk and loss of victimization in

Mozambique within a unified analytical framework. We highlight that while we have tried in Table 2 to link explanatory variables to one of the four sociological categories (exposure, guardianship, proximity, and attractiveness), there are potential ambiguities involved. A particular variable may in theory be associated with more than one category. This was referred to above when we discussed income. Another example is that gender may be associated with exposure (at the individual and household level) as well as guardianship. Females may not be quite as effective in protecting their household as males. The same kind of consideration may go for age. It might also be speculated that employment status and education are linked to attractiveness in addition to exposure and proximity, respectively. While this effect is at least in part controlled for by having income in the attractiveness category, these kinds of ambiguities form part of the motivation of the empirical part of this paper and are further discussed in Section 4.

3. METHODOLOGY, DATA AND DESCRIPTIVE STATISTICS

Based on the literature survey in the previous section we take a reduced form approach to modeling the probability of an individual being victimized. Formally,

$$\Pr(y_{ijc} = 1 \mid x_{ijc}, z_{jc}, q_c) = f(x_{ijc}, z_{jc}, q_c, e_{ijc}) \qquad (1)$$

where y_{ijc} is an indicator variable showing whether an individual *i*, who is a member of family *j* that lives in community *c*, was a victim of crime. The dependent variable takes on a value of one if the individual was victim of a crime and zero otherwise. x_{ijc}, z_{jc}, q_c are vectors of respectively individual, household and community characteristics, whereas ε_{ijc} is an individual error term. We also estimate (1) at the household level, the dependent variable y^h_{jc} indicating whether any member of the household was victimized.

We use a probit model as our preferred specification, and interpret (1) as derived from an underlying latent variable model. In this model, we assume that $y_{ijc} = 1[\alpha_0 + x_{ijc}\alpha_1 + z_{jc}\alpha_2 + q_c\alpha_3 + e_{ijc} > 0]$ with e_{ijc} being normally distributed. In

Section 5 we test the robustness of our results with respect to the variables included. Choosing a logit or linear probability model instead of the probit specification would not affect the results reported.

In the analysis of the relative loss from property crimes at the household level we rely on Heckman's selection framework. For household *j* the relative loss, l_{jc}, can be expressed as:

$$l_{jc} = \beta_0 + z_{jc}^{*h}\beta_1 + q_j^{*h}\beta_2 + u_{1j} \tag{2}$$

where superscript *h indicates that the vectors z_{jc}^{*h} and q_j^{*h} are not necessarily identical to their counterparts in the household level version of (1). A loss is only observed if:

$$\gamma_0 + z_{jc}^{h}\gamma_1 + q_j^{h}\gamma_2 + u_{2j} > 0,\ (u_{1j}, u_{2j}) \sim binorm(0,0,\sigma,1,\rho) \tag{3}$$

Equation (3) is our selection equation, and it is the household level equivalent of the underlying latent variable model above. Note that the samples used in estimating (1) and (3) differ. Households, who suffer a loss, are a subset of all households victimized.[4]

In all estimations, appropriate household weights are used, taking into account the survey design (i.e. stratification of the survey sample and the clustering of enumeration areas).

The data come from a nationally representative household survey (IAF) conducted in Mozambique during 2002 and 2003 by the National Institute of Statistics (INE). The survey took place over the space of a year, beginning in July 2002 and ending in June 2003. Data collection was carried out in clusters of nine and 12 households in respectively rural and urban regions using a stratified sampling process with 21 strata (consisting of 10 provinces, each divided into a rural and an urban zone, plus Maputo city).[5] A total of 858 clusters make up the sample of 8,700 households. After data collection, INE constructed household weights so as to ensure that the sample is representative at the national, regional and rural/urban levels in accordance with the 1997 census.

The survey contains detailed information on individual characteristics including victimization entries on robbery, assault and larceny for around 43,000 individuals distributed among the 8,700 households. The survey instrument also includes questions on general characteristics of the individual and the household (including whether or not individuals have been victimized), daily expenses and home consumption, possession of durable goods, gifts and transfers received. Other expenses, which tend to occur with lower frequency than daily expenditures, such as school fees or purchases of clothing are covered as well. Additional details on the survey can be found in MPF *et al.* (2004). Full documentation of all aspects of the 2002-03 IAF survey is available in Portuguese from the National Institute of Statistics (INE, 2004).

In our analysis we consider individuals aged 12 and above, but variables measured at the household level include information on the complete household, i.e. including members aged less than 12 years. A number of households had to be excluded due to missing information, so our final sample consists of 25,594 individuals distributed among 8,515 households.

The questionnaire includes a novel and detailed section on victimization of each member of the family as well as related questions at the household level. These are the data on which we focus in this paper, and by way of background, we note that 6.4% of the respondents are of the view that criminal offenses are the main social problem in Mozambique at the moment. Moreover, 19% of the households answered that crime in their residential area had increased during the past 12 months. About half of the households in our sample felt unsafe when walking alone at night, even though only around 27.5% of the households have experienced a household member being victimized. Table 3 gives an overview of the types of crimes faced by the households. About two-thirds of the crimes can be characterized as some kind of theft or robbery, whereas rape, other sexual abuse, assault and domestic violence account for 4.8%. Interestingly, bribery does not come across as particularly serious in Mozambique.

Table 3: Victimization statistics

Type of crime	Percent
A. Purse snatching	6.2
B. Tried or took an object of value	3.4
C. Robbed of a bicycle	4.1
D. Robbed of any type of vehicle	0.3
E. Cattle stolen	5.5
F. Victim of other theft	47.7
G. Victim of rape	0.4
H. Insult or offensive	7.3
I. Threats	1.8
J. Assault	2.7
K. Sexual abuse	0.3
L. Domestic Violence	1.4
M. Bribery	0.5
N. Other	18.5
Total	100

Locality of crime	Percent
At home	66.9
On roads	8.9
In public transport	1.6
In the market	3.2
At work	3.2
In places of leisure	1.2
Other	14.9
Total	100

From Table 3 it is also clear that most offences happen within the household premises. One third of crimes take place in the public domain, including in particular on roads (8.9%), in the market (3.2%) and at work (3.2%).

In Section 2, we identified a number of potential determinants for being victimized, and information on these determinants can be obtained from the IAF survey questionnaire. The variables used in the analysis are listed together with descriptive statistics in Table 4 at the individual and at the household level. Most variables come straight out of the survey, but a few had to be constructed as explained below. To ease our brief overview of the determinants, they are grouped according to the classification in Table 2.

The first set of determinants considered here is those in the attractiveness group. The average monthly individual (nominal) income is 0.92 million Meticais (around 37 US$) and the household income is on average 2.77 million Meticais (approximately 111 USD).[6] This figure might suggest underreporting of income. At the household level the real annual consumption is on average 14.1 million Meticais (562 US$), which as expected is somewhat above the GDP per capita of 212 US$ in 2000 reported in WDI (2003). Last, possession of durable goods, which is expected to make individuals more prone to being a victim, is measured by household dummy variables for having at least one TV, radio and bicycle in the household. Around half of households own a radio, whereas bicycle and TV ownership is more limited at 26.8% and 11.9%, respectively.

The second group of determinants includes those in the exposure group, including gender (individual and household head), age (individual and household head), employment status (individual and household head), and marital status. At the individual level, 53.3% of the sample consists of women, but only around 22.8% (27.2% measured at the household (HH) level) of individuals has a woman as household head. The average age of individuals is 31.1 years, and for household heads this figure is 45.5 (43.4 years measured at the HH level). As already pointed out in Section 2, employment status is potentially an important determinant of victimization. At the individual level, 11.7% of the sample is registered as being without work, and for household heads this figure is 7.0%. This corresponds with the average for sub-Saharan Africa and the information on Mozambique in WDI (2005). The final determinant at individual level in the exposure group considers the marital status of the individuals. Married or cohabiting partners make up 51.4% of the sample, 37.1% are single and the rest are either divorced or widowed.

Table 4: Summary statistics

Group	Classification	Variable	Individual		HH	
			Mean	Std. Dev.	Mean	Std. Dev.
Individual and HH level	Type of crime	Victim	0.089	0.284	0.266	0.442
		Burglary	0.055	0.228	0.169	0.375
		Larceny	0.028	0.164	0.086	0.281
		Assault	0.013	0.113	0.039	0.194
Individual	Attractiveness	Individual income (monthly)	0.924	7.295		
Individual	Exposure	Gender	0.533	0.499		
		Age	31.106	16.489		
		Employment status: Employed	0.667	0.471		
		Employment status: Studying	0.216	0.411		
		Employment status: Unemployed	0.117	0.322		
		Marital status: Single	0.371	0.483		
		Marital status: Married	0.101	0.301		
		Marital status: Married Polygam	0.074	0.262		
		Marital status: Cohabit	0.339	0.473		
		Marital status: Divorced	0.059	0.236		
		Marital status: Widow	0.056	0.230		
Individual	Proximity	Education: Educ0	0.247	0.431		
		Education: Educ1	0.382	0.486		
		Education: Educ2	0.207	0.405		
		Education: Educ3	0.107	0.309		
		Education: Educ4	0.057	0.233		
Household	Attractiveness	HH income (monthly)			2.777	12.896
		HH consumption (yearly)			14.056	32.427
		Possession of durable goods: TV	0.172	0.378	0.119	0.324
		Possession of durable goods: Radio	0.553	0.497	0.494	0.500
		Possession of durable goods: Bicycle	0.296	0.457	0.268	0.443
Household	Exposure	HH gender	0.228	0.420	0.272	0.445
		HH age	45.467	14.513	43.432	15.317
		HH employment status	0.930	0.254	0.942	0.234
Household	Proximity	HH education 0	0.239	0.427	0.278	0.448
		HH education 1	0.351	0.477	0.359	0.480
		HH education 2	0.209	0.407	0.187	0.390
		HH education 3	0.103	0.303	0.093	0.291
		HH education 4	0.098	0.298	0.082	0.275
Household	Guardianship	Household size	6.441	3.407	5.067	2.794
		Adult male share	0.243	0.170	0.245	0.207
		Distance to police station 1	0.441	0.496	0.373	0.484
		Distance to police station 2	0.123	0.329	0.124	0.330
		Distance to police station 3	0.069	0.254	0.076	0.265
		Distance to police station 4	0.073	0.260	0.087	0.281
		Distance to police station 5	0.294	0.456	0.340	0.474
Community	Proximity	Unemployment rate	0.157	0.140	0.136	0.129
		Distribution of Income: Inequality	0.394	0.095	0.385	0.099
		Average level of educational attainment	1.420	0.590	1.340	0.573
		Population density	1.181	3.515	0.860	2.938
		Integration	0.018	0.006	0.018	0.006
N		Total observations	25,594		8,515	

Proximity determinants such as education and the community variables are listed in the third group. Measures of the individual level of human capital, i.e. education, is represented by five dummy variables.[7] At the level of the household head, education level is measured by a categorical variable with values from zero to four, constructed from the five dummy variables at the individual level.[8] Some 24.7% of the sample has no education at all, and in general the educational level is as expected quite low. The next variable in the proximity group is the community unemployment rate, which is constructed from the survey data. It is measured at the district level (a geographical and administrative unit below the province level) by looking at the share of adults aged 18 years and above, who stated in the IAF that they did not have work and were not studying. With a mean unemployment rate of 15.7%, this community indicator is above the individual unemployment level referred to above.

Turning to inequality at the district level we use the Gini-coefficient of real expenditure (i.e. spatially and temporally deflated) extracted from the survey data.[9] The Gini-coefficient in our sample is 0.39. This is around the average for sub-Saharan Africa (WDI, 2005), and the reported 0.40 for Mozambique in 1997. The average population density is also recorded at the district level and is based on the 1997 census. In the sampled areas the average population density is around 1,181 persons per square kilometer. As a crude proxy for how integrated each community is we use the information in the 1997 census on the number of foreigners living in each district to form the share of foreigners in total district population. The average share of foreigners in our sample is quite low at 1.8%.

Finally, we have descriptive statistics concerning the characteristics related to the guardianship characteristics, such as household size, family composition (share of adult males over the age of 18 in the household) and household distance to a police station. The average household size in the sample is 6.4 individuals (5.1 at the household level), and the adult male share is around 24.3%. Distance to the police station is reported by the household as a categorical variable corresponding to different lengths of time it takes to reach the nearest police station. A third of the population has less then half an hour to the nearest police station by foot, but variation is large and 29.4% of the sample has more than a 120 minutes walk to the police.[10]

4. RESULTS

Some 26.6% of the households in our sample and 8.9% of people at the individual level experienced as shown in Table 4 one or another kind of criminal act(s) during the past 12 months according to the Mozambican IAF.[11] At the individual level, the 8.9% were victimized at least once during 2002/03, but only 1.3% of the 25,594 observations were physically assaulted. Most of the crimes registered in the survey were burglaries (5.5%), whereas cases of larceny were reported for 2.8% of the sample.

(a) Determinants of Victimization

Tables 5 and 6 present the main findings of our econometric analysis of the probability of being victimized (marginal effects at the mean of the data); and the discussion in what follows is organized in accordance with the four groups of determinants identified in Section 2. We start with the attractiveness and end with the guardianship variables, and possible ambiguities in classifying the various explanatory variables are alluded to. Table 5 documents the baseline formulation at the individual level including the 25,776 observations analyzing aggregate victimization as well as a disaggregation into three types, burglary, assault, and larceny. Table 6 shows results of the analysis at the household level using household consumption as a proxy for income. This has been done recognizing that income may be measured with error.[12]

Table 5: Individual level probit results

Group	Classification	Variable	Victim: Marginal effects	Victim: *t*-stats	Burglary: Marginal effects	Burglary: *t*-stats	Larceny: Marginal effects	Larceny: *t*-stats	Assault: Marginal effects	Assault: *t*-stats
Ind	Attractiveness	Income (x100) (coefficients) [a)]	1.160***	(2.57)	1.432***	(3.51)	0.548	(1.60)	0.767**	(2.03)
		Income squared (coefficients) [a)]	-0.003*	(1.68)	-0.005***	(2.95)	-0.001	(1.09)	-0.002	(1.21)
		Income (x100) (marginal effect)[b)]	0.130**	(2.54)	0.095***	(3.40)	0.019	(1.59)	0.194	(1.33)
Ind	Exposure	Gender (x10)	-0.551***	(8.93)	-0.297***	(8.19)	-0.152***	(5.04)	-0.028**	(2.12)
		Age (x100)	0.015***	(8.39)	0.107***	(9.13)	0.040***	(4.58)	-0.001	(0.23)
		Empl. status: Studying	-0.028***	(2.61)	-0.014**	(2.24)	-0.003	(0.54)	-0.005	(1.63)
		Empl. status: Unemployed	-0.022***	(3.33)	-0.011**	(2.39)	-0.006**	(1.98)	-0.002	(1.20)
		Marital status: Married	0.085***	(7.11)	0.056***	(5.95)	0.019***	(2.99)	0.003	(0.95)
		Marital status: Polygam	0.138***	(9.16)	0.075***	(6.49)	0.043***	(4.71)	0.015***	(3.36)
		Marital status: Cohabit	0.090***	(9.48)	0.048***	(7.09)	0.027***	(4.74)	0.010***	(3.33)
		Marital status: Divorced	0.161***	(8.06)	0.067***	(5.65)	0.050***	(4.29)	0.038***	(3.94)
		Marital status: Widow	0.144***	(7.06)	0.075***	(5.58)	0.053***	(3.90)	0.013**	(2.06)
Ind	Proximity	Education: Educ1 (x10)	0.295***	(4.72)	0.198***	(5.10)	0.108***	(3.63)	-0.003	(0.24)
		Education: Educ2 (x10)	0.448***	(4.64)	0.361***	(4.96)	0.114**	(2.53)	-0.003	(0.19)
		Education: Educ3 (x10)	0.446***	(3.19)	0.258**	(2.23)	0.161**	(2.47)	-0.011	(0.49)
		Education: Educ4 (x10)	0.571***	(3.10)	0.423***	(2.82)	0.206**	(2.16)	-0.014	(0.37)
HH	Attractiveness	Durable goods: TV (x10)	0.049	(0.61)	0.008	(0.19)	0.069	(1.54)	-0.029**	(2.16)
		Durable goods: Radio (x10)	-0.056	(1.30)	-0.031	(1.09)	-0.003	(0.14)	-0.019*	(1.75)
		Durable goods: Bicycle (x10)	0.055	(1.12)	0.037	(1.05)	-0.012	(0.59)	0.026**	(2.17)
HH	Exposure	HH Gender (x10)	0.405***	(5.33)	0.217***	(4.16)	0.118***	(3.48)	0.037**	(2.19)
		HH Age (x100)	-0.142***	(7.69)	-0.082***	(6.49)	-0.040***	(5.04)	-0.012**	(2.36)
		HH Employment status (x10)	-0.085	(1.02)	-0.068	(1.31)	-0.030	(0.74)	0.006	(0.30)
HH	Proximity	HH Education 1 (x10)	-0.050	(0.86)	-0.074**	(1.96)	-0.004	(0.19)	0.014	(0.88)
		HH Education 2 (x10)	0.021	(0.29)	-0.056	(1.09)	0.026	(0.82)	0.071***	(2.98)
		HH Education 3 (x10)	-0.072	(0.70)	-0.105	(1.46)	0.010	(0.22)	0.036	(1.24)
		HH Education 4 (x10)	0.010	(0.07)	-0.108	(1.42)	-0.008	(0.20)	0.163**	(2.02)
HH	Guardianship	Household size	-0.004***	(5.35)	-0.002***	(4.50)	-0.001***	(2.68)	-0.001***	(2.96)
		Adult male share	-0.042***	(2.88)	-0.028***	(2.90)	-0.002	(0.34)	-0.001	(0.17)
		Distance to police 2	0.002	(0.30)	-0.002	(0.37)	0.008*	(1.88)	-0.003**	(2.31)
		Distance to police 3	-0.006	(0.43)	-0.001	(0.13)	0.002	(0.40)	-0.001	(0.50)
		Distance to police 4	-0.017*	(1.68)	-0.009	(1.44)	-0.002	(0.43)	-0.004**	(2.12)
		Distance to police 5	-0.024***	(2.98)	-0.018***	(3.28)	0.001	(0.30)	-0.002	(1.19)
Com	Proximity	Unemployment rate	0.104	(1.49)	0.070	(1.54)	0.014	(0.46)	0.037***	(2.91)
		Inequality	0.087**	(2.12)	0.035	(1.20)	0.040***	(2.79)	0.003	(0.46)
		Average educational level	-0.006	(0.53)	-0.007	(0.89)	0.000	(0.08)	0.001	(0.49)
		Population density	0.001	(0.82)	0.001*	(1.94)	-0.001*	(1.66)	0.000	(0.11)
		Integration	-0.528	(1.04)	-0.415	(1.13)	-0.099	(0.32)	0.051	(0.48)
Including regional and rural/urban dummies			Yes		Yes		Yes		Yes	
		Observations	25594		25594		25594		25594	
		Pseudo R-squared	0.15		0.15		0.15		0.14	

Note: Probit, marginal effects at the mean of all variables. *, **, *** indicates significance at a 10%, 5% and 1% level, respectively. Including only individuals, who are living in a household reporting non-zero (positive) income. Base: Individual male, individual no education, individual employed, individual single, household head male, household head unemployed, household head no education, distance to police station 1, Maputo.

[a)] The two first rows show coefficients for the non-linear effect of income. Implied turning points are much larger than the 99 percentile of the distribution. All other rows show marginal effects at the mean of all variables.

[b)] Marginal effect of income evaluated at the mean of all variables.

Table 6: Household level probit results

Group	Classification	Variable	Victim Marginal effects	Victim *t*-stats	Burglary Marginal effects	Burglary *t*-stats	Larceny Marginal effects	Larceny *t*-stats	Assault Marginal effects	Assault *t*-stats
HH	Attractiveness	Real cons. (x100) (coefficients) [a]	0.755***	(4.36)	0.667***	(3.74)	0.538***	(3.25)	0.335*	(1.95)
		Real cons. squared (x100) (coefficients) [a]	-0.001***	(2.99)	-0.001***	(3.21)	-0.001**	(2.23)	-0.000	(0.99)
		Real cons. (x100) (marginal effects) [b]	0.151***	(2.85)	0.084**	(2.12)	0.041**	(2.04)	0.054	(1.54)
		Durable goods: TV	0.011	(0.35)	-0.004	(0.22)	0.027	(1.22)	-0.016***	(3.02)
		Durable goods: Radio	-0.015	(0.96)	-0.010	(0.76)	-0.001	(0.15)	-0.008*	(1.91)
		Durable goods: Bicycle	0.020	(1.19)	0.022	(1.53)	-0.005	(0.58)	0.009	(1.62)
HH	Exposure	HH Gender	0.050**	(2.47)	0.011	(0.69)	0.022**	(2.04)	0.015**	(2.15)
		HH Age (x100)	-0.063	(1.42)	0.012	(0.34)	-0.026	(0.94)	-0.047***	(3.04)
		HH Employment status	0.022	(0.82)	0.004	(0.22)	0.007	(0.46)	0.006	(0.86)
HH	Proximity	HH Education 1	0.060***	(3.11)	0.029**	(2.00)	0.035***	(3.34)	0.006	(0.96)
		HH Education 2	0.102***	(4.14)	0.074***	(3.72)	0.045***	(2.83)	0.023**	(2.52)
		HH Education 3	0.048*	(1.72)	0.005	(0.21)	0.045**	(2.51)	0.009	(0.89)
		HH Education 4	0.058*	(1.65)	0.026	(0.90)	0.029	(1.58)	0.037**	(2.06)
HH	Guardianship	Household size	0.010***	(3.35)	0.006***	(2.69)	0.002	(1.36)	0.001	(0.90)
		Adult male share	-0.003	(0.06)	-0.030	(0.87)	0.030	(1.38)	0.013	(1.01)
		Distance to police 2	0.047	(1.46)	0.007	(0.27)	0.049**	(2.53)	-0.012**	(2.06)
		Distance to police 3	-0.026	(0.54)	-0.005	(0.13)	0.006	(0.29)	-0.003	(0.36)
		Distance to police 4	-0.037	(0.98)	-0.029	(1.06)	0.003	(0.15)	-0.013*	(1.65)
		Distance to police 5	-0.065**	(2.23)	-0.067***	(2.82)	0.017	(1.06)	-0.008	(1.06)
Com	Proximity	Unemployment rate	0.312	(1.30)	0.226	(1.23)	0.071	(0.49)	0.160***	(3.06)
		Inequality	0.150	(1.09)	0.069	(0.57)	0.127*	(1.91)	-0.001	(0.04)
		Average educational level	-0.005	(0.14)	-0.031	(0.96)	0.015	(0.57)	0.003	(0.28)
		Population density	0.005	(1.30)	0.006**	(2.08)	-0.002	(1.18)	0.000	(0.11)
		Integration	-0.928	(0.53)	-1.495	(0.96)	0.134	(0.09)	0.531	(1.22)
Including regional and urban/rural dummies			Yes		Yes		Yes		Yes	
		Observations	8515		8515		8515		8515	
		Pseudo R-squared	0.04		0.06		0.09		0.10	

Note: Probit, marginal effects at the mean of all variables. *, **, *** indicates significance at a 10%, 5% and 1% level, respectively. Base: Household head male, household head unemployed, household head no education, distance to police station 1, Maputo.
[a] The two first rows show coefficients for the non-linear effect of income. Implied turning points are much larger than the 99 percentile of the distribution. All other rows show marginal effects at the mean of all variables.
[b] Marginal effect of income evaluated at the mean of all variables.

(i) Attractiveness

There is a statistically significant indication in the data of income being positively related to the probability of being victimized. This holds in all regressions in Tables 5 and 6 except for the larceny regression at the individual level (Table 5), although the marginal effects at the mean of the data in both assault regressions are not significant. Moreover, results suggest that there exists a non-linear relationship between income and victimization. The probability of being victimized is in Mozambique increasing in income, but at a diminishing rate. However, looking at the specific crime types this result is driven by the non-linearity between burglary and income/consumption and – for the household level – in the larceny regression. These results confirm Cohen *et al.* (1981), who concluded that the effect of income on victimization risk is highly dependent on the nature of the crime. However, we do (even for assault) find a positive relationship between income/consumption both at the individual and the household level independent of the types of crime faced.

Possession of durable goods is another attractiveness variable which is often expected to affect the risk of victimization positively. We are not able to confirm this relationship in our aggregate data. However, when disaggregating victimization by types of crime, we find a negative and significant relation between owning a TV or a radio and a positive relation with owning a bike in the assault regression. While assets may be expected to make potential victims more attractive, the ownership of a bicycle, a TV or a radio is also related to exposure, so there is ambiguity here. Accordingly, we believe our empirical result is due to the fact that (i) the TV and radio variables proxy for the amount of time individuals spend at home, and (ii) bicycle ownership is related to how much time individuals spend away from their immediate neighborhood. This implies that this particular result is probably more due to exposure associated with bicycle, TV and radio ownership, overriding the effect from the attractiveness dimension.

It follows that the results for the attractiveness variables are consistent with Fafchamps and Minten (2006). They conclude that certain forms of crime respond to economic incentives while others do not. This also reinforces the argument that both economic and sociological dimensions should be considered when analyzing victimization.

(ii) Exposure

The above observations are reinforced by looking at determinants in the exposure group. We generally find a significant influence of exposure variables on the probability of being victimized both at the individual and the household level. At the individual level we confirm that males have a higher probability of being victimized than females. However, members of female headed families have a higher probability of becoming a crime victim both in the individual and household level regressions. Males clearly tend to be more exposed than females, and the latter observation is in all likelihood driven by ambiguity vis-à-vis the guardianship dimensions of gender. Another result is that older people have a higher probability of victimization driven by property crime. This holds even though we include several attractiveness controls (people accumulate assets over time and thereby increase their attractiveness for potential offenders). All in all, this might suggest that the ambiguity vis-à-vis guardianship may be effective. Older people are in Mozambique less able to protect themselves at the individual level, *ceteris paribus*, possibly linked to their physical ability to protect themselves. In contrast, when persons live in households with an older household head this reduces as expected the risk of victimization (individual level regressions), probably linked in part to the experience embodied in age. This result is, however, only significant in the assault regression when aggregating the analysis to the household level.

In the individual level regressions we see (in accordance with the literature) that employed people have a higher probability of being victimized than their unemployed and studying counterparts. It suggests that exposure is indeed important, but we also note that this result is largely driven by the property crime (burglary) regression. We have tried to control for individual attractiveness in order to account for ambiguity in categorizing employment; but we cannot exclude that there are elements of attractiveness involved reinforcing the exposure link. Finally, single people are less victimized than their marital and divorced counterparts in Mozambique. This is opposite to what is normally found in the literature (Fajnzylber *et al.*, 2000), but our result holds for all types of victimization. This indicates that it is a common characteristic of the Mozambican case, probably related to greater exposure of these groups (noting that we have controlled for the fact that guardianship may be different for single people).

(iii) Proximity

Looking at proximity characteristics it appears that individuals, who are educated (measured vis-à-vis those without any education), are more likely to be victimized (except for assaults in the individual level regressions). This is in accordance with Gaviria and Pàges (2002). They argue that educated people are more proximate to crime than less educated people. In the Mozambican context, we do not find this convincing. It is more appropriate to suggest that being educated may transmit some kind of signal of being a more attractive target of crime, even if we highlight that we have tried to control for other factors of influence, including attractiveness. In any case, proximity and exposure (and probably also guardianship to some extent) are probably all playing a role here, the net effect being positive. Unemployment at the community level tends to increase the individual probability of being victimized. We believe this is a result that is largely and directly driven by the fact that living in areas with high community unemployment increases the probability of being assaulted.

Consistent with Bourguignon *et al.* (2003) and Soares (2004), we find indications of a positive relationship between inequality and the risk of victimization. When disaggregating by type of crime the relationship between inequality and assaults as well as burglary turn out insignificant. That is, victimization in terms of larceny is sizeable, significant and positively related to income inequality whereas the relationship turns both insignificant (for assaults and burglary) and of much smaller size (for assaults). This suggests that for these types of crime, proximity is less important, but it may also reflect to some extent that there are counter veiling effects from ambiguity in relation how to classify inequality. Finally, there is as expected an indication of high population density being associated with greater risk of being burglarized in Mozambique both in the individual and the household level analysis. In sum, proximity does indeed seem to matter, but the general picture is complex.

(iv) Guardianship

Finally, when we turn to determinants classified in the guardianship category, it is clear that household size yields different results depending on whether the analysis is done at the individual or at the household level. At the individual level, household size is significant and negatively related to victimization in accordance with typical lifestyle theories. Family members tend to look after each other and the household does seem to serve as a network of protection (Fajnzylber *et al*., 2000). We find a positive and weakly significant relationship between household size and victimization at the household level driven by property crimes. This suggests that the greater exposure caused by more members outweighs the guardianship effect, illustrating the unavoidable ambiguity in classifying a variable such as household size. Yet, this observation does illustrate the need for bringing economics and sociological approaches together.

A larger share of adult males in a household seems to reduce the risk of becoming a victim, and according to Table 5 this is especially so for burglaries, which is in all likelihood a clear guardianship effect. Turning to distance to the police station, this variable may be expected to have an effect on the probability of criminals being caught and therefore on the risk of victimization, in line with the economics approach. However, coefficients do not depict a clear picture, except that when distances are large, the risk of victimization decreases (noting that we have controlled for a series of other variables related to exposure, proximity and attractiveness). This may well suggest that police services are not altogether effective in terms of guardianship, but we wish to note that reverse causality may also be at play, i.e. police stations may be placed where criminal rates and the risk of victimization is higher. Our results correspond with the findings in Fajnzylber *et al.* (2000) for Latin America, and they do seem to suggest that the Fafchamps and Moser (2003) result for Madagascar, where crime increases with distance to urban centers, does not hold in the case of Mozambique.[13]

(b) Economic Loss from Victimization

As a measure of the severity of economic loss due to victimization (caused by property crimes, but excluding assaults), we use the logarithm of the ratio of monetary loss to yearly household income. For the 1,937 households with economic loss due to crime the mean ratio is 25.3% of yearly household income (9.2% excluding households reporting loss greater than yearly income). The similar mean for the entire population is 6.1%. The few observations with very high loss ratios (i.e. greater than 100% of yearly income) can however be included in the analysis since leaving them out has no effect on the estimates of the effect of household income on relative loss.[14] We are particularly interested in the relationship between relative loss and household assets and therefore include household consumption and household consumption squared as explanatory proxy variables together with household level variables included in Table 4 – except distance to police station. In the robustness analysis we show that the coefficients on consumption and consumption squared are not sensitive to which other explanatory variables are included. The selection equation is the same as the household level probit regression reported in Table 6. Provincial and rural/urban dummies together with community variables and distance to police station are excluded from the loss regression. Thus, we implicitly assume that these variables have no effect on the size of the economic loss from victimization when controlling for the household level variables described above.

Table 7 presents the results when controlling for selection and – for comparison – without controlling for selection.

Table 7: Loss from victimization

			Heckman		OLS	
Group	Classification	Variable	Mean	*t*-stats	Mean	*t*-stats
HH	Attractiveness	Real cons. (x100)	-1.915***	(3.98)	-1.249***	(3.10)
		Real cons. squared (x100)	0.003***	(3.71)	0.002***	(3.04)
		Durable goods: TV	-0.276	(1.10)	-0.265	(1.23)
		Durable goods: Radio	-0.294***	(2.57)	-0.286***	(2.60)
		Durable goods: Bicycle	-0.181	(1.49)	-0.141	(1.22)
HH	Exposure	HH Gender	-0.066	(0.42)	0.042	(0.28)
		HH Age	0.007*	(1.66)	0.006	(1.52)
		HH Employment status	-0.162	(0.64)	-0.136	(0.58)
HH	Proximity	HH Education 1	-0.467***	(2.78)	-0.342**	(2.22)
		HH Education 2	-0.562**	(2.37)	-0.305	(1.42)
		HH Education 3	-0.788***	(3.19)	-0.626**	(2.52)
		HH Education 4	-0.756**	(2.20)	-0.567*	(1.90)
HH	Guardianship	Household size	-0.100***	(4.42)	-0.078***	(3.64)
		Adult male share	-0.283	(0.74)	-0.268	(0.79)
Observations			1937	Uncensored (8409)	1937	
Wald test of independent equations			Chi2(1) = 5.69	P-value: 0.02		

Note: Dependent variable is the logarithm of the loss ratio (see main text). *, **, *** indicates significance at a 10%, 5% and 1% level, respectively. Base: Household head male, household head unemployed, household head no education. The coefficient estimates for the selection equation are not shown. Due to identical specification they are very close to the estimates reported for the household level probit in Table 5.

The test for dependent equations reported in Table 7 illustrates the need for using a selection equation framework. Our interest centers on the estimates of the consumption terms. The convex relationship has an estimated turning point of around 300 million Meticais (around 12,000 USD). This means that for all but a few households, the expected marginal loss ratio is decreasing in income. Poorer households, though less at risk of being victimized, lose a relatively large share of their income when they are victimized. This highlights that the vulnerability of the poor is also in this area of social and economic life a dimension that deserves careful attention by policy makers. Helping combat crime is of particular importance to the poorest which are hardest hit, in relative terms.

5. ROBUSTNESS ANALYSIS

In this section, we analyze the robustness of the results presented in Section 4. We investigate how the results are affected when one or more of the variables previously identified as potential determinants of victimization are omitted. In an analogous way we then turn to investigating whether our results on the relation between economic loss from victimization and income are robust to changes in the specification.

Following the literature on extreme bounds analysis, we run a systematic series of probit regressions to assess the sensitivity of the estimated coefficients to omission of specific groups of variables. Specifically, we divide the variables of Table 6 into two groups. One group contains what we denote as core variables. These are included in all subsequent regressions. The remaining variables belong to the second group – denoted secondary variables. The victimization dummy is then regressed on all possible linear combinations of the secondary variables including, in all the regressions, the full set of core variables. In other words, if the group of secondary variables is said to consist of *k* variables we perform 2^k-1 regressions.

The selection of core variables can of course take different directions. Yet, our main focus is on how individual and household characteristics affect the probability of being victimized. We therefore include as core variables all individual and household characteristics, excluding the three dummy variables indicating household possession of durable goods (due to their insignificance) and the distance to police station dummies because of the possible endogeneity mentioned above. This implies that the group of secondary variables is made up of 12 variables: Possession of durable goods (three variables), distance to police station (four variables), unemployment rate, inequality, average level of educational attainment, population density, and integration.

Table 8: Sensitivity analysis –summary statistics

Group	Classification	Variable	Max	Min	Mean	AvgSTD	PercSig	Perc+	Perc-	AvgT
Ind	Attractiveness	Income (x100)	1.301	1.100	1.196	0.446	1	1	0	2.682
		Income squared (x100)	-0.002	-0.003	-0.003	0.002	0.00	0	1	1.747
Ind	Exposure	Gender (x10)	-4.656	-4.745	-4.706	0.527	1	0	1	8.932
		Age (x100)	0.139	0.135	0.137	0.017	1	1	0	8.234
		Empl. status: Studying	-0.268	-0.287	-0.277	0.109	1	0	1	2.541
		Empl. status: Unemployed	-0.202	-0.230	-0.218	0.069	1	0	1	3.158
		Marital status: Married	0.562	0.542	0.552	0.079	1	1	0	7.017
		Marital status: Polygam	0.776	0.751	0.762	0.083	1	1	0	9.200
		Marital status: Cohabit	0.671	0.657	0.664	0.071	1	1	0	9.389
		Marital status: Divorced	0.848	0.821	0.836	0.108	1	1	0	7.740
		Marital status: Widow	0.781	0.758	0.770	0.111	1	1	0	6.940
Ind	Proximity	Education: Educ1 (x10)	2.593	2.459	2.531	0.537	1	1	0	4.713
		Education: Educ2 (x10)	3.426	3.237	3.342	0.723	1	1	0	4.623
		Education: Educ3 (x10)	3.484	3.102	3.246	1.001	1	1	0	3.241
		Education: Educ4 (x10)	4.195	3.727	3.931	1.253	1	1	0	3.137
HH	Attractiveness	Durable goods: TV (x10)	0.716	0.384	0.516	0.686	0	1	0	0.752
		Durable goods: Radio (x10)	-0.333	-0.506	-0.419	0.386	0	0	1	1.084
		Durable goods: Bicycle (x10)	0.484	0.286	0.379	0.426	0	1	0	0.888
HH	Exposure	HH Gender (x10)	3.342	3.075	3.195	0.585	1	1	0	5.465
		HH Age (x100)	-1.236	-1.287	-1.260	0.167	1	0	1	7.538
		HH Employment status (x10)	-0.690	-0.870	-0.773	0.703	0	0	1	1.100
HH	Proximity	HH Education 1 (x10)	-0.353	-0.577	-0.466	0.528	0	0	1	0.883
		HH Education 2 (x10)	0.509	0.064	0.271	0.647	0	1	0	0.417
		HH Education 3 (x10)	-0.009	-0.811	-0.454	0.968	0	0	1	0.470
		HH Education 4 (x10)	0.888	-0.082	0.343	1.250	0	0.98	0.02	0.276
HH	Guardianship	Household size	-0.034	-0.040	-0.037	0.007	1	0	1	5.104
		Adult male share	-0.345	-0.382	-0.365	0.131	1	0	1	2.784
		Distance to police 2	0.108	0.004	0.067	0.076	0	1	0	0.890
		Distance to police 3	0.072	-0.081	0.012	0.114	0	0.60	0.40	0.381
		Distance to police 4	-0.005	-0.192	-0.092	0.089	0.01	0	1	0.991
		Distance to police 5	-0.129	-0.230	-0.175	0.061	1	0	1	2.872
Com	Proximity	Unemployment rate	1.292	0.657	0.981	0.601	0.12	1	0	1.635
		Inequality	0.823	0.640	0.738	0.357	0.66	1	0	2.072
		Average educational level	0.152	-0.086	0.039	0.086	0	0.71	0.29	0.727
		Population density	0.015	0.005	0.010	0.007	0.21	1	0	1.442
		Integration	-1.137	-5.661	-3.717	4.698	0	0	1	0.791

Note: See table 5 for description of base household. Max, Min and Mean are respectively the maximum, minimum and mean value of the point estimate over all regression. AvgSTD and AvgT are averages over the standard deviations and t-values, respectively. PercSig gives the percentage times the coefficient was significant. Perc+ indicates the number of times the coefficient had a positive sign and analogously for Perc-.

Table 8 shows the summary statistics from this analysis. The first three columns show the maximum, minimum and average of the point estimate over all possible regressions discussed above. Column (4) shows the average standard deviation of the point estimates. Columns (5) to (7) contain the main results from the analysis. They reflect respectively the share of regressions where the point estimate is significant at the

5% level, the share with a positive point estimate (not necessarily significant), and finally the share of regressions with a negative point estimate. Column (8) gives the average *t*-value over all regressions.

The core variables are remarkably robust. Only the dummy variable for higher education change sign in any combination with the secondary variables, and except for individual income squared, education and employment status of the household head they are always significant at the 5% level. The square of individual income is ‘on average’ significant at the 10% level. Regarding the secondary variables the results are more mixed. Only the dummy variable indicating the longest distance to a police station is significant in all the regressions, while the smallest and second longest distance dummies retain the same sign in all regressions. The attractiveness variables (TV, radio and bike) also have the same sign in all regressions, though never significant. The unemployment rate, inequality and population density have sizeable shares of significant variables and together with integration the sign never varies. The average level of educational attainment is the only other variables where the point estimate switches sign depending on which secondary variables are included in the regression.

Turning to the robustness of the economic loss from victimization we use a similar methodology. More specifically – in light of our findings above – we retain the same selection equation (3) in all the regressions. For the regression equation (2) we use the variables of primary interest, income and income squared, and household size as core variables. The group of secondary variables consists of gender, age, education (four variables), employment and the share of adult males at the household level, and possession of a TV, radio and a bicycle. Proceeding as described above leaves us with 2,047 regressions.

Table 9 summarizes our findings for the three core variables.[15] As is evident from Table 9 the coefficient estimates of the income terms are very robust with respect to the specification of the regression equation. In all but a few regressions, they are clearly significant at the 5 percent level and the variability between regressions is small.

Table 9: Selection sensitivity analysis – summary statistics

	1	2	3	4	5	6	7	8
Variable	Max	Min	Mean	AvgSTD	PercSigni	Perc+	Perc-	AvgT
Household income (x10)	-1.539	-2.969	-2.122	0.751	0.98	0	1	2.9
Household income squared (x100)	0.005	0.003	0.003	0.001	1	1	0	3.0
Household size	-0.077	-0.140	-0.104	0.027	1	0	1	3.9

Note: Max, Min and Mean are respectively the maximum, minimum and mean value of the point estimate over all regression. AvgSTD and AvgT are averages over the standard deviations and t-values, respectively. PercSig gives the percentage times the coefficient was significant. Perc+ indicates the number of times the coefficient had a positive sign and analogously for Perc-. Selection equation as specified in Table 5. Secondary variables are (see main text): gender, age education (4 variables) and employment of household head together with share of adult males in the household.

6. CONCLUSIONS

This paper departed from the observation that economics and sociology point in somewhat different directions when trying to understand victimization in developing countries. Economics suggests that focus should be on the potential offenders and their evaluation of costs and benefits of antisocial behavior. This implies, for example, that higher incomes among potential victims are, *ceteris paribus*, expected to lead to a greater risk of victimization. Moreover, economists have with few exceptions paid no attention to variation by type of crime and individual characteristics of the victims. Sociology, on the other hand, has been more concerned with the individual characteristics of potential victims, including in general a more complicated set of explanatory categories. This approach has, however, been hampered by poor links between theory and data, inadequate measures of key concepts and failure to specify clear functional forms of the relationship between various sets of variables (Meier and Miethe, 1993).

While the sociological approach may to some reflect an idea that smacked of 'blaming the victim' as formulated by Meier and Miethe (1993), it is equally correct that the economic approach may potentially suffer from its trying to move forward on 'one leg' only, ignoring that there may be need to control for influential sociological variables before conclusions are drawn up. On this background, we proceeded to studying victimization in the case of Mozambique, relying on a unified analytical framework where both economic and sociological dimensions were allowed to speak. The choice of country case is justified both with a view to the fact that Mozambique belongs to the category containing the poorest countries in the world and because a novel and revealing, nationally representative household survey with relevant information on victimization at the individual, household and community levels has recently become available.

In discussing the identification of our explanatory variables, we highlighted that there are potential ambiguities involved in their categorization according to the four main factors emphasized in the sociological literature (attractiveness, proximity, exposure and guardianship). Individual level variables such as gender and age may, just to mention some examples, belong to several categories. Nevertheless, individual level variables do turn out to matter in a statistically significant and robust manner in our integrated analysis. Thus, the sociological approach helps ensure that potentially important differences between analyzing victimization from an individual and a household perspective are accounted for, and this does appear to matter. For example:

- Males have a higher probability of being victimized, but members of female headed households are more at risk of becoming a victim, probably due to lower levels of guardianship. Preventive policies geared towards supporting female headed households emerge as important, in contrast with what follows from the former observation.
- The guardianship variable 'household size' reduces the risk of becoming a victim when controlling for individual level characteristics. At the household level the conclusion is the reverse (driven by the burglary relationship). The bigger the household, the bigger the risk of victimization. There is in other words merit in viewing guardianship as related not only to public (police) initiatives but also to private aspects of deterrence, a dimension an offender view on victimization would not capture.

Our analysis also showed that analyzing victimization from an aggregate economic point of view misses that the explanatory factors behind different types of crime may differ. We brought out a variety of such examples, including:

- Inequality affects victimization, but only through larceny. Burglary and assault victimization are not affected. Similarly, it appears that lower unemployment will only decrease the assault rate, leaving burglary and larceny unaffected. It follows that a package of preventive policies geared towards employment creation and limiting inequality would affect larceny and assaults, but leave burglary unaffected. The general implication is that one-dimensional policy action is unlikely to be effective. To achieve several different victimization goals, a series of complementary policy measure are required. On the other hand, if policy makers assign priority to for example the assault issue (as has been the case in Mozambique more recently), attention should certainly be paid to the reduction of unemployment. In contrast, lower inequality will not *per se* make assaults go away.

In sum, by drawing upon the sociological literature we believe to have uncovered a more informed set of causal relationships than would have emerged in a 'pure' economic analysis. For sure, our 'map' for identifying individuals with the highest risk of victimization is more accurate than would otherwise have been the case. This is important both in general and with a view to the fact that policy makers may have specific groups of people or specific types of crime in mind as deserving priority attention.

At the same time, our integrated analysis confirmed the critical importance of approaching victimization from an economic angle. While exposure was mainly viewed from a sociological perspective, attractiveness, proximity and guardianship are all dimensions, which are narrowly related to economic considerations. Their empirical significance is also clear, including in particular:

- The probability of being victimized is increasing in income, but at a diminishing rate. The latter aggregate effect is mainly driven by the non-linearity between property crimes and income, but it is nevertheless clear that economic attractiveness matters across the board. The implication of our analysis is for example that both private and public guardianship measures should be promoted, taking due account of the importance of individual exposure. Moreover, it would appear that institutions offering effective insurance against victimization should be considered and promoted as an integral part of an overall strategy against crime and victimization.
- Poorer households are less at risk of victimization. They exhibit lower attractiveness. They are also more vulnerable. They suffer proportionally larger losses when crime occurs. It is not straightforward to derive policy implications from this result, as a complex interpersonal metric of utility based on fractions of income are looming in the background. Yet, we do find the result of interest.

It merits careful attention by policy makers as well as further study, and we suggest that this characteristic is likely to be linked to poorer people possessing more limited guardianship than better off people. This implies *inter alia* that the design of criminal policy should take careful account of the distribution of preventive resources across social groups.

In our review of the importance of employment status (classified under exposure) and education (classified under proximity) we also noted that there are in all likelihood elements of attractiveness involved in explaining the impact of the specific empirical variables included, especially in the education case. Thus, our analysis suggests that being employed and educated increase the risk of victimization, over and above the impact captured through individual income, which we controlled for. This might induce a negative incentive towards seeking employment outside the home, so policy makers may wish to focus attention on safety while workers are commuting alongside measures addressed to combat attractiveness crime. Furthermore, education does appear to increase victimization risk. It is not obvious what should be done about this, but policy makers should be aware that this carries with it the risk that better educated workers may eventually migrate or change behavior. Different kinds of insurance may be one way of dealing with this problem.

In sum, we believe to have justified that it is indeed possible to increase our understanding of victimization following the advice of Meier and Miethe (1993) of incorporating sociological victimization theories into a unified theory of crime where functional relationships are clearly specified. We acknowledge the ambiguities involved, but wish to reiterate by way of general conclusion that the robust causal patterns identified in this paper can serve both as an input into policy formulation and as a more comprehensive starting point for future victimization research.

NOTES

[1] For further country specific context on Mozambique see for example UN (2005). Reference can also be made to the following website maintained by the UN Office for the Coordination of Humanitarian Affairs http://www.irinnews.org/frontpage.asp?SelectRegion=Southern_Africa&SelectCountry=Mozambique.

[2] See Fafchamps and Moser (2003) and Demombynes and Özler (2005) for two important recent studies where a broad range of explanatory variables are relied on in explaining crime in respectively Madagascar and South Africa. While these two articles study crime we focus in this paper on the other side of the criminal event, victimization.

[3] Looking at the available evidence on guardianship factors it is important when interpreting the results to distinguish between 'aggregate' household level and 'micro' individual level studies. Moreover, Meier and Miethe (1993) note that few studies of guardianship have included sufficient controls for other factors influencing victimization risk. We focus here on the few existing individual level studies, which are closer in nature to the objective of the present paper. See for example Gaviria and Pàges (2002) for a recent analysis at the household level.

[4] Of households being victimized 90.2% suffered a property loss. We highlight that it is only these 90.2% of the victimized, who enter the selection equation (3) with a loss. The welfare loss suffered due to non-property crime can of course be substantial, but an analysis hereof is well beyond the scope of the present paper.

[5] Mozambique has 10 administrative provinces (Cabo Delgado, Niassa, Nampula, Sofala, Zambézia, Manica, Tete, Gaza, Inhambane and Maputo) in addition to Maputo city.

[6] Exchange rate 1US$ = 25,000 Meticais or 1Metical = 0.00004 US$.

[7] Educ0 = Never went to school; Educ1 = Went to school but no grade obtained; Educ2 = literate and primary 1st completed; Educ3= primary 2nd completed; Educ4 = higher and technical educations.

[8] Thus, average education is calculated as (# of persons with edu0=1x0 + # of persons with edu1=1x1…)/# of persons.

[9] See MPF et al. (2004) for a detailed description of the construction of real consumption.

[10] The categorical values correspond to the time it takes to reach the police station on foot: 1 = 0-29 min.; 2 = 30-44 min.; 3 = 45-60 min.; 4 = 60-119 min.; 5 =120+ min. When answering the question on distance to the police station households could choose mode of transportation. For some households distance on foot had to be constructed. This was done by giving all households that did not answer 'on foot', the corresponding average value for households in the same enumeration area that answered 'on foot'. Moreover, we assumed that households which reported that they had more than 30 minutes transport by car to the nearest police station has been categorized under category 5. Similarly, households responding that they had more than 60 minutes by bicycle are put into category 5.

[11] In the sample, there are a few cases of people, who suffered more than one offense. They are however so few that they do not affect our overall results. In this paper we focus on whether people were victimized or not.

[12] Consumption figures are only available at the household level. In the analysis we also used household consumption to instrument household income. In order to capture the non-linearity in household income we created instruments interacting household real consumption with continuous exogenous regressors in our specification (household real consumption squared was a weak instrument for squared income according to our first stage regressions). We were left with eight instruments, exogenous by construction. The qualitative results reported in Table 6 do not change using this approach.

[13] We included the control variable 'distance to urban centres and markets' and this did not change our distance to police station result. Moreover, we found an indication of distance to urban centres and markets being negatively related to the risk of being victimized in all specifications.

[14] We tried several different 'cut-off' values for the relative loss (i.e. only using observations with relative loss less than 10, 30 or 50% of yearly income). All estimations produced results very similar to using the full sample.

[15] A full set of regressions is available upon request.

REFERENCES

Becker, G.S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, *76*, 169-217.

Bourguignon, F., Nunez, J. and Sanchez, F. (2003). A Structural Model of Crime and Inequality in Columbia. *Journal of European Economic Association*, *1*, 440-449.

Cohen, L.E. and Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*, 588-607.

Cohen, L.E., Kluegel, J.R. and Land, K.C. (1981). Social Inequality and Predatory Criminal Victimization: An Exposition and Test of a Formal Theory. *American Sociological Review*, *46*, 505-524.

Demombynes, G. and Özler, B. (2005). Crime and Local Inequality in South Africa. *Journal of Development Economics*, *76*, 265-292.

Dgedge, M., Novoa, A., Macassa, G., Sacarlal, J., Black, J., Michaud, C. and Cliff, J. (2001). The Burden of Disease in Maputo City, Mozambique: Registered and Autopsied Deaths in 1994. *Bulletin of the World Health Organization*, *79*, 546-552.

Ehrlich, I. (1973). Participation in Illegitimate Activities: A Theoretical and Empirical Investigation. *Journal of Political Economy*, *81*, 521-565.

Fafchamps, M. and Minten, B. (2006). Crime, Transitory Poverty, and Isolation: Evidence from Madagascar. *Economic Development and Cultural Change*, *54*, 579-603.

Fafchamps, M. and Moser, C. (2003). Crime, Isolation and Law Enforcement. *Journal of African Economies*, *12*, 625-671.

Fajnzylber, P., Lederman, D. and Loyaza, N. (2000). Crime and Victimization: An Economic Perspective. *Economia*, *1*, 219-302.

Fajnzylber, P., Lederman, D. and Loyaza, N. (2002). Inequality and Violent Crime. *Journal of Law and Economics*, *45*, 1-40.

Fajnzylber, P., Lederman, D. and Loyaza, N. (2002). What Causes Violent Crime? *European Economic Review*, *46*, 1323-1357.

Gaviria, A. and Pages, C. (2002). Patterns of Crime Victimization in Latin American Cities. *Journal of Development Economics*, *67*, 181-203.

Hindelang, M.J., Gottfredson, M.R. and Garofalo, J. (1978). *Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization*. Cambridge: Ballinger.

Instituto Nacional de Estatística (2004). *Inquérito Nacional aos Agregados Familiares Sobre Orçamento Familiar 2002/3*. Maputo: INE.

Lederman, D., Loyaza, N. and Menedez, A.N. (2002). Violent Crime: Does Social Capital Matter? *Economic Development and Cultural Change*, *50*, 509-539.

Meier, R.F. and Miethe, T.D. (1993). Understanding Theories of Criminal Victimization. *Crime and Justice*, *17*, 459-499.

Ministry of Planning and Finance (MPF), International Food Policy Research Institute (IFPRI) and Purdue University (2004). *Poverty and Well-being in Mozambique: The Second National Assessment*, Southern African Regional Poverty Network (SARPN) Working paper.

Ministry of the Interior (2004). Statistical Tables on Criminal Situation. Mimeo, Maputo.

Nizamo, H., Meyrowitsch, D.W., Zacarias, E. and Konradsen, F. (2006). Mortality Due to Injuries in Maputo City, Mozambique. *International Journal of Injury Control and Safety Promotion*, *13*, 1-6.

Smith, D.A. and Jarjoura, G.R. (1989). Household Characteristics, Neighborhood Composition and Victimization Risk. *Social Forces*, *68*, 621-640.

Soares, R.R. (2004). Development, Crime and Punishment: Accounting for the International Differences in Crime Rates. *Journal of Development Economics*, *73*, 155-184.

UN (2005). *Crime and Development in Africa*. United Nations Office on Drugs and Crime, Vienna.

WDI (2005). World Development Indicators. CD-Rom.

WHO (2002). *World Report on Violence and Health*. WHO, Geneva.

Zenou, Y. (2003). The Spatial Aspects of Crime. *Journal of European Economic Association*, *1*, 459-467.

APPENDIX

The sociological categories in the routine activity framework are defined as follows (Cohen *et al.*, 1981):

a) Exposure

The physical visibility and accessibility of persons or objects to potential offenders at any given time or place.

b) Proximity

The physical distance between areas where potential targets of crime reside and areas where relatively large populations of potential offenders are found.

c) Guardianship

The effectiveness of persons or objects in preventing violations from occurring, either by their presence alone or by some sort of direct or indirect action.

d) Target Attractiveness

The material or symbolic desirability of persons or property targets to potential offenders, as well as the perceived inertia of target against illegal treatment.

e) Properties of Crimes

The features of specific crimes that act to constrain strictly instrumental actions by potential offenders. For example, many larcenies are less difficult to commit and require less knowledge of victim routine activities than do burglaries.

Orphans and Discrimination in Mozambique

Mikkel Barslund[1]
University of Copenhagen, Development Economics Research Group (DERG)

Channing Arndt
Purdue University

Virgulino Nhate
Ministry of Planning and Development (MPD), Mozambique

Katleen Van den Broeck
University of Copenhagen, Development Economics Research Group (DERG)

July 2005
Revised: September 2006

Abstract

An increasing number of orphans is one of the most unfortunate consequences of the AIDS pandemic in Africa. High HIV prevalence in Mozambique motivates this study. The projected 800,000 AIDS related adult deaths over the period 2004-2010 will leave significant numbers of orphans in their wake. Of these, many will reside in families where the household head is not their biological parent. We analyze the extent of discrimination in resource allocation within households against children who are not the biological descendant of the household head in Mozambique. Both Deaton's outlay equivalence method and the related Engel curve approach are applied using a nationally representative survey from 2002/03. Results point to discrimination in the intra-household allocation of resources against children that are not direct biological descendants of the household head *in poor households*. Discrimination is identified in both the rural and urban sub-samples. In non-poor households, resource allocations between biological and non-biological children do not differ significantly.

Keywords: Mozambique, AIDS, Orphans, Outlay equivalence, Engel curves, Discrimination

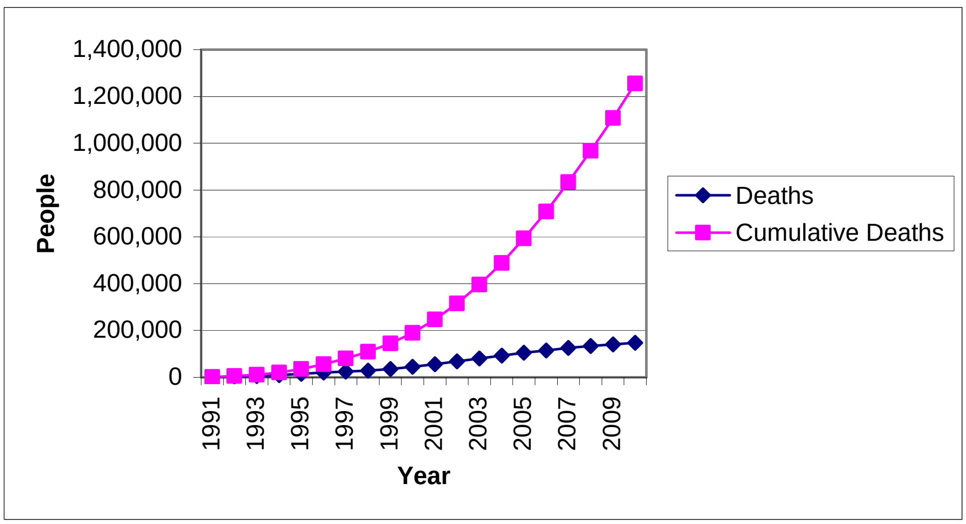
JEL classification: J13, D13

[1] Address for correspondence: Mikkel.Barslund@econ.ku.dk. This is a much expanded version of an article with the same title appearing in *American Journal of Agricultural Economics 88, No. 5*.

1. Introduction and background

High HIV prevalence in many parts of Africa motivates this study. For example, in Mozambique, the prevalence of HIV among adults aged 15-45 years in 2005 is estimated to be about 16.2 percent and is projected to climb (INE et al, 2004). Figure 1 illustrates estimated annual and cumulative adult AIDS deaths from 1991 to 2010. As shown in the figure, nearly 400,000 Mozambican adults were estimated to have died of AIDS related causes by 2003, the year of the survey. Worse, AIDS deaths are projected to grow rapidly through the rest of the decade. In fact, more than twice as many adults are projected to die in the period 2004-2010.

Figure 1: Adult AIDS deaths in Mozambique.



Due to the tendency of the pandemic to strike young adults, AIDS related deaths leave significant numbers of orphans in their wake. A demographic and health survey (DHS) carried out in 2003 found that, for children under 15 years of age, approximately one child in ten had been orphaned (paternal, maternal, or dual) (INE, 2004). Demographic projections based on a time series of HIV prevalence data estimate an orphaning rate of more than 16% in 2003 for children below 18 years of age (INE et al., 2004). The difference in age categories (0-14 versus 0-17) explains part, but not all, of the difference in the rates. Reluctance on the part of surveyed households to admit the death of the biological mother of the child could account for the remaining difference and would explain the relatively low ratio of maternal to paternal orphans in the DHS data relative to the demographic projections. Overall, despite some differences in quantity and nature, both sources of data point to significant orphaning. Furthermore, the number of orphans appears set to climb dramatically.[2]

Mozambican national policy specifically favors the integration of orphans into substitute or extended families (GM, 2004). This mirrors policy in other highly afflicted African countries such as Botswana, Zimbabwe, Zambia, and Uganda (UNUSIDA, 1999). The approach has the advantage that orphans remain integrated within a family. This also implies that the resources available to families that accept orphans and the allocation of those resources within the household become of policy interest.

Generally, resources are exceedingly tight within Mozambican households. In 1996-97, 72% of children (aged 0-17) lived in households characterized as absolutely poor using a consumption based metric. By 2002-03, this share had improved considerably but remained very high with 58% of all children living in households characterized as absolutely poor. Although non-biological children tend to concentrate in households that are on average slightly better off (Nhate 2004), resource availability remains distinctly limited. Because of the severe limitation of available resources, difficult decisions regarding resource distribution have to be made. As noted by Hamilton (1964) biological bonds are important in the distribution of resources within the household implying the potential for discrimination against non-biological children. Some evidence of discrimination in Mozambique has already been found. Nhate (2004) found that children that are not biological descendants of the household head were significantly less likely to attend school in both rural and urban areas holding constant other factors. A qualitative survey commissioned by the World Bank building on interviews with parents as well as school managers found evidence of orphans more often being kept at home for domestic tasks (World Bank, 2004).

[2] In 2003 there were around 470,000 maternal orphans in Mozambique. In 2010 the number is expected to reach 900,000 (National Statistics, Mozambique).

It is important to point out that, similar to Nhate (2004), this analysis compares children who are biological versus non-biological descendants of the household head rather than orphans specifically. The available database on consumption does not permit the identification of 'true' orphans. For the age group 15 and under, about one child in four is not the biological descendant of the household head and the AIDS pandemic can be expected to add considerably to this group of children over the next decade.[3]

Nevertheless, an important subset of children who are not the biological descendant of the household head is not likely to be at risk for discrimination. In particular, weak geographic coverage of higher primary and secondary school causes some families living in areas without access to schools to send children to live with relatives or friends in areas where primary schools are available. It may be plausibly assumed that children who are sent by their parents to live with another family in order to attend school are less likely to be discriminated against than the target group of interest children, such as orphans, who are forced into fostering due to some negative shock. As we are not capable of distinguishing between these two groups of children in our sample, the results obtained here could be viewed as a lower bound on the degree of discrimination within families against the target group of interest.

A number of other African countries are facing large increases in orphans due to the AIDS pandemic. Therefore, the issue of how orphanhood affects important outcomes, like e.g. schooling, has recently been studied in a number of countries. Case et al. (2004) use DHS surveys for 10 sub-Saharan African countries to investigate if orphans are less likely to be enrolled in school compared to non-orphans once wealth is controlled for. They find that although orphans are on average poorer than non-orphans, this wealth gap does not explain the lower enrollment rate observed for orphans. Further, they find that closeness of the caretakers biological ties with the orphan affect the degree to which orphans are 'under enrolled'. The closer the biological ties, the smaller is the enrollment gap. Sharma (2006) uses a panel for Malawi to assess the impact of orphanhood on schooling and finds that

[3] For an unknown but likely substantial fraction of these children, the circumstance of being fostered reflects stress, such as the death of a parent, resulting in placement of the child with another family. We hypothesize that these children are at risk of being discriminated against.

they are less likely to complete as many years as non-orphans. Beegle et al. (2006) find that orphans in Northwestern Tanzania lose out on schooling relative to non-orphans. Also, there are indications that orphanhood is associated with height deficiencies although longer spanning longitudinal data is needed to see if the effect is permanent. Case and Ardington (2006) look at data from South Africa's KwaZulu-Natal province and find a negative effect on schooling outcomes from being a maternal orphan, but no effects for paternal orphans. Evans and Miguel (2005) find similar results in their analysis from a district in rural Kenya.

In the present study we focus on possible discrimination in intra-household allocation of resources in general and related to expenditure on education and children's clothing. The issue of how resources are allocated within households has become an important focus of poverty analysis. Unfortunately, where individual consumption data is not available, intra household resource allocations are difficult to measure directly; and standard household consumption surveys rarely attempt to do so. To counter this difficulty, indirect measures have been developed. In particular, Deaton (1989a) proposed a method, labeled 'outlay equivalence', whereby spending on children is measured indirectly via spending on adult goods. The intuition is that the addition of a child should imply increased spending on goods for children. If total consumption levels are inflexible (i.e. the budget constrain is binding), the budget constraint must then imply reduced spending on adult goods. Since, particularly in developing countries, pure adult goods are often easier to identify than pure children's goods, the method has become popular. Where aggregate household expenditure has been available for pure children goods (i.e. education and clothing expenditure) a method known as the Engel curve approach has been applied. The intuition behind this approach is that once total expenditure and household size have been controlled for the composition of the households' children in biological and/or non-biological children should not influence the expenditure share on children's goods. Below we utilize both designs in order to detect possible discrimination.

A large number of applications have often focused on whether female children are discriminated against relative to their male counterparts. Although, some studies have found significant differences between boys and girls in Asia using either or both of the two methods (i.e. Miller, 1981; Deaton, 1989b; Behrman, 1990; Gibson and Rozelle, 2004; and Kingdon, 2005), researcher have often been puzzled by the failure of the methods to detect differences even where their presence are strongly insinuated by other indicators (Case and Deaton, 2002). In African countries, studies tend not to find statistically significant evidence of discrimination against girls (Deaton, 1989b; Haddad and Reardon, 1993). Based on analysis of an individual level data set, Kingdon (2005) argues that two effects could account for the failure of the Engel curve methodology. First, aggregation might dilute discrimination sufficiently to render it undetectable, and second, the functional form usually estimated may not be up to the job. She suggests a modification which works well in her application and we pursue it here.

The remainder of this paper is structured as follows. Section 2 discusses both data and both the outlay equivalence and Engel curve methods. Section 3 presents and compares the results. The final section presents conclusions.

2. Data and Methodology

2.1. Methodology

2.1.1. Outlay equivalence analysis

The first part of the analysis of orphan discrimination follows the methodology outlined by Deaton (Deaton, 1989b. See also Gibson and Rozelle, 2004, for a recent application). The methodology is briefly described below.

The first step is to identify a bundle of adult goods and test if the proposed adult goods are demographically separable from children demographic groups, meaning that children should have only an income effect but not substitutability effect. This is equivalent to testing the joint significance of the coefficients on the children demographic variables in the following linear model:

$$p_i q_i = \alpha_{0i} + \alpha_{1i} X_G + \sum c_{ij} n_j + d_i . z + \varepsilon_i \tag{1}$$

where $p_i q_i$ is expenditure on the candidate adult good *i*, X_G is total expenditures on adult goods, n_j is the number of members in demographic category *j*, *z* is a vector of other explanatory variables included in the model, and ε_i is the error term. Given total expenditures on adult goods, children should not influence the distribution of spending across adult goods.[4] If the goods included are really adult goods, children will not have any effect in equation (1). Therefore, the coefficients, c_{ij}, should be jointly different from zero for demographic groups related to children in order for demographic separability to hold.

Following the test of existence of adult goods using equation (1), we calculate the outlay equivalent ratios. The equivalent ratio, π_{ir}, for a normal adult good *i* and demographic category *r*, can be calculated as:

$$\pi_{\mathrm{ir}} = \frac{\partial(p_i q_i)/\partial n_r}{\partial(p_i q_i)/\partial x} * \frac{n}{x} \tag{2}$$

where *x* is total household expenditure, *n* total household size and n_r number of household members of demographic category *r*. π_{ir} measures the effect of the addition of a member of demographic type *r* on total expenditure on good *i* in terms of the change in total expenditure that would be necessary to produce the same effect on adult good demand with this change presented as a share of per capita expenditure. For adult goods, one would expect an additional child to have the same effect as a reduction in income (measured by total expenditure) and hence a negative value for π_{ir}.

Following Deaton, (1989a), the expenditure outlay equivalent ratios in (2) can be calculated using the coefficients estimated from a standard Engle curve, specified in the following way:

$$w_i = \frac{p_i q_i}{x} = \alpha_i + \beta_i \ln(\frac{x}{n}) + \eta_i \ln n + \sum_{j=1}^{j-1} \gamma_{ij} (\frac{n_j}{n}) + \delta_i . z + \mu_i \tag{3}$$

[4] In the empirical analysis a Durbin-Wu-Hausman test was done to test the endogeneity of total adult goods expenditures. If the test of exogeneity was rejected, 2SLS with total expenditure as an instrument was employed.

where w_i is the budget share of the i^{th} adult good, and z is a vector of control variables. The estimated parameters in equation (3) can be used to calculate:

$$\pi_{ir} = \frac{(\eta_i - \beta_i) + \gamma_{ir} - \sum_{j=1}^{j} \gamma_{ij} (\frac{n_j}{n})}{\beta_i + w_i}$$

(4)

Estimates of these ratios are obtained by substituting the parameters with their respective estimates (from equation 3) and substituting w_i and the fraction n_j/n by their mean values in the sample. The test of equal treatment between biological and non-biological children in each age group and for all adult goods is equivalent to testing the hypothesis

$$Ho : \pi_{ij} = \pi_{ik}$$

(5)

where j refers to biological children and k to non-biological children in the same age group. The test is simple to implement by testing the equality of the demographic coefficients in (3) via a t-test.

Standard errors for the π ratios were derived using the non-parametric bootstrap methodology (Cameron and Trivedi, 2005). The bootstrap method involves drawing synthetic samples of the same size as the original sample by sampling *with replacement* from the original sample.[5] Hence, an arbitrary observation from the original sample may appear not at all, once, or multiple times within a given synthetic sample. Regressions using equation (3) were run on 1000 synthetic samples and the π ratios were calculated in each instance. Standard errors are then calculated from this sample of 1000 π ratios. An alternative approach to deriving standard errors for the π ratios is described in Deaton et al. (1989a). The bootstrap approach has the advantage of accommodating the non-linear nature of the π ratios as a function of the estimated parameters as well as allowing for clustered data sampling.

[5] The method for drawing synthetic samples paralleled the approach for drawing the original sample. In particular, since data is clustered, clusters are sampled rather than observations.

An additional test for demographic separability is available as a further check on the validity of the chosen adult goods (Deaton et al., 1989a). If separability holds then the addition of a child will act as an equally sized income effect on all adult goods, thus, the outlay equivalent ratios should be equal across goods for a given (children) demographic group *k*. If π_k denotes the vector of π-ratios for all *M* adult goods in demographic group *k*, and $\bar{\pi}_k$ the average of the π-ratios over the *M* adult goods, then testing equality of the π-ratios is equivalent to testing the *M-1* linear restrictions: $\pi_{ik} = \bar{\pi}_k$ for all $i = 1,..,M$.[6] To form the appropriate Wald statistics, construct the matrix $A = I - (ii'/M)$ where *I* is an *MxM* identity matrix and *i* is a *Mx1* unit vector. The set of linear restrictions can now be expressed as $A\pi_k$. The Wald statistics are distributed as χ^2 with *M-1* degrees of freedom under the null hypothesis and given by

$$W_r = \pi_k' A' \left[A' V(\pi_k) A \right]^{-1} A\pi_k . \quad (6)$$

$V(\pi_k)$ is the variance-covariance matrix for the *M* π-ratios for demographic group *k*. The empirical variance matrix is obtainable from the bootstrapped sample of π-ratios.

2.1.2. Engel curve approach

The Engel curve methodology relies on the estimation of Engel curves for children's goods controlling for household demographic composition. In absence of differential treatment, the share of biological children (in total household size) should have the same effect on the expenditure share as the share of non-biological children in the same age group. Consequently, the test of equal coefficients provides a test for equal treatment. As is aptly summarized in Kingdon (2005) previous studies have estimated Engel curves of the form defined in equation (3) by OLS, despite a large number of households having zero purchases on the good. An issue which also pertains to the present data set as evident from Table 1. Using an individual level data set from India showing discrimination against girls, Kingdons analysis suggests that accounting for zero consumption might improve the Engel curves ability to detect discrimination. We follow her recommendation and specify a hurdle model of the form (Wooldridge, 2002):

---

[6] Note the last restriction is redundant, hence the *M-1* degrees of freedom.

$$P(s = 0 \mid Z) = 1 - \Phi(Z\gamma) \tag{7}$$

$$\log(s) \mid (Z, s > 0) \sim N(Z\beta, \sigma^2) \tag{8}$$

The share of children goods in total expenditure is denoted s, Z is a vector including the full set of explanatory variables from equation (3), γ and β are conformable coefficient vectors. Φ is the standard normal cumulative distribution function. Embedded in the formulation (7) and (8) is that whether a household has expenditure on the analyzed good or not is determined by a probit model. For households, where the expenditure share is positive, the logarithm of the expenditure share follows a normal distribution. Maximum likelihood estimation is laid out in Wooldridge (2002) and is equivalent to first estimate (7) by a probit, and then OLS estimation of the logarithmic expenditure share on explanatory variables. Using probabilities of the lognormal distribution it can be shown that the unconditional expectation of s and the marginal effect of a variable $z \in Z$ are given by:

$$E(s \mid Z) = \Phi(Z\gamma)\exp(Z\beta + \sigma^2 / 2) \tag{9}$$

$$\frac{\partial E(s \mid Z)}{\partial z} = \exp(Z\beta + \sigma^2 / 2)\left[\gamma\phi(Z\gamma) + \beta\Phi(Z\gamma)\right] \tag{10}$$

Similar to the method of obtaining standard errors for the outlay equivalence ratios we adopt a bootstrap approach to allow for the clustered nature of the data set.[7]

2.2. Data

The data used in this study comes from the Mozambique national household survey about living conditions (IAF) undertaken by the National Institute of Statistics (INE). This survey is representative at the national, provincial, and rural/urban levels. The survey was conducted between July 2002 and June 2003. The year long interview period was programmed in order to capture potential seasonality in household consumption. The survey covered 8,700 households corresponding to about 44,000 individuals. Enumerators visited each household at least three times over the period of a week to collect consumption and other information. The survey collected expenses on 863 different goods (food and non-food).

[7] A 1,000 random draws with replacement are performed. For each draw marginal effects are calculated for the relevant demographic groups.

For the purpose of the outlay equivalence method we are specifically interested in identifying adult goods, i.e. goods that children do not consume. If a good is a true adult good, the addition of a child (with the concomitant expenses necessary to support that child) acts in a manner analogous to a reduction in income with respect to spending on adult goods. For the case of normal goods, consumption should decline. Six candidate adult goods were identified including: adult clothes; alcoholic beverages (inside and away from home); personal care (hair treatment, nail products, lipstick, "mulala", lotion, etc,); public and private transportation services; tobacco; and food and soft drinks away from home.

For the Engel curve approach we identify two children goods. Engel curves are estimated separately for these two goods. If no discrimination is present then the composition of the children in a household should not affect the expenditure share. The survey included questions on yearly expenditure on primary education (EP1 and EP2) – including school uniforms – and monthly expenditures on children's clothing and footwear.[8] While a monthly recall question should give a good estimate of households' expenditure on children's clothing and footwear (at least for items deemed by the respondent to belong to these groups), the yearly recall question on education expenditure may not be particularly accurate. At the time of the survey the structure of school fees in Mozambique was rather erratic with enrollment, examination, graduation and add-hoc fees (World Bank, 2004).[9] Anecdotal evidence also points to some extent of informal fees being asked for children to be promoted to the next class. The questionnaire covers 'fees' and it is not clear if all types of fees are included. In particular, it is doubtful if informal/illegal fees are covered. Even in the best of circumstances, fees are paid irregular throughout the year, perhaps making precise recall difficult. Nevertheless, since we are interested in all indicators which can illuminate if discrimination is present, educational expenditure is used cautiously.

Note the advantage of comparing biological and non-biological children of the same gender versus boys and girls. Whereas differences found in boy-girl comparisons using adult goods or children's goods can be due to differential norms regarding the dressing or teaching of

[8] In the Portuguese questionnaire there are two entries: 'Calçado para criança' (footwear) and 'Vestuario para criança' (clothing).
[9] School fees have since been abolished for primary schools.

boys and girls, this does not apply to differences between biological and non-biological children of the same gender.[10]

Fejl! Henvisningskilde ikke fundet. presents relevant data for this study. The analysis will be conducted for rural and urban zones in order to capture differential characteristics of rural and urban families. Furthermore, the analysis will also be performed separately for poor and non-poor household. Poor households are defined as those living below a poverty line that reflects basic needs (Ministry of Planning and Development et al, 2004). Resource constraints in these households living below the poverty line are severe and may influence intra-household resource allocation decisions. In other words the budget constraint is more likely to be binding. Finally, following general practice, 1046 households without any children and 538 households with only a single household member were excluded from the sample leaving a total of 7116 households with at least one child present in the final sample.

The average budget share of the candidate adult goods as a group is 13 percent. Tobacco and adult clothes are the goods that have the highest share among all adult goods. Each of these two goods represents about 4 percent in total of expenditure. The groups "food and soft drinks consumed away from home" and "personal care" represent small shares of total expenditures (0.2 and 0.6 percent, respectively). Generally, budget shares for adult goods are higher in urban than in rural areas. In urban areas, these goods represent 15 percent of total expenditures compared with 11 percent in rural areas. Differences between rural and urban samples are most marked with respect to transportation and personal care products. Overall, the shares for adult goods observed in Mozambique are similar to values found in other developing countries. In Burkina Faso, for example, Haddad et al. (1993) found that these goods represented 15 percent of total expenditures. In Papua New Guinea, Gibson and Rozelle (2004) found that candidate adult goods represented 12 percent of total of expenditure.

[10] There could still be differences between biological and non-biological children. For example, if biological children are more astute due to better circumstances as a small child, and therefore work and play more, clothing expenditures might be higher for biological than non-biological children without the difference being attributable to discrimination. However, we believe the size of such effects to be negligible.

Table 1. Descriptive statistics (mean values).

Variables	Full sample	Urban sample	Rural sample	Full Sample	Rural Sample
Expenditure share:					
Total adult goods	0.125	0.153	0.114	7,116	3,805
Of which: Alcohol	0.010	0.011	0.010	7,116	3,805
Tobacco	0.043	0.049	0.041	7,116	3,805
Adult clothing	0.043	0.043	0.043	7,116	3,805
Transportation	0.022	0.035	0.016	7,116	3,805
Food and soft drinks consumed away from home	0.002	0.004	0.001	7,116	3,805
Personal care	0.006	0.011	0.004	7,116	3,805
Education (EP1, EP2, S1, S2)	0.007	0.012	0.003	5694	2,957
Education (EP1, EP2, S1, S2) (households with expenditure>0)	0.009	0.013	0.005	4538	2080
Children's clothing	0.005	0.005	0.005	5694	2,957
Children's clothing (households with expenditure>0)	0.0487	0.050	0.047	594	294
Log of total household expenditures	9.151	9.496	8.851	7,116	3,805
Log of household size	1.556	1.632	1.491	7,116	3,805
Demographic composition: Proportion of .. (group)					
(1) Biological children aged 0-5 years	0.150	0.128	0.170	7,116	3,805
(2) Non-biological children aged 0-5 years	0.040	0.042	0.038	7,116	3,805
(3) Biological children aged 6-10 years	0.104	0.098	0.109	7,116	3,805
(4) Non-biological children aged 6-10 years	0.032	0.032	0.032	7,116	3,805
(5) Biological children aged 11-15 years	0.079	0.082	0.076	7,116	3,805
(6) Non-biological children aged 11-15 years	0.031	0.035	0.028	7,116	3,805
(7) People aged 16-20 years	0.110	0.130	0.092	7,116	3,805
(8) People age d 21-25 years	0.075	0.087	0.065	7,116	3,805
(9) People aged 26-59 years	0.320	0.320	0.319	7,116	3,805
(10) People with more than 60 years of age	0.059	0.045	0.072	7,116	3,805
Proportion of households headed by women	0.252	0.266	0.239	7,116	3,805
Educational level of household head	1.106	1.884	0.432	7,116	3,805
The mean age of the household head	42.937	42.696	43.146	7,116	3,805
Proportion of household with a person working in agriculture or fishing	0.756	0.503	0.976	7,116	3,805
Proportion of household with a person working in commerce	0.180	0.313	0.0065	7,116	3,805
Proportion of household with a person working in the services sector	0.142	0.270	0.030	7,116	3,805

Source: IAF2002/03 as described in the main text.

Educational expenditures include fees for lower primary one (EP1, grade 1-5), upper primary (EP2, grade 6-7), lower secondary (SE1, grade 8-10) and upper secondary (SE2, grade 11-12) for both public and private schools together with spending on school uniforms. The expenditure share is generally low, and a large number of households do not have any educational expenditure. This reflects in part children not being enrolled in school, and partly that non-paying children are sometimes allowed to stay in school (World

Bank, 2004).[11] Expenditure on children's clothing is overall quite low due to the large number of households not purchasing in the recall period. However, for households incurring expenses it rises to around 5 percent.

To study the influence of demographic effects the household members were divided into 10 groups. The first six groups, comprised of people less than 15 years of age, are the ones of primary interest for this study. The remaining four groups include adults that are used for the confirmation of the presence of adult goods. The groups were divided in the following way: biological children aged 0-5 years (group 1), non-biological children aged 0-5 years (group 2), biological children aged 6-10 years (group 3), non-biological children aged 6-10 years (group 4), biological children aged 11-15 years (group 5), non-biological children aged 11-15 years (group 6). For the rest of the age groups, the categorizations were as follows: people aged 16-20 years (group 7), people aged 21-25 years (group 8), people aged 26-59 years (group 9), people 60 years and older (group 10). The largest demographic category is biological children aged 0-5 years in rural areas. Of the total rural population, nearly 17 percent are biological children aged 0-5 years old. In urban areas, the same group represents about 13 percent of the total population. Non-biological children in the same age group represent only about 4 percent of the total population.

In the study sample, about 25 percent of the households are headed by women with a slightly higher percentage in urban areas compared to rural areas (27 and 24 percent respectively). In terms of productive activities, 76 percent of the households have one or more members active in agriculture and fishing. Agriculture and fishing utterly dominates activities in rural areas with 98% of households having one member at least part-time active in this sector. Agriculture remains important in urban areas with 50 percent of households having an individual identifying it as a primary activity. In urban areas, 31 percent of households reported having a member working in trading/commerce. For service activities the number is 27 percent.

[11] General problems with the reporting of education expenditures (as discussed above) may also explain the low share.

3. Results

For both the outlay equivalence method and the Engel curve approach, the analysis is performed at the rural and urban levels with households further divided by socio-economic status (poor and non-poor households). The results are first described for poor and then for non-poor households.

3.1 Analysis for poor households

3.1.1 Outlay equivalence method

Table 2 presents p-values from the tests for identification of adult goods based on equation (1) for the sub-set of poor households. The last column shows the p-values for the joint tests (in italics). The results for this sub-set of population indicate that all six candidate adult goods qualify, although tobacco is a borderline case.

Table 3 presents π-ratios and standard errors for the analysis conducted at rural and urban levels respectively, for poor households. As stated above, negative π ratios indicate compression of expenditure on the associated adult good due to the addition of a child in a given age group. There are seven goods (the six adult goods plus the results for all six goods combined) and three age classes resulting in 21 comparisons at each of the two levels of analysis (rural and urban) or 42 comparisons overall. However, the crucial comparison is with respect to the aggregate of all six adult goods. For this case, the relationship is as hypothesized (greater compression of expenditure on adult goods with respect to biological children) in five of six instances. For urban 0-5 year olds the result indicate possible reverse discrimination. The results from the Wald tests of equal π-ratios across adult goods (equation 6) are shown in Table 4. For all children demographic groups in rural and urban areas the hypothesis of equal π-ratios cannot be rejected, thus confirming the results from the F-tests presented in Table 2.

Table 2. Poor households: Test for 'true' adult goods (p-values).

Adult goods candidates	Biological 0-5	Non-biological 0-5	Biological 6-10	Non-biological 6-10	Biological 11-15	Non-biological 11-15	Join test of excluding all children groups
				P- values			
				Urban			
Alcohol	0.223	0.163	0.865	0.418	0.411	0.767	*0.831*
Tobacco	0.273	0.221	0.704	0.477	0.243	0.072	*0.514*
Adult cloth	0.452	0.370	0.510	0.532	0.280	0.670	*0.827*
Transportation	0.875	0.163	0.873	0.05	0.954	0.548	*0.169*
Meal and soft drink away home	0.192	0.494	0.468	0.172	0.15	0.036	*0.207*
Personal care	0.527	0.458	0.715	0.539	0.876	0.963	*0.925*
				Rural			
Alcohol	0.058	0.501	0.079	0.085	0.315	0.016	*0.379*
Tobacco	0.185	0.839	0.008	0.241	0.024	0.099	*0.060*
Adult cloth	0.469	0.255	0.423	0.518	0.997	0.217	*0.813*
Transportation	0.261	0.202	0.108	0.412	0.305	0.069	*0.624*
Meal and soft drink away home	0.329	0.548	0.519	0.443	0.038	0.96	*0.420*
Personal care	0.21	0.177	0.86	0.22	0.151	0.585	*0.113*

Notes: Columns labeled with demographic groups present p-values from t-test of individual coefficients in the adult goods regression (Equation 1). The last column shows the p-values for the F-tests that all children's demographic groups are jointly zero.

Table 3. Poor households: Outlay equivalence ratios.

Adult goods	Biological 0-5	Non-biological 0-5	Biological 6-10	Non-biological 6-10	Biological 11-15	Non-biological 11-15
			Urban			
	0.093	0.027	0.206	-0.216	1.199	0.094
Alcohol	(0.647)	(0.604)	(0.778)	(0.592)	(0.900)	(0.732)
	0.715	0.325	-0.216	-0.638	-0.723*	*0.287*
Tobacco	(0.767)	(0.766)	(0.503)	(0.779)	(0.372)	(0.798)
	0.028	-0.496*	-0.680**	*0.144*	-0.886***	*0.279*
Adult clothing	(0.309)	(0.299)	(0.303)	(0.404)	(0.269)	(0.637)
	0.027	-0.195	-0.512*	*0.074*	-0.401	*-0.136*
Transportation	(0.256)	(0.340)	(0.301)	(0.350)	(0.284)	(0.461)
	-0.153	-0.620	0.028	-0.472	-0.382	-1.408**
Meal and drink away home	(0.449)	(0.668)	(0.654)	(0.672)	(0.645)	(0.669)
	-0.234	-0.488	-0.213	*-0.103*	-0.461*	*0.079*
Personal Care	(0.258)	(0.353)	(0.238)	(0.349)	(0.307)	(0.549)
	0.180	-0.179	-0.418**	*-0.168*	-0.580***	*0.096*
All 6 goods	(0.200)	(0.245)	(0.187)	(0.259)	(0.149)	(0.310)
			Rural			
	-0.745*	*-1.014***	-0.591	*0.560*	-0.715*	*-1.129***
Alcohol	(0.395)	(0.453)	(0.389)	(0.827)	(0.423)	(0.504)
	-0.157	*0.919**	0.070	0.369	-0.092	*0.485*
Tobacco	(0.346)	(0.550)	(0.386)	(0.580)	(0.434)	(0.632)
	0.059	0.211	-0.032	*0.053*	0.032	0.260
Adult clothing	(0.190)	(0.328)	(0.199)	(0.307)	(0.217)	(0.362)
	-0.388	*-0.011*	-0.425	*0.662*	-0.111	-0.786**
Transportation	(0.316)	(0.486)	(0.292)	(0.527)	(0.388)	(0.390)
	1.991	-0.264	0.057	0.416	-1.887**	*-0.028*
Meal and drink away home	(1.373)	(0.633)	(0.611)	(1.177)	(0.834)	(0.785)
	-0.555**	*0.043*	-0.387	*0.793*	-0.674**	*-0.339*
Personal Care	(0.247)	(0.718)	(0.259)	(1.416)	(0.332)	(0.559)
	-0.168	*0.223*	-0.144	*0.300*	-0.152	*-0.025*
All 6 goods	(0.133)	(0.204)	(0.128)	(0.242)	(0.165)	(0.248)

Notes: Outlay equivalence ratios calculated from equation 4. Standard errors (in parenthesis) are obtained by bootstrap with 1,000 replications, cf. the main text. Entries in italics indicate a larger non-biological than biological outlay equivalence (reverse discrimination).
*, **, *** denotes significance levels of 10, 5 and 1 percent, respectively.

Table 5 presents the results of t-tests for equality of π ratios between biological and non-biological children at the rural and urban levels for each adult good. Again, the crucial tests are the ones for all six goods combined. For this aggregate, the greater compression of expenditure on adult goods with respect to biological children was found to be statistically significant for three of the six possible cases. Muddying the waters somewhat, the one case with an unexpected sign (more compression of household expenditures for non-biological children than biological in the case of children from 0-5 years old in urban areas) is also statistically significant at the 10% level.

Table 4. Poor households: Wald tests for equality of π-ratios across adult goods.

Demographic group	Rural		Urban	
	Test statistic	p-value	Test statistic	p-value
(1) Biological (0-5)	8.262	0.14	1.861	0.87
(2) Non-biological (0-5)	8.377	0.14	2.066	0.84
(3) Biological (6-10)	3.359	0.64	3.002	0.70
(4) Non-biological (6-10)	1.410	0.92	1.403	0.92
(5) Biological (11-15)	8.957	0.11	6.225	0.28
(6) Non-biological (11-15)	8.313	0.14	4.491	0.48
General category (16-20)	9.663	0.09	6.717	0.24
General category (21-24)	7.048	0.22	11.374	0.04
General category (25-59)	8.491	0.13	4.958	0.42

Notes: Wald statistic calculated from equation 6, distributed χ^2 with 5 degrees of freedom. Reported p-values equal the probability of observing a Wald statistic larger than the reported test statistic under the null of equality of π-ratios.

Table 5. Poor households: T-tests for equality of π-ratios for each age group.

Adult goods	Urban			Rural		
	Children 0-5	Children 6-10	Children 11-15	Children 0-5	Children 6-10	Children 11-15
	P- Values					
Alcohol	0.92	0.59	0.32	0.56	0.26	0.50
Tobacco	0.41	0.68	0.26	0.04**	0.60	0.38
Adult clothing	0.22	0.08*	0.10*	0.67	0.83	0.63
Transportation	0.63	0.12	0.56	0.45	0.08*	0.23
Meal and soft drink away home	0.44	0.52	0.18	0.22	0.62	0.08*
Personal Care	0.54	0.73	0.28	0.29	0.37	0.57
All 6 goods	0.06*	0.40	0.05**	0.09*	0.06*	0.71

Notes: P-values from t-test of equality of demographic coefficients from equation (3) for each good.
*, **, *** denotes significance levels of 10, 5 and 1 percent, respectively.

3.12 Engel curve approach

Table 6 shows the results from estimating Engel curves for children's clothing expenditure for poor household in the rural and urban sub-samples. Starting with the rural sample, the first column gives estimated coefficients and their standard errors for a standard Engel curve specification estimated by OLS. In the two next columns the estimated coefficients from the hurdle model (equation 7 and 8) are presented. The last column in the rural section shows marginal effects on the expenditure share coming from the hurdle model. The last four columns display analogous results from the urban sample. Our interest centers on differences in coefficients across demographic groups belonging to the same age category. The last three rows summarize the differences for each column together with significance levels. A plus sign indicates a difference which is consistent with discrimination, i.e. more is spent on children's clothing for biological children than for non-biological children controlling for other factors. For both the rural and urban samples the pure OLS estimates give some indication of discrimination, albeit in different age groups. In rural areas discrimination is significant at 10 percent for the 6 to 10 years old and at 5 percent in the age category 11 to 15 years old. For the urban areas there is – based on expenditures on children's clothing – weak evidence (significant at 10 percent) of discrimination for the age group 0-5 years. For the hurdle model significant differences (5 percent) only show up in the rural sample for the oldest age group (11 to 15 years). Note, that even though some of the differences in coefficients between demographic groups are consistent with reverse discrimination (indicated by a minus sign) – none of these are significant. To the extent that discrimination is present in the data, the hurdle model does not seem to be better than a pure OLS estimation to detect them. The fact that both models find significant discrimination for the 11 to 15 year old children in rural areas is taken as a sign of robustness of the result.

Table 7 is equivalent to Table 6 but for educational expenditure. Despite a majority of individual level regressions showing significant coefficients on the demographic groups, we find no significant differences between the impact of biological and non-biological demographic groups for school age children.

Table 6. Poor households: Engel curves for clothing expenditure.

Variable	Rural				Urban			
	OLS (Exp. share) (x 100)	Probit[a)] (Exp. share > 0)	OLS (Log exp. share)	Probit, OLS combined [b)] (x 100)	OLS (Share) (x 100)	Probit [a)] (Share > 0)	OLS (Log exp. share)	Probit, OLS combined [b)] (x 100)
Log pr. capita expenditure	0.17	0.43***	-0.48*	0.14	-0.03	0.23	-0.50	0.00
	(0.10)	(0.16)	(0.25)	(0.12)	(0.16)	(0.17)	(0.40)	(0.11)
Log household size	-0.36	-0.03	-0.66	-0.24	-0.41	-0.16	-0.77	-0.24
	(0.22)	(0.33)	(0.48)	(0.27)	(0.27)	(0.36)	(1.05)	(0.28)
Demographic group:								
[1] Biological 0-5	0.17	-0.95	1.02	-0.27	3.31*	1.30	6.54*	1.95
	(0.54)	(0.88)	(2.05)	(0.88)	(1.83)	(1.27)	(3.36)	(1.21)
[2] Non-biological 0-5	0.83	-0.29	0.53	0.03	1.19	-0.08	5.40	1.16
	(0.79)	(1.06)	(2.36)	(0.97)	(1.24)	(1.40)	(3.83)	(1.22)
[3] Biological 6-10	0.71	-1.20	3.08	0.16	2.94**	1.80	4.50	1.70
	(0.69)	(0.93)	(2.05)	(0.94)	(1.31)	(1.26)	(3.55)	(1.18)
[4] Non-biological 6-10	-0.17	-2.25**	2.91	-0.59	3.43*	1.92	5.56	1.89
	(0.64)	(1.10)	(2.23)	(1.06)	(1.99)	(1.50)	(3.61)	(1.28)
[5] Biological 11-15	0.01	-1.41	1.57	-0.39	0.97	-0.12	4.50	0.94
	(0.59)	(0.98)	(2.40)	(1.02)	(1.17)	(1.39)	(3.26)	(1.14)
[6] Non-biological 11-15	-0.72	-2.90***	-0.47	-1.95*	2.04	0.72	6.70	1.68
	(0.57)	(1.10)	(2.48)	(1.13)	(1.37)	(1.47)	(3.40)	(1.28)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. obs. (No. clusters)	1723 (476)	1723 (476)	124 (94)	1723 (476)	1449 (309)	1449 (309)	117 (78)	1449 (309)
R^2 (Probit,, McFadden pseudo R^2)	0.03	0.10	0.54	..	0.04	0.14	0.60	..
Test for discrimination of orphans (equality of coefficients): (Significance levels marked with asterisks)								
(0 - 5 years) [1] – [2]	–	–	+	–	+*	+	+	+
(6 - 10 years) [3] – [4]	+*	+	+	+	–	–	–	–
(11 - 15 years) [5] – [6]	+**	+*	+*	+**	–	–	–	–

Notes: Standard errors in parenthesis. *, **, *** denotes significance levels of 10, 5 and 1 percent, respectively. Additional controls include the variables described in table 1: Provincial dummies, dummy for female headed household, dummy for education level of EP2 of household head, Age and its square of the household head, dummies for occupation (none, agriculture, commerce and service).

a) Coefficients (not marginal effects).

b) Marginal effects at the mean value of explanatory variables. Standard errors obtained by a bootstrap with 1,000 replications as described in the main text.

Table 7. Poor households: Engel curves for educational expenditure.

Variable	Rural				Urban			
	OLS (Exp. share) (x100)	Probit[a)] (Exp. share > 0)	OLS (Log exp. share)	Probit, OLS combined [b)] (x100)	OLS (Share) (x100)	Probit [a)] (Share > 0)	OLS (Log exp. share)	Probit, OLS combined [b)] (x 100)
Log pr. capita expenditure	-0.40***	0.28**	-0.95***	-0.22***	-0.20	1.02***	-0.44***	-0.36***
	(0.13	(0.12)	(0.09)	(0.04)	(0.14)	(0.20)	(0.10)	(0.12)
Log household size	0.44**	0.26	0.70***	0.21***	-0.19	-0.70*	0.26	0.19
	(0.20	(0.26)	(0.21)	(0.07)	(0.30)	(0.41)	(0.22)	(0.27)
Demographic group:								
[1] Biological 0-5	-0.16	0.38	-1.08**	-0.23	0.74	-0.35	0.13	-0.08
	(0.37	(0.75)	(0.54)	(0.19)	(0.99)	(1.03)	(0.86)	(0.92)
[2] Non-biological 0-5	-0.31	0.32	-0.72	-0.16	-0.95	-1.94	-1.28	-1.66*
	(0.41	(0.75)	(0.69)	(0.23)	(1.06)	(1.25)	(0.86)	(0.91)
[3] Biological 6-10	0.95*	2.13***	0.43	0.38**	2.92***	1.48	1.26	1.66*
	(0.53	(0.71)	(0.53)	(0.19)	(1.02)	(1.25)	(0.81)	(0.87)
[4] Non-biological 6-10	0.67	2.17**	0.22	0.32	3.06**	0.84	1.63*	1.88**
	(0.54	(0.94)	(0.61)	(0.23)	(1.23)	(1.52)	(0.91)	(0.96)
[5] Biological 11-15	1.18**	3.62***	1.43**	0.81***	3.77***	2.31**	2.18***	2.64***
	(0.46	(0.79)	(0.55)	(0.20)	(0.96)	(1.17)	(0.71)	(0.76)
[6] Non-biological 11-15	1.25**	2.83***	1.09*	0.64***	3.34***	1.39	2.06**	2.47**
	(0.52	(0.89)	(0.62)	(0.23)	(1.22)	(1.38)	(0.94)	(1.04)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. obs. (No. clusters)	1723 (476)	1723 (476)	1227 (422)	1723 (476)	1449 (309)	1449 (309)	1274 (306)	1449 (309)
R^2 (Probit,, McFadden pseudo R^2)	0.19	0.16	0.50	..	0.22	0.27	0.35	..
Test for discrimination of orphans (equality of coefficients): (None of the differences are significant)								
(6 - 10 years) [3] – [4]	+	–	+	+	–	+	–	–
(11 - 15 years) [5] – [6]	–	+	+	+	+	+	+	+

Notes: Standard errors in parenthesis. *, **, *** denotes significance levels of 10, 5 and 1 percent, respectively. Additional controls include the variables described in table 1: Provincial dummies, dummy for female headed household, dummy for education level of EP2 of household head, Age and its square of the household head, dummies for occupation (none, agriculture, commerce and service).

[a)] Coefficients (not marginal effects).

[b)] Marginal effects at the mean value of explanatory variables. Standard errors obtained by a bootstrap with 1,000 replications as described in the main text.

Although the outlay equivalence analysis on poor households contains the one counterintuitive result that biological children are discriminated relative to non-biological children for the age group of 0 to 5 year olds in urban areas, overall, the results indicates discrimination in intra-household resource allocation against children that are not the biological descendant of the household head for poor households. The findings in the outlay equivalence analysis are somewhat corroborated by Engel curve analysis of expenditures on children's clothing and education. In particular, discrimination is significant for children aged 6 to 10 years in rural areas for both the outlay equivalence and Engel curve analysis on clothing expenditure.

3.2 Analysis for non-poor households

3.2.1 Outlay equivalence method

Table 8 presents results of the tests for identification of adult goods based on equation (1) for the subset of non-poor households. Three out of the six candidate adult goods fail to pass this test, namely adult clothing, tobacco and transportation. Since the method is invalid with non-adults goods, the analysis preceded using only the three goods (alcohol, Meal/drink away from home and personal care) that qualified as adult goods (Table 8). π-ratios and their associated standard errors are presented in Table 9 and the Wald tests for equality of π-ratios across the three adult goods are given in Table 10. Results show that the hypothesis of the three good being adult goods cannot be rejected. Looking at all three adult goods combined (Table 9), four of the six comparisons of coefficients between biological and non-biological demographic groups show the outlay equivalence ratio to be smaller for biological than non-biological children, thus, indicating discrimination. However, Table 11 – which shows p-values from the test of differences between π-ratios – fails to find any significant difference between biological and non-biological children.

3.2.2 Engel curve approach

The results from the Engel curve analysis of respectively children's clothing and educational expenditure for non-poor households are given in Table 12 and 13. Starting with children's clothing, the majority of coefficient comparisons are consistent with discrimination (last three rows).

Table 8. Non-poor households: Test for 'true' adult goods (p-values).

Adult goods	Biological 0-5	Non-biological 0-5	Biological 6-10	Non-biological 6-10	Biological 11-15	Non-biological 11-15	Join test of excluding all children groups
				P- values			
				Urban			
Alcohol	0.371	0.132	0.463	0.730	0.539	0.342	*0.584*
Meal and soft drink away home	0.185	0.187	0.492	0.089	0.535	0.311	*0.275*
Personal care	0.807	0.648	0.128	0.309	0.736	0.839	*0.351*
				Rural			
Alcohol	0.879	0.827	0.147	0.254	0.803	0.064	*0.262*
Meal and soft drink away home	0.077	0.662	0.949	0.377	0.240	0.981	*0.525*
Personal care	0.557	0.878	0.126	0.156	0.917	0.051	*0.302*

Notes: Columns labeled with demographic groups present p-values from t-test of individual coefficients in the adult goods regression (Equation 1). The last column shows the p-values for the F-tests that all children's demographic groups are jointly zero.

Table 9. Non-poor households: Outlay equivalence ratios.

Adult goods	Biological 0-5	Non-biological 0-5	Biological 6-10	Non-biological 6-10	Biological 11-15	Non-biological 11-15
			Urban			
	-0.472	*-0.096*	-0.715	*-0.566*	0.730	0.643
Alcohol	(0.561)	(0.726)	(0.556)	(0.822)	(0.684)	(0.740)
	-2.050**	*-1.345*	-1.159	-2.481***	-0.323	*0.821*
Meal and soft drink away home	(0.814)	(1.663)	(0.825)	(1.368)	(0.853)	(2.037)
	-0.576	-1.263	0.298	0.799	-0.564	*0.702*
Personal care	(0.431)	(1.276)	(0.366)	(1.325)	(0.472)	(0.818)
	-0.865*	*-0.747*	-0.543	-0.653	0.116	0.719
All 3 goods	(0.467)	(0.726)	(0.364)	(0.629)	(0.429)	(0.613)
			Rural			
	-0.413	*-0.360*	-0.386	-1.107	-0.142	*-0.107*
Alcohol	(0.529)	(0.702)	(0.483)	(0.686)	(0.514)	(0.795)
	-1.609*	*5.113*	-0.929	-1.909**	-1.748**	*-1.654**
Meal and soft drink away home	(0.891)	(3.815)	(0.741)	(0.901)	(0.773)	(0.890)
	-0.493	*0.337*	0.154	0.354	-0.207	*1.166*
Personal care	(0.344)	(0.865)	(0.475)	(0.754)	(0.368)	(1.545)
	-0.497	*0.125*	-0.312	-0.860*	-0.254	*0.065*
All 3 goods	(0.402)	(0.510)	(0.335)	(0.519)	(0.383)	(0.668)

Notes: Outlay equivalence ratios calculated from equation 4. Standard errors (in parenthesis) are obtained by bootstrap with 1,000 replications, cf. the main text. Entries in italics indicate a larger non-biological than biological outlay equivalence (reverse discrimination).
*, **, *** denotes significance levels of 10, 5 and 1 percent, respectively.

Table 10. Non-poor households: Wald tests for equality of π-ratios across adult goods.

Demographic group	Rural: Test statistic	Rural: p-value	Urban: Test statistic	Urban: p-value
(1) Biological (0-5)	1.499	0.47	2.949	0.23
(2) Non-biological (0-5)	2.214	0.33	0.894	0.64
(3) Biological (6-10)	1.628	0.44	4.169	0.12
(4) Non-biological (6-10)	4.207	0.12	3.150	0.21
(5) Biological (11-15)	3.503	0.17	2.490	0.29
(6) Non-biological (11-15)	3.170	0.20	0.008	0.99
General category (16-20)	0.301	0.86	5.160	0.08
General category (21-24)	2.476	0.29	3.155	0.21
General category (25-59)	1.813	0.40	3.209	0.20

Notes: Wald statistic calculated from equation 6, distributed χ^2 with 2 degrees of freedom. Reported p-values equal the probability of observing a Wald statistic larger than the reported test statistic under the null of equality of π-ratios.

However, only three of the differences in coefficients are significant and two of them point towards discrimination of biological children. Only the pure OLS model finds discrimination of non-biological children. This result is found for the youngest age group (0-5 years). Looking at Table 13 no significant results indicating discrimination in education expenditure are found and in general no consistent picture emerges from the (insignificant) differences in coefficients.

Based on the methods employed we were not able to find any evidence of discrimination in the non-poor sample of households.

Table 11. Non-poor households: T-tests for equality of π-ratios for each age group.

	Urban			**Rural**		
Adult goods	Children 0-5	Children 6-10	Children 11-15	Children 0-5	Children 6-10	Children 11-15
			P- Value			
Alcohol	0.64	0.81	0.73	0.96	0.39	0.93
Meal drink away from home	0.65	0.34	0.67	0.16	0.19	0.84
Personal Care	0.45	0.61	0.27	0.35	0.77	0.28
All 3 goods	0.99	0.83	0.52	0.22	0.39	0.64

Notes: P-values from the test of equality of π-ratios in each age group. P-values obtained from t-test of equality of means in the bootstrapped sample.
*, **, *** denotes significance levels of 10, 5 and 1 percent, respectively.

Table 12. Non-poor households: Engel curves for clothing expenditure.

Variable	Rural				Urban			
	OLS (Exp. share) (x 100)	Probit[a)] (Exp. share > 0)	OLS (Log exp. share)	Probit, OLS combined [b)] (*100)	OLS (Share) (x 100)	Probit [a)] (Share > 0)	OLS (Log exp. share)	Probit, OLS combined [b)] (x 100)
Log pce	-0.09	0.02	-0.34*	-0.12	0.09	0.22**	-0.17	0.05
	(0.17)	(0.17)	(0.19)	(0.15)	(0.13)	(0.11)	(0.19)	(0.10)
Log household size	-0.35	0.13	-0.79	-0.21	-0.86*	-0.43	-1.28**	-0.59*
	(0.37)	(0.39)	(0.48)	(0.35)	(0.52)	(0.28)	(0.54)	(0.30)
Demographic group:								
[1] Biological 0-5	1.59	2.83***	1.35	2.52**	5.57**	5.61***	7.25***	5.59***
	(1.03)	(1.02)	(2.13)	(1.15)	(2.22)	(1.56)	(2.30)	(1.65)
[2] Non-biological 0-5	1.23	2.30*	0.82	1.97	2.64*	3.99**	4.88*	4.10**
	(1.83)	(1.35)	(2.33)	(1.49)	(1.56)	(1.67)	(2.90)	(1.64)
[3] Biological 6-10	2.35***	2.23**	2.56	2.55**	5.22***	5.55***	7.51***	5.91***
	(0.72)	(0.92)	(2.09)	(1.11)	(1.89)	(1.65)	(2.27)	(1.61)
[4] Non-biological 6-10	1.69	1.79*	2.32	2.12*	4.91**	5.83***	8.49***	6.31***
	(1.05)	(1.07)	(2.45)	(1.23)	(1.97)	(1.77)	(2.74)	(1.91)
[5] Biological 11-15	-0.35	-0.07	1.23	0.47	2.76*	4.59***	5.17**	4.45***
	(0.83)	(1.00)	(2.25)	(1.19)	(1.41)	(1.42)	(2.12)	(1.52)
[6] Non-biological 11-15	0.71	1.38	1.85	1.58	2.76*	3.90**	4.81*	4.05***
	(0.99)	(1.15)	(2.15)	(1.21)	(1.51)	(1.58)	(2.55)	(1.51)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. obs. (No. clusters)	1234 (394)	1234 (394)	170 (124)	1234 (394)	1288 (311)	1288 (311)	183 (120)	1288 (311)
R^2 (Probit,, McFadden pseudo R^2)	0.08	0.15	0.39	..	0.05	0.12	0.46	..
Test for discrimination of orphans (equality of coefficients): (Significance levels marked with asterisks)								
(0 - 5 years) [1] – [2]	+	+	+	+	+**	+	+	+
(6 - 10 years) [3] – [4]	+	+	+	+	+	–	–	–
(11 - 15 years) [5] – [6]	–*	–**	–	–	–	+	+	+

Notes: Standard errors in parenthesis. *, **, *** denotes significance levels of 10, 5 and 1 percent, respectively. Additional controls include the variables described in table 1: Provincial dummies, dummy for female headed household, dummy for education level of EP2 of household head, Age and its square of the household head, dummies for occupation (none, agriculture, commerce and service).

a) Coefficients (not marginal effects).

b) Marginal effects at the mean value of explanatory variables. Standard errors obtained by a bootstrap with 1,000 replications as described in the main text.

Table 13. Non-poor households: Engel curves for educational expenditure.

Variable	Rural				Urban			
	OLS (Exp. share) (x 100)	Probit [a] (Exp. share > 0)	OLS (Log exp. share)	Probit, OLS combined [b] (x 100)	OLS (Share) (x 100)	Probit [a] (Share > 0)	OLS (Log exp. share)	Probit, OLS combined [b] (x 100)
Log pce	-0.10*	0.14	-0.64***	-0.09***	-0.24*	0.28*	-0.56***	-0.49***
	(0.05)	(0.14)	(0.14)	(0.03)	(0.13)	(0.17)	(0.09)	(0.09)
Log household size	0.17	0.98***	0.34	0.12**	0.31	0.91**	0.65***	0.59***
	(0.11)	(0.31)	(0.30)	(0.06)	(0.37)	(0.39)	(0.24)	(0.21)
Demographic group:								
[1] Biological 0-5	-0.66	-0.97	-1.50**	-0.30**	1.60*	1.12	-0.12	-0.16
	(0.40)	(0.79)	(0.70)	(0.12)	(0.94)	(1.43)	(0.90)	(0.81)
[2] Non-biological 0-5	-0.74**	-2.17**	-2.35***	-0.51***	0.48	2.02	0.01	-0.05
	(0.38)	(1.05)	(0.81)	(0.16)	(1.04)	(1.72)	(0.86)	(0.83)
[3] Biological 6-10	-0.47	-0.01	0.47	0.08	3.42***	6.29***	1.86*	1.85**
	(0.40)	(0.77)	(0.63)	(0.11)	(1.08)	(2.37)	(1.01)	(0.90)
[4] Non-biological 6-10	-0.60	0.74	-0.06	0.05	2.91***	6.47***	2.19**	2.12**
	(0.49)	(0.87)	(0.72)	(0.13)	(0.93)	(2.49)	(1.03)	(0.95)
[5] Biological 11-15	-0.24	1.65**	0.43	0.19*	3.28***	4.89***	2.65***	2.42***
	(0.33)	(0.74)	(0.66)	(0.11)	(1.07)	(1.57)	(0.92)	(0.87)
[6] Non-biological 11-15	-0.47	1.56*	-0.23	0.08	3.75***	5.26***	2.71***	2.45***
	(0.51)	(0.92)	(0.83)	(0.15)	(1.22)	(2.00)	(0.95)	(0.90)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. obs. (No. clusters)	1234 (394)	1234 (394)	853 (336)	1234 (394)	1288 (311)	1288 (311)	1184 (305)	1288 (311)
R^2 (Probit,, McFadden pseudo R^2)	0.07	0.16	0.34	..	0.16	0.36	0.31	..
Test for discrimination of orphans (equality of coefficients): (Significance levels marked with asterisks)								
(6 - 10 years) [3] – [4]	+	–	+	+	+	–	–	–
(11 - 15 years) [5] – [6]	+	+	+*	+	–	–	–	–

Notes: Standard errors in parenthesis. *, **, *** denotes significance levels of 10, 5 and 1 percent, respectively. Additional controls include the variables described in table 1: Provincial dummies, dummy for female headed household, dummy for education level of EP2 of household head, Age and its square of the household head, dummies for occupation (none, agriculture, commerce and service).

[a] Coefficients (not marginal effects).

[b] Marginal effects at the mean value of explanatory variables. Standard errors obtained by a bootstrap with 1,000 replications as described in the main text.

4 Conclusions

The weight of evidence points to discrimination in the intra-household allocation of resources against children that are not direct biological descendants of the household head in poor households. The outlay equivalence analysis found significant discrimination for younger children (aged 0-10) in rural households and older children (aged 11-15) in urban households. For the rural households this was confirmed for the 6-10 year olds by the Engel curve approach (children's clothing expenditure), which also found significant discrimination for the oldest age group in rural areas.

There is no evidence that non-poor households discriminate against children that are not the biological descendant of the household head. There are two likely reasons underpinning the dichotomy of results between poor and non-poor households. First, resources are more severely constrained in poor versus non-poor households forcing more difficult choices in resource allocation. Non-biological children may experience discrimination under these harsher economic conditions. Second, our inability to identify the reason for the presence of a non-biological child within a family may also play a role. The available evidence indicates that wealthier households are more likely to host children in order for them to attend school (Nhate, 2004). Hence, the bias from mixing together children that are likely to be discriminated against (AIDS orphans for example) with children that are not (those living with friends or relatives in order to attend school) under a single rubric "non-biological children" may be substantially more profound in the non-poor sub-set of the population. As indicated earlier, the results obtained are likely a lower bound on the discrimination against the target group of children.

Unfortunately, AIDS is likely to aggravate the problem over the next five to ten years by substantially increasing the number of children requiring care from neighbors, friends, and/or relatives due to the death of one or more of their parents. As the overall burden on communities grows, few would hypothesize that the tendency for non-biological children to reside in better off households would become more pronounced or the degree of discrimination against non-biological children would decline. Rather, the inverse seems more likely.

If one wishes to target some assistance at particularly disadvantaged groups, then children living in poor households that are not the biological descendant of the household head, especially those that do not attend school or attend school only sporadically, would appear to be a logical choice. The results also indicate that the policy of placing orphans in families of neighbors, friends or relatives functions less well, in terms of the interests of the orphans, than would occur in a world free of discrimination. Further, the policy may perform even more poorly as the burden grows. Nevertheless, the result does not necessarily imply that the policy should be abandoned. This decision can only be reached through comparison with potential substitute policies. While the analysis of potential substitute policies merits further attention, the available evidence indicates that attractive substitute policies are few to non-existent. Despite discrimination, the current policy may be the best available alternative.

References

Beegle, K., De Weerdt, J. & Dercon, S., ’Orphanhood and the Long-Run Impact on Children’, *American Journal of Agricultural Economics* 88, No. 5, pp. 1266-72.

Behrman, J., 1990. *Intrahousehold Allocation of Nutrients and Gender Effects: A Survey of Structural and Reduced-Form Estimates.* Oxford: Oxford University Press.

Cameron, A.C. & Trivedi, P.K., (2005). *Microeconometrics – Methods and Applications.* Cambridge University Press, New York, USA.

Case, A. & Ardington, C., 2006, ‘The Impact of Parental Death on School Outcomes: Longi-tudinal Evidence from South Africa’, *Demography* 43, No. 3, pp 401-420.

Case, A. & Deaton, A, 2002, ‘Consumption, Health, Gender and Poverty’, WP

Case, A., Paxson, P., & Ableidinger, J., 2004, ‘Orphans in Africa: Parental Death, Poverty, and School Enrollment’, *Demography* 41, No. 3, pp. 483-508.

Deaton, A., Ruiz-Castilo, J. & Thomas, D., 1989a, ‘The influence of Household Composition on Household Expenditure Patterns: Theory and Spanish Evidence.’ *Journal of Political Economy,* vol. 97, No.1.

Deaton, A., 1989b. ‘Looking for Boy-Girl Discrimination in Household Expenditure Data’. *World Bank Economic Review,* vol.3, No.1.

Evans, D., Miguel, A.M., 2005, ‘Orphans and Schooling in Africa: A Longitudinal Analysis’, WP, Center for International and Development Economics Research, University of California, Berkeley

Gibson, J. & Rozelle, S., 2004, ‘Is it Better to be a Boy? A Disaggregated Outlay Equivalent Analysis of Gender Bias in Papua New Guinea.’ *The Journal of Development Studies*, Vol. 40, No. 4, pp. 115-136.

GM (Governo de Moçambique); 2004a. Plano Nacional para a Crianca. Draft. Versao de 23 de Junho de 2004.

Haddad, L. & Reardon T, 1993, ‘Gender Bias in the Allocation of Resources Within Household in Burkina Faso: A Disaggregated Outlay Equivalent Analysis’. *The Journal of Development Studies*, Vol.29, No. 2, pp. 260-276.

Haddad, L., Pena, C., Nishida, C., Quisumbing A. & Slack, A., 1996, ‘Food Security and Nutrition. Implications of Intrahousehold Bias: A Review of Literature.’ FNCD Discussion Paper No. 19. IFPRI, Washington, DC.

Hamilton, W.D., 1964, ‘The genetical evolution of social biology’, *Journal of Theoretical Biology* 7, 1-16.

INE (Instituto Nacional de Estatistica), 2004. 'Inquerito Nacional ao Agregados Familiares Sobre Orcamento Familiar 2002/03.' Maputo.

INE (National Institute of Statistics), Ministry of Health, Ministry of Planning and Finance, National Council of Fighting HIV/AIDS, the Center for Population Studies and Faculty of Medicine of Eduardo Mondlane University; 2004 . 'Impacto Demográfica do HIV/SIDS em Moçambique. Actualizacao*'*. Maputo.

Kingdon, G.G., 2005, 'Where has the Bias Gone? Detecting Gender Bias in the Intrahousehold Allocation of Educational Expenditure', *Economic Development and Cultural Change*, Vol. 53, No. 2, pp. 409-452.

Ministry of Planning and Finance, International Food Policy Reaserch Institute and Eduardo Mondlane University; 1998. 'Understanding Poverty and Well Being in Mozambique. The First National Assessment'. Maputo - Mozambique.

Miller, B., 1981, *The Endangered Sex: Neglect of Female Children in Rural North India*. Ithaca, N.Y.: Cornell University Press.

Nhate, V., 2004, 'Orphans in Mozambique: Vulnerability, Trends, Determinants, and Programme Responses.' Ministry of Planning and Development. Mozambique. Maputo.

Sharma, M.P., 2006, 'Orphanhood and Schooling Outcomes in Malawi', *American Journal of Agricultural Economics* 88, No. 5, pp. 1273-1278.

UNUSIDA, 1999. Crianças Orfãos Devido Sida. Respostas da Vanguarda da Africa Oriental e Austral.

Wooldridge, J.M., 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT press, Massachusetts, US.

World Bank, 2004, 'Mozambique – School Fees and Primary School Enrollment and Retention', World Bank. Available from: http://siteresources.worldbank.org/INTPSIA/Resources/490023-1120841262639/ch4_mozambique.pdf.

Regional differences in food consumption in urban Mozambique: A censored demand system approach

Mikkel Barslund[a]
University of Copenhagen, Development Economics Research Group (DERG)

March 2006
Revised: January 2007

Abstract

A nationwide household survey for Mozambique (IAF 2002/03) is used to estimate a large censored food demand system with 12 food groups for the sample of urban households. Using the translog indirect utility approach, the censored nature of the data is addressed by estimating a system of Tobit equations with a recently suggested quasi maximum likelihood estimator. Augmenting the system with demographic and geographical variables in a theoretically consistent way, I find that differences in elasticities between regions are significant. The results show that regional variation has to be taken into account when evaluating policies or employing CGE models. Further, the approach employed here can be applied to a number of developing countries with varying geographic conditions.

JEL classification: D12, O12, O18

Keywords – Censored demand system, Elasticities, Mozambique, Food demand, Regional differences.

[a] Address correspondence to: mikkel.barslund@econ.ku.dk. Part of this work was done while I was working in the Ministry of Planning and Development, Maputo, for the International Food Policy Research Institute (IFPRI). I would like to thank participants at the seminar in the Ministry of Planning and Development, Maputo, the NNDE conference in Oslo, Carol Newman and Christoffer Sonne-Schmidt for helpful suggestions. Thomas Barnebeck Andersen, Finn Tarp and Katleen Van den Broeck have given a wealth of comments for which I am particularly grateful. The usual disclaimer applies.

1. Introduction

Over the last ten years Mozambique has enjoyed a good spell of macroeconomic growth and – measured by a consumption metric – the level of poverty has fallen steadily from 69 percent in 1996 to 54 percent in 2002 (MPD, 2004). Other poverty indicators show similar improvements. Despite these achievements which had broad geographic coverage Mozambique is still a poor country with limited integrated markets and a strong need for further poverty reduction policies (Fox et al., 2005). The task of policy formulation is complicated by the geography of Mozambique. The two main centres of economic activity (Maputo, the capital in the south, and Beira in the central part of the country) are separated by more than a 1,000 kilometres. Further to the north of Beira the Zambezi River cuts off the northern part of the country and only poor infrastructure link the two parts, limiting economic integration.

Detailed knowledge of households' preference structure is a valuable tool for improving policy advice and evaluating the effects of existing policies. Important policy areas where such knowledge can improve policy advice span a range from tax reform and transfers to public goods provision. In poor countries, such as Mozambique with a per capita income of 1,300 USD (PPP) in 2005, where the bulk of expenditures are directed towards food consumption, studying food demand is important. This is so both as a goal in itself to allow policy makers to study the effect of policies directly affecting food demand (i.e. price subsidies) and as an essential building block in economy wide models, such as general equilibrium models (e.g. Jensen & Tarp, 2004).

Not surprisingly, given the importance of price and income response parameters, identifying and estimating these have a long tradition in applied economics. Because price variation is inherently necessary for any successful attempt at identifying consumers' responses to price changes, estimation of price and income elasticities in developed countries has usually relied on aggregate time series data. Since prices normally vary little within developed countries at any given point in time, time series data is essential to identify price variation. However, in developing countries where markets are less integrated and transportation is often burdensome in terms of both monetary and time costs, prices can be expected to vary considerably between locations. If nationally representative (or regionally representative but dispersed) cross sectional household survey data is available the price differences between different locations can be exploited to obtain estimates of price response parameters. This was first exploited by Deaton (1987) to estimate own and cross price elasticities from a cross section of households from Cote d'Ivoire.[1] An additional benefit from using household survey data is the availability of demographic variables at the household level. This makes it possible to (partly) control for household heterogeneity in the parameters. In addition, household surveys usually have a sample size which permits estimation of a large number of parameters.

This paper attempts to shed light on households' response to relative price changes in food products in urban areas in Mozambique, a geographically weakly integrated country.[2] The contribution of the present paper is twofold. First, I use the most recent national representative household survey conducted in Mozambique in 2002/03 (IAF02) to estimate a large complete demand system for 12 food groups augmented with demographic and locational variables using the urban household sample. This is done using a translog demand system approach. Using survey data at the household level both poses challenges and yields benefits. A recurring problem when estimating demand systems using micro data is the likely wide presence of households reporting zero consumption of one or more of the commodities analyzed. As will be evident, the problem of censoring is severe in the sample of urban Mozambican households considered here. Ignoring it would bias our estimates. The approach taken in the present paper accounts for censoring by estimating a system of Tobit equations as proposed by Yen, Lin and Smallwood (2003).

Second, location dummy variables for households living in the southern, central and northern part of the country are utilized in the demand system (together with demographic

[1] See Kedir (2005) for a recent application to Ethiopia.
[2] While the majority of Mozambican household reside in rural areas the urban definition applied in this study is quite broad covering some 30 percent of the population of households.

variables) to focus on differences in elasticities between these regions. In particular, the paper investigates to what extent differences are due to preferences (i.e. differences in parameters) or whether they are attributable to regional price differences. Given the geography of Mozambique, location is likely to be an important determinant of demand patterns and elasticities. The results are interesting in themselves and should be a valuable input into ongoing efforts at the Mozambican Ministry of Planning and Rural Development to construct both national and regional applied general equilibrium models. Further, the methodology used here can be applied to other countries where regions differ in geography.

For the purposes of this paper I am fortunate to use a data set which is both nationally representative and have a spatial time dimension in the way the data was collected. Because it was collected throughout a full year there is ample price variation over and above what exists between villages as a result of lack of market integration. Thus, elasticities are expected to be estimated with a better precision than is usually obtainable from cross-sectional data.

The remainder of the paper is structured as follows: in section 2 the data together with some descriptive statistics are presented. This is followed by an outline of the methodology employed in section 3, while section 4 presents the results. Finally, section 5 concludes.

2. Data and descriptive statistics

The data source for this study is the 2002/03 nationally representative household survey of Mozambican households (IAF). It contains detailed information on food consumption for a random sample of 8,700 households in Mozambique, as well as information on general characteristics of the household, daily expenses and consumption from home production, possession of durable goods, gifts and transfers received. All aspects of survey implementation and a set of summary statistics are available from the National Institute of Statistics (INE 2004). The interviewers were in the enumeration area for a week, during which three household visits were programmed in order to administer questionnaires and assist households in keeping track of daily consumption. Thus, to the extent it is possible food consumption should be very well covered within the survey period.

The survey was designed with an explicit view to be representative in time as well as space. Data collection was done over the space of one year divided into quarters. For each sub-group of the population, the survey was designed to represent, one quarter of the households were interviewed in each period.

The geography of Mozambique and the fact that 'around the year' price information is available should allow sufficient price variation to identify price elasticities with good precision. This is an exercise which is often difficult when only cross section data is available. It is natural to divide the 11 provinces of Mozambique into three distinct regions; south, central and north. The south is made up of the provinces Maputo City, Maputo province, Gaza and Inhambane. The provinces of Sofala, Manica, Tete and Zambezia constitute the central part of Mozambique. Lastly, the north includes Nampula, Niassa and Cabo Delgado.[3]

The food demand system estimated includes all expenditures on food products – divided into 11 separate food groups and a residual category; vegetables, maize flour, fish, bread, rice, meat, oil & fats, fruits, sugar, beans, other staples and the residual group other foods. Other staples consist of cassava and potatoes and the residual group includes beverages, spices and meals eaten outside the house. Maize, bread and rice are the main staples of Mozambican households in urban areas, with some also consuming cassava and potatoes (other staples). As an artefact of the geography of Mozambique fish is also widely consumed. Meat is composed of beef, pork and chicken meat. In nutritional terms beans are an important protein substitute for meat and fish. Fruits are consumed throughout Mozambique. A large component of oil and fats is cooking oil, but a limited number of households also consume butter.

[3] In the following the regions are simply referred to as: south, central and north.

To avoid the problems inherent in evaluating the value of home produced goods the scope is limited to the urban part of the sample.[4] This sample consists of 4,005 urban households interviewed in 335 clusters. However, households were removed where the unit price and expenditure information looked dubious or were missing, very large households, and households which had zero purchases for more than eight goods. Specifically, deleted households were those with more than 10 members, households with logarithmic income and prices more than four standard deviations from the mean sample value. Unit prices were obtained by averaging over all consuming households in each enumeration area. If no households in the enumeration area consumed the good the average over households interviewed in the same quarter in the same region (north, central or south) was used. Unit prices for bundles of goods are obtained by weighting individual good prices with the expenditure share. The final sample includes 3,543 households.

The first part of Table 1 presents expenditure shares on the 12 food groups for the south, central and the north separately. Expenditure shares clearly differ between regions. Vegetables are much more widely consumed in the south, with the highest expenditure share there, compared with central and north. On the other hand, maize flour which makes up around 24 percent of the budget in the central region is less important in the north and only accounts for 3 percent of expenditure in the south. Fish and other staples – mostly cassava and potatoes – are the most important food product for households located in the north, whereas these food groups are less important elsewhere, although fish is widely consumed. In the north sorghum is a significant part of the other staple category. Overall, it is clear that there are large regional differences in food consumption patterns which need to be accounted for in the estimation. In addition, Table 1 indicates the need for estimating a large demand system with many goods when the focus is on regional differences. Aggregating some of the categories further risks blurring regional differences in food consumption. Thus, the approach of limiting the number of demographic variables in favour of more food groups seems warranted.

[4] Excluding rural household to some extent limits the usefulness of the elasticity estimates obtained here for nationwide policy analysis. However, including rural households would add further complications without adding much value to the analysis of regional differences.

Table 1. Food consumption for urban Mozambican households.

	South	Central	North	Full Sample	
Food group	Share in total food expenditure (%)			Households consuming (%)	Mean expenditure share (%)
Vegetables	19.4	10.7	5.3	91.7	11.9
Maize flour	3.1	23.6	15.5	52.9	13.4
Fish	11.5	13.2	20.5	90.1	15.2
Bread	13.0	5.0	3.9	65.9	7.5
Rice	8.6	9.8	7.0	50.2	8.4
Meat	9.1	6.6	4.9	33.7	6.9
Oil & fats	3.6	5.4	2.6	58.8	3.8
Fruits	12.0	3.9	5.3	83.3	7.3
Sugar	3.5	3.3	3.4	52.1	3.4
Beans	3.6	5.7	4.6	54.8	4.6
Other staples	4.7	6.1	21.1	66.0	10.9
Other food	8.0	6.5	6.0	72.4	6.8
No. Obs. (N)	1720	1102	721	3543	3543

Number of goods consumed	Number of household consuming (full sample)	Share of household consuming (%) (full sample)
4	227	6.4
5	334	9.4
6	520	14.7
7	606	17.1
8	594	16.8
9	526	14.9
10	370	10.4
11	268	7.6
12	98	2.8

Source: IAF2002/03. Sample as explained in the main text.
Note: Shares in columns with the heading 'Share in total food expenditure' and 'Mean expenditure share' may not sum to 100 due to rounding.

The two last columns of the both the upper and lower part of Table 1 illustrate the need for a censored approach to estimate food demand for urban Mozambican households. While two food groups (vegetables and fish) are consumed by more than 90 percent of the households, most food groups have a substantial number of households with zero-purchases. Looking at the second half of Table 1 reveals that only around 3 percent of the households consume all 12 goods; again highlighting the relevance of taking zero consumption explicitly into account.

Apart from the dummy variables for location, south and north (central is the base specification), household size is also included as an additional explanatory variable to account for economies of scale in food preparation. There are other potential demographic variables which could be included, but their inclusion would add limited value to the focus of this analysis, and at the same time expand the already large number of parameters to be estimated.

Table 2 lists some summary statistics for the demographic and location dummy variables. The mode of the distribution of household size is five members, which constitutes 16 percent of all households. The sample is roughly equally divided between the three geographical areas.

Table 2. Sample means of demographic variables.

Variable	Description	Mean	Min	Max
Household size	Household size normalized at the mode of the sample distribution	5.06	1	10
South	Household located in south (omitted category)	0.36	0	1
Center	Household located in center (=1)	0.29	0	1
North	Household located in north (=1)	0.35	0	1

Source: IAF2002/03. Sample as explained in the main text.

3. Methodology

3.1 The translog demand system

The theoretical point of departure is the n-good indirect translog demand system proposed by Christensen, Jorgensen & Lau (1975)[5]. The flexible indirect utility function is logarithmic quadratic in normalized prices and has the form

$$\log \ V(p,x) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \log \frac{p_i}{x} + 1/2 \sum_{i}^{n} \sum_{j}^{n} \gamma_{ij} \log \frac{p_i}{x} \log \frac{p_j}{x} \tag{1}$$

where x, p_j is total expenditure and the price of good j, respectively. The unknown parameters are α_0, α_i and $\gamma_{ij}, i, j = 1,.....,n$. Since all prices are normalized by income, homogeneity of degree zero in prices and income is guaranteed. Share equations can be obtained by applying a logarithmic version of Roy's identity

$$w_i = \frac{\alpha_i + \sum_{j=1}^{n} \gamma_{ij} \log\left(p_j / x\right)}{1 + \sum_{k}^{n} \sum_{j}^{n} \gamma_{kj} \log\left(p_j / x\right)}, \quad i = 1,.....,n. \tag{2}$$

The normalisation $\sum_{j=1}^{n} \alpha_j = 1$, ensuring that the budget shares sum to one, has been imposed. The theoretical restriction of Slutsky symmetry can be implemented by the restriction $\gamma_{ij} = \gamma_{ji}$ for all *i,j*. Demographic and locational variables are incorporated through the α-parameters. Specifically, let z^h denote a *1xL* vector of household demographic and location dummy variables for household *h* with the elements denoted by z_l^h. The α-parameters can then be specified as $\alpha_i(z^h) = \alpha_{i0} + \sum_{l=1}^{L} \alpha_{il} z_l^h$. To maintain the adding-up property the following restrictions must be satisfied: $\sum_{i=1}^{n} \alpha_{i0} = 1, \quad \sum_{i=1}^{n} \alpha_{il} = 0, l = 1...L$.

[5] For recent applications of the indirect translog demand system see Yen, Fang & Su (2004) and Yen, Lin & Smallwood (2003).

Adding an error term, ε_i, and denote by Λ the set of all parameters, latent share equations can be written in the form

$$\bar{w}_i = w_i(p, x; \Lambda) + \varepsilon_i, \quad i = 1, \ldots\ldots, n,$$

where $w_i(p, x; \Lambda)$ is the deterministic component given by (2).

3.2 Accounting for censoring

Without restrictions on the error term the system of latent share equations cannot be a valid representation of observed behaviour since nothing constrains the shares to be non-negative. In particular, if a substantial number of households have zero consumption of some goods, the distribution of the error terms should allow for a positive probability of observing zero consumption. A number of methods have been proposed to deal with the issue of censoring. Wales and Woodland (1983) suggest a Kuhn-Tucker approach, whereby a utility function is maximized subject to non-negativity constraints on quantities. Lee and Pitt (1986) start from a random indirect utility function and use a dual approach related to the literature on rationing based on reservation prices, where demand for all goods depends on market prices for positively consumed goods and reservation prices for non-consumed goods. Because the rationed quantity is zero it is sometimes possible to solve for reservation prices explicitly. While theoretically appealing, these approaches suffer from the drawbacks that in the case of many non-consumed goods for some households, evaluation of multiple integrals is necessary. Further, as illustrated by Van Soest and Kooreman (1990), the issue of coherency of the solution has also to be addressed. Since the set of reservation (and market) prices that supports the observed behaviour may not be unique. In addition, for some flexible forms neither the Wales and Woodland nor the Lee and Pitt approaches are feasible.

An alternative solution, the Amemiya-Tobin approach, is inspired by Amemiya's (1974) multivariate regression with truncated normal distributions. Here the consumer is implicitly seen as maximizing a deterministic utility function and deviations from the corresponding deterministic shares are interpreted as random errors in the optimization process, measurement errors in the shares or random disturbances which influence the consumption decision (Wales and Woodland, 1983). However, like the Kuhn-Tucker approach, the implementation of Amemiya-Tobin type estimators is complicated by the need for evaluating multiple integrals in cases where censoring is severe. To circumvent the computational difficulties involved in the procedures described above, Shonkwiler and Yen (1999) suggested a two-step estimation procedure, where a probit regression is run in the first step for each equation to determine the pattern of censoring.[6] Using the estimated parameters in the second step, each latent share equation is augmented so as to take into account the censored nature of the data. The augmented system can then be estimated consistently with the (transformed) errors being normally distributed. The two-step nature of the estimator makes inference complicated by the need to adjust the covariance matrix as devised by Murphy and Topel (1985). Even so, the technique has been widely used in empirical applications (e.g. Yen, Kan and Su 2002).[7]

In this paper I follow the Amemiya-Tobin approach and specify the system of expenditure shares as a system of Tobit equations (see Yen, Lin and Smallwood 2003, Dong, Gould and Kaiser 2004, Yen and Lin 2002).[8] The method has the advantage of having potential fewer parameters for a given number of share equations than for example the two-step approach. For the study of regional difference in Mozambique a large (in terms of different food groups) demand system is essential in order to avoid that the aggregation of food groups is not obscuring possible differences among regions. To overcome the computational burden of simulating multiple integrals, a quasi maximum likelihood estimation technique recommended by Yen, Lin and Smallwood (2003) is employed.

[6] Alternatively, a multivariate probit can be run to facilitate cross equation correlation between the errors in the censoring mechanism, however, multivariate probit estimation requires evaluation of multiple integrals (or simulations hereof).
[7] The three approaches considered here have received much attention. However, other methods have recently been suggested in the literature, see Golan, Perloff and Shen (2001) and Perali and Chavas (2000).
[8] This treatment is not entirely innocuous. In particular, the Tobit formulation suffers the well-known limitation that the same process determines both the probability of censoring and the size of the expenditure share. Thus, a variable is constrained to influence the probability of censoring and the expenditure share in the same direction. Even with its drawbacks, this set-up has the advantage of saving on parameters.

3.3 Quasi maximum likelihood estimation

Write the observed shares, w_i^*, as a system of Tobit equations where joint normality of the error term, ε_i, is assumed

$$w_i^* = \max(w_i(p,x;\Lambda) + \varepsilon_i, 0), \quad i = 1,..,n \tag{3}$$

Even if the adding up condition is imposed on the deterministic shares, it may not hold for the observed shares in the model. To overcome this Pudney (1989) suggests treating the n'th good as a residual good and then obtain the elasticities from the following identity,[9]

$$w_n^* = 1 - \sum_{i=1}^{n-1} w_i^* .$$

To construct the likelihood function for the *n-1* estimated share equations, let the first k goods be the ones which are consumed in positive quantities by household h, and partition the *(n-1)x1* error vector into

$$\left[e_1 : e_2\right] \equiv \left[\varepsilon_1, \varepsilon_2, .., \varepsilon_k : \varepsilon_{k+1}, \varepsilon_{k+2}, ...\varepsilon_{n-1}\right]$$

and assume they are normally distributed with covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where Σ_{11} is a *kxk* matrix, Σ_{21} is a *(n-1-k)xk* matrix and Σ_{22} is a *(n-1-k)x(n-1-k)* matrix. The joint probability density function (pdf) of the errors can be written in terms of the joint marginal pdf for the first k errors and the pdf of the last *n-1-k* errors conditional on the first k errors, i.e.

$$f(e_1, e_2) = g(e_1) \times h(e_2 \mid e_1).$$

where $g(e_1)$ is the joint marginal pdf for the first k errors, distributed k-dimensional normal with zero mean and covariance matrix Σ_{11}. The joint pdf of e_2 conditional on e_1, $h(e_2 \mid e_1)$, can be shown to be *(n-1-k)*-dimensional normal with mean and covariance matrix given by (Greene, 2000)

$$\mu_{2.1} = \Sigma_{21}\Sigma_{11}^{-1} e_1$$
$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}'$$

---

[9] This has the consequence that the estimated parameters are not invariant to which good is excluded.

The likelihood contribution of household *h*, being in consumption regime *c*, i.e. where the first *k* goods are consumed, can be stated as

$$L_c^h = g(e_1) \int_{-\infty}^{-w_{n-1}(\Lambda)} \dots \int_{-\infty}^{-w_{l+2}(\Lambda)} \int_{-\infty}^{-w_{l+1}(\Lambda)} h(u_{k+1}, u_{k+2}, .., u_{n-1} \mid e_1) du_{k+1} du_{k+2} .. du_{n-1}$$

Using the expressions for the mean and covariance matrix given above, it is possible to evaluate the multiple integral as a standard *(n-k-1)*-dimensional normal pdf. Define the indicator function I_c^h as being one if household *h* is in consumption regime *c* and zero otherwise. The likelihood function for the sample of *N* households can now be written as

$$L = \prod_{h=1}^{N} \prod_{c} L_c^h (w_h(\Lambda))^{I_c^h} \tag{4}$$

Where $w_h(\Lambda)$ denotes the *(n-1)x1* vector of deterministic shares for household *h*.

While methods exist to simulate the likelihood function derived above (see Yen, Lin & Smallwood (2003) for a simulated maximum likelihood approach), they are computationally intensive when the number of non-consumed goods are large for a sizable part of the sample, as is the case for the present sample. Instead I adopt the quasi maximum likelihood (QML) procedure where the true likelihood function is approximated by linking bivariate Tobit models across the equations. More explicitly, the likelihood function given by (4) is approximated by multiplying all possible pair-wise combinations of bivariate Tobit likelihood functions. Define for household *h* and equation *i* and *j* the normalized errors; $u_{ih} = \left(w_{ih}^* - w_i(\Lambda)\right)/\sigma_i$ and $u_{jh} = \left(w_{jh}^* - w_j(\Lambda)\right)/\sigma_j$ where w_{ih}^* is the observed share for good *i* and σ_i is the standard deviation of the *i*'th error term. The likelihood contribution from the joint share equations *i,j* of household *h* takes the form of a bivariate Tobit likelihood function

$$
\begin{aligned}
L_{ij}^{h} = & \left[\Psi(u_{ih}, u_{jh}, \rho_{ij})\right]^{I(w_{ih}^{*}=0, w_{jh}^{*}=0)} \times \left[\sigma_j^{-1}\sigma_i^{-1}\psi(u_{ih}, u_{jh}, \rho_{ij})\right]^{I(w_{ih}^{*}>0, w_{jh}^{*}>0)} \times \\
& \left[\sigma_j^{-1}\phi(u_{jh})\Phi\left((u_{ih} - \rho_{ij}u_{jh})/(1-\rho_{ij}^{2})^{1/2}\right)\right]^{I(w_{ih}^{*}=0, w_{jh}^{*}>0)} \times \\
& \left[\sigma_i^{-1}\phi(u_{ih})\Phi\left((u_{jh} - \rho_{ij}u_{ih})/(1-\rho_{ij}^{2})^{1/2}\right)\right]^{I(w_{ih}^{*}>0, w_{jh}^{*}=0)}
\end{aligned}
$$

I(.) is an indicator function taking the value one if it evaluates to true, ϕ, Φ, are the standard normal pdf and standard normal cumulative distribution function (cdf), respectively. Similar, the bivariate standard normal pdf and cdf are denoted ψ and Ψ. The quasi likelihood function for the *N* households over all pair wise bivariate Tobits is given by

$$
L = \prod_{h=1}^{N}\prod_{i=1}^{n-2}\prod_{j=i+1}^{n-1} L_{ij}^{h} \tag{5}
$$

Cross-equation parameter restrictions on the demand system are easily accommodated via the linking of the pair wise Tobit likelihoods. Estimation proceeds by maximizing (5). The parameters obtained are consistent but will be less efficient than full information maximum likelihood estimation. In Monte Carlo simulations, Barslund (2006a) shows that the quasi maximum likelihood estimator yields parameter estimates very close to those of a simulation based full information maximum likelihood method.

3.4 Elasticities and decomposition of demographic effects

As pointed out by Lazaridis (2004), in a system of censored demand equations, price and expenditure elasticities should take into account not only the direct effect on the latent share but also the indirect effect from a possible change in the nature of censoring. The unconditional mean of observed shares is given by (Wooldridge 2002)

$$
E(w_i^{*}) = \Phi\left[w_i(\Lambda)/\sigma_i\right]w_i(\Lambda) + \sigma_i\phi\left[w_i(\Lambda)/\sigma_i\right]
$$

Income and uncompensated price elasticities for the system of equations can be written as

$$\eta_i^X = \frac{1}{Ew_i^*}\hat{\Phi}_i \cdot \frac{-\sum_{j=1}^{n}\hat{\gamma}_{ij} + w_i(\hat{\Lambda})\cdot\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\gamma}_{ij}}{1+\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\gamma}_{ij}\overline{\log}\left(p_j/x\right)}+1 \text{ and } e_{ij}^M = \frac{1}{Ew_i^*}\hat{\Phi}_i \frac{\hat{\gamma}_{ij} - w_i(\hat{\Lambda})\cdot\sum_{k=1}^{n}\hat{\gamma}_{kj}}{1+\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\gamma}_{ij}\overline{\log}\left(p_j/x\right)} - \delta_{ij}$$

where a hat denotes estimated parameters and $\hat{\Phi}_i = \Phi\left[w_i(\hat{\Lambda})/\hat{\sigma}_i\right]$. δ_{ij} is the kronecker delta taking the value one if $i=j$ and zero otherwise. A bar indicates that the logarithm is evaluated at the mean price and income for the location and demographic group in question and the superscript M denotes uncompensated (Marshallian) elasticities. Compensated price elasticities, ε_{ij}^C, can be calculated using the Slutsky equation $\varepsilon_{ij}^C = \varepsilon_{ij} + \eta_i^X Ew_i^*$, for all $i, j = 1,...,n.$

Differences in elasticities between demographic and geographical groups can be obtained by utilizing the augmentation of the α_i parameter (as described above). Let superscript r and a signify a reference and alternative demographic group, respectively, and make the convenient normalization of prices and incomes such that $\overline{\log}^r\left(p_j/x\right)=0$ for the reference group. The difference in expenditure elasticities for good i is then

$$\partial\eta_i^{X,r,a} = \eta_i^{X,r} - \eta_i^{X,a}$$

$$= \frac{1}{Ew_i^{*,r}}\hat{\Phi}_i^r \cdot \left[-\sum_{j=1}^{n}\hat{\gamma}_{ij} + w_i^r(\hat{\Lambda})\cdot\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\gamma}_{ij}\right] - \frac{1}{Ew_i^{*,a}}\hat{\Phi}_i^a \cdot \frac{-\sum_{j=1}^{n}\hat{\gamma}_{ij} + w_i^a(\hat{\Lambda})\cdot\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\gamma}_{ij}}{1+\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\gamma}_{ij}\overline{\log}^a\left(p_j/x\right)} \quad (6)$$

Analogous for own- and cross price elasticities. Confidence intervals facilitating statistical inference are obtained by the delta method.[10]

It is possible to decompose the elasticity differences into two components; a part stemming from demographic price differences and a part due to differences in α parameters. The relative importance of the first part can be judged by taking the difference between the

[10] An alternative method to get standard errors for the elasticities and their differences is to use a bootstrap approach. However, given the computational burden in estimating the system, this is not feasible here.

elasticity for the reference demographic group and the elasticity for the alternative demographic group evaluated at the parameters pertaining to the reference group and prices for the alternative group. Similarly, the difference associated with differences in α parameters is calculated by taking the difference in (6), but with the elasticity for the alternative demographic group evaluated at reference group prices. Because the elasticities are non-linear the two components will not in general sum to the total difference defined by equation 6 above. However, the ratio of the two components will illustrate their relative importance in contributing to the total difference. The results will focus on differences in elasticities with respect to geographical location, but differences between any two demographic groups can be analysed using the framework presented here.

An obvious alternative to analysing regional differences in elasticities would be to estimate three models separately and compare the estimated elasticities. An advantage is that it would not be necessary to restrict the price response parameters, γ_{ij}'s, in (2) to be equal over regions.[11] Although attractive, this would reduce the sample considerably for each region and, contrary to the method pursued here, it would not allow for a statistical test of difference between the regions.

4. Results

4.1 Elasticities for central Mozambique

All estimations are carried out using household weights provided by the National Statistical Office in Mozambique. Standard errors are robust to clustering at the enumeration area, and stratification of the sample.[12] In total, 188 parameters were estimated. A full list of coefficients is relegated to appendix B.

[11] In principle it is possible to augment each γ_{ij} in (2) with regional dummy variables but the number of parameters to estimate would be unmanageable in practical terms.
[12] All estimations were done in the software package Stata 9.2 using the command qmldemand_tl.ado (see appendix A).

The reference household considered in the following has five members and comes from the central part of the country. I compare the elasticities from this group of households with two households with an equal number of members in respectively the south and north of the country. Income and all prices are normalised such that the average over the households in the reference group is one (i.e. the logarithm of income, prices and prices over income are zero).

As expected, a substantial number of parameters are estimated with good precision. In total 111 of the 188 (56 %) estimated parameters are significant at the five percent level. Of the 78 price response parameters (γ's) 35 (45 %) are significant at the five percent level, and 45 out of 66 estimated covariance parameters (ρ's and σ's) are significant. Inclusion of the demographic variables is also warranted from the results; 20 out of 33 are significant. All but one of the 11 α_{i0}'s are significant at 5 percent. Table 3 shows the size and significance of the demographic variables.

Table 3. Size and significance of demographic and geographic variables.

Equation	Household size	South	North
Vegetables	0.005***	0.078***	-0.063***
	(0.001)	(0.008)	(0.010)
Maize flour	-0.003	-0.296***	-0.041
	(0.004)	(0.029)	(0.026)
Fish	0.005***	-0.062***	0.072***
	(0.002)	(0.011)	(0.015)
Bread	0.004***	0.049***	-0.009
	(0.001)	(0.011)	(0.014)
Rice	-0.005*	-0.049*	-0.023
	(0.003)	(0.026)	(0.024)
Meat	0.011	0.043	-0.074
	(0.008)	(0.026)	(0.050)
Oil & fats	0.001	-0.056***	-0.040***
	(0.001)	(0.008)	(0.007)
Fruits	0.002***	0.086***	0.033***
	(0.001)	(0.011)	(0.008)
Sugar	-0.001	-0.024***	0.002
	(0.001)	(0.001)	(0.008)
Beans	0.004***	-0.036***	-0.037***
	(0.001)	(0.010)	(0.012)
Other staples	-0.010***	0.010	0.132***
	(0.003)	(0.021)	(0.029)

Note:
*, **,*** denotes significance at 10, 5 and 1 percent level. Standard errors in parenthesis. Household size normalised at 5 members (the mode of the distribution). That is, effects of household size are calculated relative to a household with 5 members. Effects are on the deterministic shares (equation 2).

Each column shows the effect from the demographic variable on the deterministic share (see equation 2) relative to a reference household.[13] Looking horizontally across the table it is clear that inclusion of household size and the dummy variables for location is warranted. Only the meat food group is unaffected by any of the three variables (none of them enters significantly). Focusing on the differences between the central part and the north and south of Mozambique the table illustrates some striking differences. For vegetables and fish both the north and south dummy variable are significant and with opposite signs, reflecting the observations in Table 1 and signifying that there are significant differences in consumption shares between the three regions for these food items even when possible price differences have been accounted for. The expenditure share for oil and fats, fruits, and beans in the north and south is significantly different from the central part of Mozambique, but since they have the same sign it is not possible to assess if they also differ between the north and south. In the south all other food groups except meat and other staples have significantly different expenditure shares relative to the central part. The results listed in Table 3 point in the direction of different elasticities between the regions in Mozambique. This is explored further below.

The precision with which the coefficients are estimated is expected to carry over to small standard errors around the estimated elasticities. Table 4, which shows estimated elasticities and their standard errors for a reference household in central Mozambique, confirms this.

Of the 121 estimated price elasticities 52 are significant at the 5 percent level and a further 11 are significant at the 10 percent level.[14] All 11 expenditure elasticities are significantly different from zero. However, more interestingly; seven are different from one at the 5 percent significance level.

[13] Recall from (3) that an increase in the deterministic share (and as a result the latent share) has two effects; it increases the probability of the household consuming the good (non-censoring) and it increases the expenditure share given consumption of the good.

[14] Since focus is exclusively on food consumption all elasticities are conditional elasticities and, thus, expenditure elasticities are measured with respect to total food expenditures.

Table 4. Estimated uncompensated elasticities for a reference household.[a)]

	Vegetables	Maize flour	Fish	Bread	Rice	Meat	Oil & fats	Fruits	Sugar	Beans	Other staples
Vegetables	-0.84***	-0.08	0.12***	0.23***	-0.10	0.02	0.02	0.03	0.00	-0.03	-0.04
Maize flour	-0.09***	-0.59***	-0.02	-0.04	-0.10*	0.06*	-0.04	0.01	-0.07***	-0.17***	-0.15***
Fish	0.07***	0.05	-1.05***	0.05	-0.18***	0.01	0.04	-0.04**	-0.03	0.17***	-0.06
Bread	0.28***	-0.15	0.05	-1.37***	-0.54***	-0.10*	0.02	-0.03	-0.01	0.12	0.35***
Rice	-0.11**	-0.20**	-0.21***	-0.27***	-0.82***	0.12***	-0.08	-0.09**	0.07	0.06	0.16***
Meat	0.03	0.01	0.01	-0.06**	0.07**	-1.10***	0.00	0.03	0.01	0.05*	-0.15***
Oil & fats	0.01	-0.10	0.06	0.03	-0.16	0.03	-1.34***	0.09**	0.02	0.11*	0.11***
Fruits	0.05	0.08	-0.09**	-0.03	-0.22**	0.07	0.11***	-1.00***	0.05	0.15***	-0.16***
Sugar	-0.03	-0.34***	-0.10*	-0.01	0.17	0.06	0.01	0.04	-1.29***	-0.03	0.11*
Beans	-0.05	-0.45***	0.29***	0.11*	0.14	0.09**	0.09*	0.11***	-0.02	-1.22***	0.04
Other staples	-0.05	-0.33***	-0.10*	0.22***	0.21***	-0.20***	0.05**	-0.09***	0.05*	0.02	-0.91***
Expenditure elasticities	0.66***	1.31***	0.86**	1.23***	1.30***	1.10	1.07	0.92	1.29***	0.89**	1.18

[a)] Reference household: 5 household members, located in central Mozambique.

Notes:

The table show percentage points change in demand for the row good when the price of the column good changes by 1 percent.

*, **,*** denotes significance at 10, 5 and 1 percent level. For price elasticities the significance is with respect to difference from 0.

For expenditure elasticities the significance is with respect to differences from 1. All expenditure elasticities are significantly different from zero at any conventional level.

All standard errors calculated by the delta method. They are not reported but available from the author.

Looking at the estimated uncompensated own price elasticities, bread, oil and fats, sugar, beans, fish, meat and fruits are price elastic – with the first four food groups significantly greater than minus one in absolute terms – while fruits, fish and meat have uncompensated own prices around minus one. Vegetables and maize flour are both significantly price inelastic. While the point estimate for rice and other staples suggest they are price inelastic, although the confidence interval is too large to say so significantly. The cross price elasticities are generally smaller than own price elasticities in absolute value. However, for some goods there are sizeable cross price effects. This is especially valid for maize flour, rice, beans and other staples, which all have relatively large and significant cross price effects. A majority of food groups are gross complements as is often found in food demand studies (see Yen, Lin & Smallwood 2003, Dong, Gould & Kaiser 2004) and the absolute size also conforms well to these studies.

Most food groups have one or more gross substitutes except for maize flour and sugar. That might be expected given the importance of this ingredient in the food basket for households located in the central part of the country, cf. Table 1. Sugar has in general few natural substitutes. Notable significant gross substitutes are found between vegetables and bread and fish. Fish and beans together with rice and other stables are also gross substitutable. There is no evidence of meat and fish being (significant) gross substitutes as might be expected by these two important sources of protein. The other main protein source, beans is a gross substitute for both fish and meat.

The food expenditure elasticities reveal four food groups, namely vegetables, fish, fruits and beans to be necessities and the remaining seven, maize flour, bread, rice, meat, oil and fats, sugar and other staples to be luxury food items. As noted earlier seven of these pass significance tests (in terms of luxuries and necessities) at five percent. These findings conform reasonably well to prior expectations. The fact that maize flour and rice come out as luxuries reflect the small incomes which, despite the good economic performance of the Mozambique economy recently, many urban families still have to get by on (MPD 2004). It is surprising that the category other staples, containing mainly cassava and potatoes, is not a necessity. It might be that the category is so broad that items are used for food variety by wealthier households. It also reflects to some extent the choice of the central part of the country as the reference location. Table 1 revealed that other staples is a much bigger category in the northern part of Mozambique and Table 3 showed a large significant coefficient on the northern location dummy for the food category of other staples.

Turning to the compensated price elasticities, these are displayed in Table 5 for completeness. The compensated own price effects are all negative and smaller (in absolute value) than their uncompensated counterpart as would be expected from demand theory. While most food products are gross complements Table 5 shows that if households are compensated for price changes most products become net substitutes.

4.2 Regional differences in elasticities

I now proceed to evaluate the impact of geography on food demand elasticities. As noted above for a majority of food groups the dummy variables for location (North and South) are significant, and therefore expenditure shares differ between regions once possible relative price differences have been accounted for. Thus, given the utility function, the estimated parameters can be used for partial welfare analysis of relative price changes and the effects are allowed to differ between the demographic groups included in the estimation. However, for policy analysis in general and for equilibrium analysis in CGE models in particular, the parameters of interest are often the elasticities. It therefore becomes of interest to investigate to what extent the share differences carry over to differences in regional elasticities and whether these differences are significant. This is pursued here using the methodology discussed in section 3.3.

Table 6 shows expenditure and own price elasticities for central and southern Mozambique measured for a reference household (i.e. with five household members), their differences and the standard error of the differences. The difference is calculated by (6) and associated standard errors are obtained by the delta method.

Table 5. Estimated compensated elasticities for a reference household.[a)]

	Vegetables	Maize flour	Fish	Bread	Rice	Meat	Oil & fats	Fruits	Sugar	Beans	Other staples
Vegetables	-0.77	0.07	0.21	0.27	-0.04	0.07	0.06	0.06	0.03	0.01	0.01
Maize flour	0.05	-0.30	0.15	0.03	0.03	0.15	0.04	0.06	-0.02	-0.09	-0.05
Fish	0.17	0.24	-0.93	0.09	-0.10	0.07	0.09	0.00	0.01	0.23	0.00
Bread	0.42	0.12	0.22	-1.30	-0.42	-0.02	0.09	0.02	0.04	0.19	0.44
Rice	0.03	0.09	-0.03	-0.20	-0.69	0.21	-0.01	-0.04	0.12	0.14	0.26
Meat	0.15	0.25	0.16	0.00	0.18	-1.03	0.07	0.07	0.05	0.12	-0.07
Oil & fats	0.12	0.14	0.20	0.09	-0.05	0.10	-1.27	0.14	0.06	0.18	0.19
Fruits	0.15	0.28	0.04	0.02	-0.12	0.13	0.16	-0.96	0.08	0.21	-0.09
Sugar	0.11	-0.06	0.07	0.06	0.30	0.15	0.08	0.10	-1.24	0.05	0.20
Beans	0.05	-0.25	0.41	0.16	0.23	0.15	0.14	0.15	0.02	-1.17	0.11
Other staples	0.07	-0.08	0.06	0.28	0.33	-0.11	0.12	-0.04	0.10	0.10	-0.82

[a)] Reference household: 5 household members, located in central Mozambique.

Notes:

The table show percentage points change in compensated demand for the row good when the price of the column good changes by 1 percent.

Table 6. Elasticity differences between the central and southern part of Mozambique.[a)]

	Expenditure elasticities					Own price elasticities				
	Central	South	Difference (Total)	Difference (Price effect)	Difference (Parameter effect)	Central	South	Difference (Total)	Difference (Price effect)	Difference (Parameter effect)
Vegetables	0.66*** (0.04)	0.70*** (0.08)	-0.04 (0.11)	0.00	-0.03	-0.84*** (0.04)	-0.91*** (0.03)	0.07 (0.13)	0.01	0.07
Maize flour	1.31*** (0.05)	2.03 (1.09)	-0.72 (1.09)	-0.08	-0.58	-0.59*** (0.09)	-0.53*** (0.16)	-0.06 (0.08)	-0.02	-0.07
Fish	0.86** (0.06)	0.88 (0.11)	-0.02 (0.17)	0.02	-0.06	-1.05 (0.05)	-1.05 (0.05)	0.01 (0.01)	-0.01	0.02
Bread	1.23*** (0.05)	1.00 (0.06)	0.23** (0.11)	0.15	0.14	-1.37** (0.17)	-1.20** (0.10)	-0.17** (0.07)	-0.10	-0.09
Rice	1.30*** (0.05)	1.36*** (0.06)	-0.06 (0.10)	0.03	-0.10	-0.82 (0.15)	-0.82 (0.15)	-0.00 (0.01)	0.00	-0.00
Meat	1.10 (0.24)	1.08 (0.30)	0.02 (0.53)	-0.02	0.05	-1.10 (0.09)	-1.10 (0.08)	-0.00 (0.01)	0.00	-0.00
Oil & fats	1.07 (0.04)	1.19** (0.08)	-0.12 (0.11)	0.06	-0.25	-1.34*** (0.09)	-1.42*** (0.11)	0.09*** (0.03)	-0.06	0.16
Fruits	0.92 (0.08)	0.80*** (0.07)	0.12 (0.14)	0.03	0.12	-1.00 (0.09)	-1.00 (0.04)	0.00 (0.04)	0.00	0.00
Sugar	1.29*** (0.05)	1.35*** (0.07)	-0.06 (0.12)	0.05	-0.13	-1.29*** (0.11)	-1.31*** (0.11)	0.02 (0.02)	-0.02	0.05
Beans	0.89** (0.05)	1.01 (0.06)	-0.12 (0.11)	-0.04	-0.07	-1.22*** (0.08)	-1.28*** (0.10)	0.06** (0.03)	0.02	0.04
Other staples	1.18 (0.13)	1.25* (0.13)	-0.07 (0.13)	-0.09	0.02	-0.91 (0.07)	-0.90 (0.08)	-0.00 (0.01)	-0.00	0.00

[a)] Measured at the reference household, i.e. for each region a household with: 5 household members.

Notes:

Standard errors are calculated by the delta method and shown in parenthesis. *, **,*** denotes significant different from zero at 10, 5 and 1 percent level. For price elasticities the significance is with respect to difference from 0. For expenditure and own price elasticities the significance is with respect to differences from 1. All expenditure and own price elasticities are significantly different from zero at any conventional level.

Looking first at expenditure elasticities (the five first columns), note that roughly the same goods are necessities and luxuries in both areas – with the exception of beans and possibly bread. Neither bread nor beans are estimated with enough precision in the south to reject that bread is a luxury and beans a necessity as is the case in central Mozambique. There are sizeable differences in elasticities (absolutely greater than 0.1) for maize flour, bread, fruits, beans and oil and fats. However, only for bread is the difference significant at 5 percent. In the south the point estimate of the expenditure elasticity of maize flour is very high at 2.03 – but the confidence interval around this estimate is correspondingly large (not shown) implying no significant difference between the central and southern estimates. For vegetables, meat and fish the differences are small and insignificant. Note though, that the point estimates for vegetables still differ with around 6 percent between central and south. For maize flour, bread, oil and fats, fruits and beans the point estimates differ with 10 percent or more, making it clear that policies aimed at or implying price changes for these food groups will have different impact in the two regions.

The last two columns of the first part of Table 6 attempt to separate the total differences in expenditure elasticities into respectively a price effect and a parameter effect as described in section 3.3. Informally, the price effect can be interpreted as the difference that would have been observed had the equations been estimated without the location dummy variables.[15] It is the effect of households being at different points on the same demand curve due to price differences alone. Although the two effects do not add to the total differences – as expected because of the non-linearities in the system – they nonetheless seem sensible in size and direction. As an example take sugar, where the price effect suggests that the expenditure elasticity should have been higher in the central than in the southern part of the country. The observed difference is negative and to reconcile that with the positive price effect the parameter effect must be large and negative relative to the price

[15] Estimating the model without location dummy variables would have resulted in another set of parameter estimates, possibly, with differences of another magnitude, thus clearly, one has to be cautious in interpreting the reported differences. On the other hand as argued in section 3.3, the relative sizes of the price effects and the parameter effects still convey useful information as to the source of the observed differences.

effect – as is indeed the case. In general the parameter effects are much larger than the price effects suggesting that differences in expenditure elasticities are due to differences in parameters rather than to differences in prices. For bread and other staples this is not the case, though. Here relative price differences play a role.

The second part of Table 6 concentrates on own price elasticities. While the same exercise has been done for cross price elasticities only differences in own price elasticities are reported to keep the focus and to have a manageable amount of output. Price elastic and inelastic food groups do not differ between the two regions. In general the size of the differences is smaller than for expenditure elasticities and for fish, rice, meat, fruits, sugar and other staples there are virtually no differences between elasticities. However, for two food groups, bread and oil and fats, differences are sizeable and significant. For beans the difference is small, but estimated with good precision so that the difference in own price elasticities is significant at 5 percent. Parameter effects explain most of the differences. Only for bread and oil and fats is the magnitude of the price effect non-negligible. Since own price elasticities constitute a combined substitution and income (expenditure in the terminology here) effect it would be preferable to observe large price effects in the differences between own price elasticities where large price effects were observed for expenditure elasticities; at least for some goods. This to some extent is the case as illustrated by the food groups of bread and oil and fats. However, for maize flour and other staples the price effect fail to show up in the differences in own price elasticities. This could be because the substitution and income effects cancel out the price effect difference.

To sum up, differences in expenditure and own price elasticities between south and central Mozambique are generally small in absolute size. For a few goods, notably vegetables and bread, the size and significance of the absolute differences warrant the use of different elasticities for the two regions.

Table 7. Elasticity differences between the central and northern part of Mozambique.[a)]

	Expenditure elasticities					Own price elasticities				
	Central	North	Difference (Total)	Difference (Price effect)	Difference (Parameter effect)	Central	North	Difference (Total)	Difference (Price effect)	Difference (Parameter effect)
Vegetables	0.66***	0.63***	0.03	0.01	0.00	-0.84***	-0.74***	-0.10	0.01	-0.09
	(0.04)	(0.09)	(0.11)			(0.04)	(0.06)	(0.14)		
Maize flour	1.31***	1.43***	-0.12	-0.08	-0.06	-0.59***	-0.54***	-0.05**	-0.03	-0.01
	(0.05)	(0.08)	(0.13)			(0.09)	(0.11)	(0.02)		
Fish	0.86**	0.82*	0.04	0.00	0.04	-1.05	-1.04	-0.01	0.01	-0.01
	(0.06)	(0.11)	(0.16)			(0.05)	(0.04)	(0.01)		
Bread	1.23***	1.31***	-0.08	-0.01	-0.03	-1.37**	-1.43**	0.06**	0.02	0.02
	(0.05)	(0.08)	(0.12)			(0.17)	(0.18)	(0.03)		
Rice	1.30***	1.40***	-0.10	-0.04	-0.05	-0.82	-0.81	-0.01	0.00	0.00
	(0.05)	(0.06)	(0.10)			(0.15)	(0.17)	(0.02)		
Meat	1.10	1.23	-0.12	-0.06	-0.09	-1.10	-1.12	0.02	0.02	0.01
	(0.24)	(0.34)	(0.58)			(0.09)	(0.13)	(0.04)		
Oil & fats	1.07	1.25***	-0.18	0.00	-0.16	-1.34***	-1.48***	0.14***	0.02	0.11
	(0.04)	(0.07)	(0.11)			(0.09)	(0.13)	(0.04)		
Fruits	0.92	0.86	0.07	0.00	0.07	-1.00	-1.00	0.00	0.00	0.00
	(0.08)	(0.09)	(0.17)			(0.09)	(0.08)	(0.01)		
Sugar	1.29***	1.32***	-0.03	-0.02	0.01	-1.29***	-1.31***	0.02	0.01	0.00
	(0.05)	(0.09)	(0.13)			(0.11)	(0.11)	(0.02)		
Beans	0.89**	0.93	-0.03	0.05	-0.07	-1.22***	-1.26***	0.04	-0.02	0.04
	(0.05)	(0.06)	(0.11)			(0.08)	(0.10)	(0.03)		
Other staples	1.18	0.98	0.20	0.00	0.19	-0.91	-0.91*	0.01	-0.01	0.01
	(0.13)	(0.13)	(0.17)			(0.07)	(0.05)	(0.02)		

[a)] Measured at the reference household, i.e. for each region a household with: 5 household members.

Notes:

Standard errors are calculated by the delta method and shown in parenthesis. *, **,*** denotes significant different from zero at 10, 5 and 1 percent level. For price elasticities the significance is with respect to difference from 0. For expenditure and own price elasticities the significance is with respect to differences from 1. All expenditure and own price elasticities are significantly different from zero at any conventional level.

Table 7 is equivalent to Table 6 and shows differences between central and northern Mozambique. Luxuries and necessities do not differ between the two regions – except for other staples, where the point estimate shows it to be a necessity in the north. With a point estimate of 0.98 and a standard error of 0.13 this is a borderline case. There are sizeable differences in expenditure elasticities for oil and fats and other staples (respectively, -0.18 and 0.20), although the confidence intervals surrounding them are too large to make them significant at any level. This to some extent reflects the rather strong data requirement to estimate differences with precision. Even when they are not significant the size of the differences in point estimates are worth taking into account when analysing regional changes in food consumption – whether it is by using different regional expenditure elasticities or as part of a sensitivity analysis. The price effect is of relative importance for maize flour, meat and beans, whereas to the extent there are differences in expenditure elasticities for the other food groups these are driven by parameter effects. As for own price elasticities, the size of the differences resemble those found between the central and southern part of Mozambique. For maize flour, bread and oil and fats the differences are significant. The price effects are small for all food groups and nowhere larger than 0.03 in absolute sizes. Except for vegetables (where the total difference is not significant) and oil and fats the parameter effects are equally quite small.

5. Conclusion

Even though Mozambique has managed to reduce poverty considerably during the past decade, it is still a poor country where around 50 percent of expenditures are directed towards food consumption. Therefore food prices, food demand and nutrition are among the key elements when discussing household welfare, and in particular changes herein. However, little is known about demand elasticities for core food groups in Mozambique. In this paper a large food demand system has been estimated for a nationally representative sample of urban Mozambican households. The issue of censoring of the expenditure shares has been addressed by a recently suggested maximum likelihood estimator. Because of Mozambique's varied geography and limited economic integration across regions the focus is on regional differences. The novel estimates illuminate some interesting differences in expenditure shares between demographic and geographic groups. Further, expenditure and own price elasticities are presented for respectively northern, central and southern Mozambique, and a test for statistical significance between regions is developed. In particular for own price elasticities there are significant – both in a statistical and a quantitative sense – differences between the central and the south and north of Mozambique. Apart from being the first estimates using micro data for Mozambique the findings are useful for developers of CGE models with a regional aspect and for the evaluation of policies that alter relative food prices.

References

Amemiya, T., 1974, ‘Multivariate Regression and Simultaneous Equation Models when the Dependent Variables Are Truncated Normal’, *Econometrica* 42, Iss. 6, pp. 999-1012.

Barslund, M., 2006a, ‘Estimation of Tobit type censored demand systems: A comparison of estimators’, WP.

Barslund, M., 2006b, ‘Qmldemand_x: A Stata command to estimate indirect translog and almost ideal censored demand systems using Quasi Maximum Likelihood (QML)’, WP.

Christensen L.R., Jorgenson, D.W. & Lau, L.J., 1975, ‘Transcendental Logarithmic Utility Functions’, *American Economic Review 65*, pp. 367-383.

Deaton, A., 1987, ‘Estimation of own and cross price elasticities from household survey data’, *Journal of Econometrics.* 36, pp. 7-30.

Dong, D., Gould, BW. & Kaiser, HM., 2004, ‘Food Demand in Mexico: An Application of the Amemiya-Tobin Approach to the Estimation of a Censored Food System’, *American Journal of Agricultural Economics* 86, Iss. 4, pp. 1094-1107.

Fox, L., Bardasi, E. & Van den Broeck, K., 2005, ’Poverty in Mozambique: Unravelling changes and determinants’, *World Bank, Africa Region Working Paper Series*, No. 87.

Golan, A,, Perloff, J.M. & Shen, E.Z., 2001, ‘Estimating a demand system with non-negativity constraints: Mexican Meat Demand’, *The Review of Economics and Statistics* 83, Iss. 3, pp. 541-550.

Greene, W.H., 2000, *‘Econometric Analysis’*, Prentice Hall, USA.

INE, Instituto Nacional de Estatística, 2004. *Inquérito Nacional aos Agregados Familiares Sobre Orçamento Familiar 2002/3*. Maputo: INE.

Jensen, H.T. & Tarp, F., 2004, ‘On the Choice of Appropriate Development Strategy: Insights from CGE Modelling of the Mozambican Economy’, *Journal of African Economies* 13, Iss. 3, 2004, pp. 446-478.

Kedir, A.M., 2005, ‘Estimation of own- and cross-price elasticities using unit values: Econometric issues and evidence from urban Ethiopia’, *Journal of African Economies* 14, Iss. 1, pp. 1-20.

Lazaridis, P., 2004, ‘Demand elasticities derived from consistent estimation of Heckman-type models’, *Applied Economics Letters* 11, pp. 523-527.

Lee, F. & Pitt, M.M., 1986, ‘Microeconometric demand systems with binding nonnegativity constraints: the dual approach’, *Econometrica* 54, No. 5 (September), pp. 1237-1242.

MPD, 2004, ‘Poverty and well-being in Mozambique: The second national assessment’, Ministry of Planning and Development, Discussion Paper 3E, Maputo.

Murphy, K.M. & Topel, R.H., 1985, ‘Estimation and Inference in Two-Step Econometric Models, *Journal of Business & Economic Statistics* 3, Iss. 4, pp. 370-379.

Perali, F. & Chavas, J-P., 2000, ‘Estimation of Censored Demand Equations from Large Cross-Section Data’, *American Journal of Agricultural Economics* 82, Iss. 4, pp. 1022-1037.

Pudney, S., 1989, '*Modelling Individual Choice: Econometrics of Corners, Kinks and Holes*', Blackwell, 1989.

Shonkwiler, J.S. & Yen, S.T., 1999, 'Two-Step Estimation of a Censored System of Equations', *American Journal of Agricultural Economics* 81, Nov, pp. 972-82.

Stata Base Reference Manual, Stata Press, Stata Corporation.

Van Soest, A. & Kooreman, P., 1990, 'Coherency of the indirect translog demand system with binding nonnegativity constraints', *Journal of Econometrics 44*, pp. 391-400.

Wales, T.J. & Woodland A.D., 1983, 'Estimation of consumer demand systems with binding non-negativity constraints', *Journal of Econometrics* 21, pp. 263-285.

Wooldridge, J.M., 2002, '*Econometric analysis of cross section and panel data*', MIT press.

Yen, S.Y., Fang, C.. & Su, S-J., 2004, 'Household food demand in urban China: a censored system approach', *Journal of Comparative Economics* 32, pp. 564-585.

Yen, S.Y., Kan, K. & Su, S-J., 2002, 'Household demand for fats and oils: two-step estimation of a censored demand system', *Applied Economics* 14, pp. 1799-1806.

Yen, S.T. & Lin, B-H, 2002, 'Beverage consumption among US children and adolescents: full-information and quasi maximum-likelihood estimation of a censored system', *European Review of Agricultural Economics* 29*,* Iss. 1, pp. 85-103.

Yen, S.T., Lin, B-H. & Smallwood, D.M., 2003, 'Quasi- and simulated-likelihood approaches to censored demand systems: food consumption by food stamp recipients in the United States', *American Journal of Agricultural Economics* 85, Iss. 2, pp. 458-478.

APPENDIX A

QMLdemand_tl: A Stata command to estimate indirect translog demand system using Quasi Maximum Likelihood (QML)

This appendix documents the Stata command qmldemand_tl, which implement a quasi maximum likelihood routine for estimating the indirect translog. The censoring is assumed to be of the multivariate Tobit type. Adding-up is guaranteed by treating the last share as a residual. The command supports augmenting the parameters with demographic variables. For a more indept treatment see Barslund, 2006b. The command estimates a system of equations of the form given in the main text (equation 3) using the QML estimator explained therein (equation 5). Below follows the syntax with examples.

Syntax

The syntax differs slightly from the general syntax for Stata commands. All parameters are identified from the knowledge of observed expenditure shares (*exp_share#*), logarithmic prices (*log_p#*) and total logarithmic expenditure (*log_exp*).

qmldemand_*x exp_sharel exp_share2 ..... exp_shareM log_p1 logp2 ... logpM log_exp*
[weight] [if exp] [in range]
[,alpha(*varlist*) beta(*varlist*) first from(*matname*) robust cluster(*varname*)
constraints(*numlist*) *maximize_options*]

Weights: pweights, fweights, aweights, and iweights are allowed.

Options

`alpha(varlist)` specifies a list of variables to augment the α-parameter. Each variable specified will increase the number of estimated parameters by the number of equations minus one.

`beta(varlist)` (only for qmldemand_ai) specifies a list of variables to augment the β-parameter. Each variable. specified will increase the number of estimated parameters by the number of equations minus one.

`first` tells Stata to run an auxiliary system estimation without price response coefficients ($\gamma's$) to obtain starting values before running the full system. Useful if no good initial values are available.

`from(matname)` specify a matrix of starting values. Options for `ml init` can be specified inside the parenthesis. If `first` is specified, from(matname) relates to starting values for the auxiliary system estimation without price response coefficients (see `first`).

Other options are normal Stata maximum likelihood options (See Stata Base Reference Manual).

`robust` specifies that the Huber/White/sandwich estimator of variance is used.

`cluster(varname)` allows non-independent observations within groups (defined by varname). Observations are independent between groups. Cluster implies robust.

`constraint(numlist)` specifies a list of linear parameter constraints.

`maximize options` general ML options available in Stata (see Stata Base Reference Manual).

Examples

qmldemand_tl s_1 s_2 s_3 s_4 s_5 log_p1 log_p2 log_p3 log_p4 log_p5 lnx [aw=hhweight], alpha(hhsize gender) first cluster(ea)

qmldemand_ai s_1 s_2 s_3 s_4 s_5 log_p1 log_p2 log_p3 log_p4 log_p5 lnx [aw=hhweight], alpha(hhsize gender) beta(hhsize) from(start, copy) cluster(ea)

Note that all M shares have to be specified, although only $M-1$ are actually used in the estimation procedure.

Appendix B. Estimated parameters.

Coefficient	Estimate	Std. Err.	p-value
γ11	0.021	0.005	0.000
γ12	-0.023	0.007	0.003
γ13	0.014	0.005	0.003
γ14	0.025	0.005	0.000
γ15	-0.019	0.009	0.033
γ16	0.004	0.008	0.646
γ17	0.001	0.003	0.683
γ18	0.004	0.006	0.556
γ19	-0.002	0.003	0.571
γ110	-0.004	0.004	0.366
γ111	-0.007	0.005	0.169
γ112	0.005	0.005	0.287
γ22	0.094	0.025	0.000
γ23	-0.007	0.010	0.454
γ24	-0.016	0.008	0.039
γ25	-0.039	0.017	0.023
γ26	0.018	0.018	0.315
γ27	-0.013	0.007	0.053
γ28	0.003	0.006	0.608
γ29	-0.025	0.006	0.000
γ210	-0.048	0.010	0.000
γ211	-0.048	0.011	0.000
γ212	-0.021	0.008	0.007
γ33	-0.008	0.007	0.275
γ34	0.004	0.006	0.512
γ35	-0.036	0.009	0.000
γ36	0.002	0.010	0.804
γ37	0.004	0.004	0.308
γ38	-0.006	0.002	0.011
γ39	-0.007	0.004	0.082
γ310	0.028	0.004	0.000
γ311	-0.013	0.008	0.078

Coefficient	Estimate	Std. Err.	p-value
γ312	0.022	0.005	0.000
γ44	-0.033	0.015	0.026
γ45	-0.049	0.011	0.000
γ46	-0.009	0.006	0.120
γ47	0.001	0.006	0.811
γ48	-0.003	0.003	0.378
γ49	-0.001	0.009	0.897
γ410	0.010	0.006	0.096
γ411	0.030	0.007	0.000
γ412	0.014	0.005	0.002
γ55	0.029	0.026	0.270
γ56	0.021	0.007	0.003
γ57	-0.015	0.009	0.097
γ58	-0.016	0.007	0.016
γ59	0.011	0.010	0.307
γ510	0.011	0.011	0.302
γ511	0.028	0.009	0.001
γ512	0.015	0.007	0.035
γ66	-0.022	0.012	0.060
γ67	0.003	0.003	0.326
γ68	0.005	0.005	0.269
γ69	0.004	0.003	0.124
γ610	0.009	0.005	0.095
γ611	-0.027	0.007	0.000
γ612	-0.003	0.009	0.751
γ77	-0.026	0.007	0.000
γ78	0.007	0.003	0.009
γ79	0.001	0.006	0.921
γ710	0.008	0.005	0.080
γ711	0.007	0.003	0.029
γ712	0.007	0.002	0.000
γ88	0.000	0.006	0.962

Coefficient	Estimate	Std. Err.	p-value
γ89	0.003	0.003	0.326
γ810	0.011	0.003	0.000
γ811	-0.012	0.004	0.004
γ812	0.006	0.003	0.044
γ99	-0.020	0.008	0.007
γ910	-0.002	0.004	0.587
γ911	0.007	0.004	0.120
γ912	0.010	0.003	0.003
γ1010	-0.022	0.008	0.006
γ1011	0.003	0.005	0.500
γ1012	-0.002	0.003	0.569
γ1111	0.013	0.010	0.202
γ1112	-0.008	0.007	0.207
γ1212	0.001	0.015	0.937
σ1	0.079	0.003	0.000
σ2	0.256	0.011	0.000
σ3	0.118	0.005	0.000
σ4	0.096	0.003	0.000
σ5	0.203	0.007	0.000
σ6	0.306	0.029	0.000
σ7	0.070	0.003	0.000
σ8	0.081	0.004	0.000
σ9	0.079	0.003	0.000
σ10	0.102	0.004	0.000
σ11	0.166	0.013	0.000
ρ12	-0.188	0.039	0.000
ρ13	-0.087	0.028	0.002
ρ14	-0.020	0.031	0.515
ρ15	-0.193	0.039	0.000
ρ16	-0.096	0.035	0.006
ρ17	0.094	0.034	0.006
ρ18	0.054	0.039	0.163

Coefficient	Estimate	Std. Err.	p-value	Coefficient	Estimate	Std. Err.	p-value	Coefficient	Estimate	Std. Err.	p-value
ρ19	-0.078	0.037	0.036	ρ511	-0.155	0.042	0.000	α16	0.011	0.008	0.174
ρ110	0.030	0.032	0.349	ρ67	-0.004	0.041	0.927	α17	0.001	0.001	0.496
ρ111	-0.029	0.045	0.514	ρ68	-0.294	0.033	0.000	α18	0.002	0.001	0.005
ρ23	-0.186	0.034	0.000	ρ69	-0.181	0.054	0.001	α19	-0.001	0.001	0.647
ρ24	-0.363	0.058	0.000	ρ610	-0.089	0.028	0.001	α110	0.004	0.001	0.004
ρ25	-0.018	0.048	0.710	ρ611	-0.324	0.074	0.000	α111	-0.010	0.003	0.002
ρ26	-0.292	0.063	0.000	ρ78	-0.124	0.033	0.000	α21	-0.063	0.010	0.000
ρ27	-0.069	0.043	0.111	ρ79	0.195	0.038	0.000	α22	-0.041	0.026	0.117
ρ28	-0.117	0.043	0.007	ρ710	0.050	0.033	0.135	α23	0.072	0.015	0.000
ρ29	-0.084	0.054	0.120	ρ711	-0.236	0.034	0.000	α24	-0.009	0.014	0.510
ρ210	-0.067	0.043	0.116	ρ89	0.037	0.028	0.189	α25	-0.023	0.024	0.331
ρ211	-0.256	0.036	0.000	ρ810	-0.098	0.041	0.018	α26	-0.074	0.050	0.135
ρ34	0.066	0.028	0.017	ρ811	0.005	0.036	0.885	α27	-0.040	0.007	0.000
ρ35	-0.154	0.034	0.000	ρ910	0.021	0.039	0.581	α28	0.033	0.008	0.000
ρ36	-0.130	0.039	0.001	ρ1011	-0.130	0.043	0.002	α28	0.002	0.008	0.795
ρ37	0.016	0.034	0.651	ρ1112	-0.138	0.035	0.000	α210	-0.037	0.012	0.002
ρ38	-0.075	0.032	0.018	α01	0.105	0.007	0.000	α211	0.132	0.029	0.000
ρ39	-0.001	0.033	0.973	α02	0.183	0.018	0.000	α31	0.078	0.008	0.000
ρ310	-0.129	0.034	0.000	α03	0.127	0.006	0.000	α32	-0.296	0.029	0.000
ρ311	-0.112	0.038	0.003	α04	0.026	0.007	0.000	α33	-0.062	0.011	0.000
ρ45	-0.180	0.043	0.000	α05	0.037	0.018	0.039	α34	0.049	0.011	0.000
ρ46	0.062	0.039	0.109	α06	-0.128	0.031	0.000	α35	-0.049	0.026	0.063
ρ47	-0.017	0.032	0.599	α07	0.046	0.005	0.000	α36	0.043	0.026	0.108
ρ48	0.003	0.036	0.941	α08	0.017	0.006	0.007	α37	-0.056	0.008	0.000
ρ49	0.248	0.039	0.000	α08	0.014	0.005	0.002	α38	0.086	0.011	0.000
ρ410	-0.008	0.036	0.815	α010	0.040	0.006	0.000	α38	-0.024	0.011	0.028
ρ411	-0.239	0.044	0.000	α011	0.014	0.015	0.342	α310	-0.036	0.010	0.000
ρ56	-0.250	0.041	0.000	α11	0.005	0.001	0.000	α311	0.010	0.021	0.624
ρ57	0.129	0.031	0.000	α12	-0.003	0.004	0.577				
ρ58	-0.010	0.033	0.767	α13	0.005	0.002	0.003				
ρ59	0.203	0.035	0.000	α14	0.004	0.001	0.000				
ρ510	-0.076	0.040	0.057	α15	-0.005	0.003	0.093				

Estimation of Tobit type censored demand systems: A comparison of estimators *

Mikkel Barslund
Department of Economics, University of Copenhagen
Development Economics Research Group (DERG)

October 2006
Revised: January 2007

Abstract

Recently a number of authors have suggested to estimate censored demand systems as a system of Tobit multivariate equations employing a Quasi Maximum Likelihood (QML) estimator based on bivariate Tobit models. In this paper I study the efficiency of this QML estimator relative to the asymptotically more efficient Simulated ML (SML) estimator in the context of a censored Almost Ideal demand system. Further, a simpler QML estimator based on the sum of univariate Tobit models is introduced. A Monte Carlo simulation comparing the three estimators is performed on three different sample sizes. The QML estimators perform well in the presence of moderate sized error correlation coefficients often found in empirical studies. With absolute larger correlation coefficients, the SML estimator is found to be superior. The paper lends support to the general use of the QML estimators and points towards the use of simple etimators for more general censored systems of equations.

Keywords: Censored demand system, Monte Carlo, Quasi maximum likelihood, Simulated maximum likelihood.

JEL Classifications: D12, C15, C34

*Address for correspondence: Mikkel Barslund, Department of Economics, Studiestraede 6, DK-1455 Copenhagen K, Denmark. Email: Mikkel.Barslund@econ.ku.dk. I would like to thank Katleen Van den Broek, Carol Newman and Heino Bohn Nielsen for useful comments and suggestions.

1 Introduction

Analysis of individuals and households consumption patterns and their response to relative price changes has a long tradition in economics and goes back at least to Engels seminal work on expenditure shares. Systems of flexible functional forms such as the translog and almost ideal demand systems (Jorgensen et al. 1979, Deaton and Muellbauer 1980) and the advance of fast computers have made estimation of price response coefficients in large demand systems with many goods based on household survey data feasible. Hence, a large literature has grown. However, until recently the problem of censoring of the expenditure shares (i.e. the minimum consumption share is zero) was largely ignored or only addressed in systems with a small number of goods (see Wales & Woodland 1983, Lee & Pitt 1986).

To account for censoring a model which allows for a positive probability of observing zero consumption must be estimated. Thus, whether implicit or explicit, the model should accommodate a market participation decision and a consumption decision. Further, the estimation procedure must be capable of accommodating cross-equation restrictions, making joint estimation of all equations necessary. If errors are normal and assumed to covary between the decisions to consume each good, then - with multiple goods not consumed for some households - the contribution to the likelihood function will require evaluation of multiple integrals over a multivariate normal density function. As an example, consider the case of a five good demand system where a non-neglible number of households only consume two of the five goods, thus equations for the three non-consumed goods are censored. For these households part of the likelihood contribution will be the probability that the three error terms fall within a range consistent with observed censoring of these three goods. Difficulties associated with evaluating multiple integrals over the multivariate normal density function explain why accounting for censoring in applications of large demand system is rare.

One way to account for censoring which has been used in the literature is to model the consumption shares as a multivariate Tobit model (see Yen, Lin and Smallwood 2003), such that implicitly the participation decision and the consumption decision are determined by the same process. In the context of demand system estimation two maximum likelihood based estimators have recently been used to estimate multivariate Tobit systems. Harris and Shonkwiler (1997) proposed a Quasi Maximum Likelihood (QML) estimator based on linking bivariate Tobit models to avoid evaluating high dimensional integrals. More recently Yen, Lin and Smallwood (2003) have used a Simulated Maximum Likelihood (SML) estimator of a similar system.[1] While both estimators are consistent, the SML estimator is asymptotically more efficient, since it uses more sample information than the QML estimator. However, the relative performance of the estimators in applications with empirically relevant sample sizes is unknown.

The contribution of this paper is twofold. First - inspired by the idea of linking bivariate Tobit models - I introduce a simpler QML estimator based on the maximization of the sum of univariate Tobit models over all equations. Although the proposed estimator does not identify the error correlation across equations this is of secondary importance in a demand system context since error correlations are not used to calculate elasticities or other quantities of interest. Second, I compare the three estimators using Monte Carlo simulations in a setup with four simultaneous equations subject to a large degree of censoring. Their performance is assessed in three different sample sizes with respectively 200, 1,000 and 3,000 observations. The sample sizes are chosen to resemble a 'small' sample of households (200 observations), a larger sample (1,000 observations) and a typical

[1] Shonkwiler & Yen (1999) propose a two step estimator where the participation decision is modelled as a univarite probit on each equation (or alternatively, a multivariate probit over all equations). In the second step, the equations determining the expenditure shares are augmented to take account of the censoring and errors are assumed multivariate normal. The estimator is consistent but less efficient relative to the SML estimator due to the two-step nature. It is not considered in the present work, since it is not suitable for estimation of Tobit type models.

(sub)-sample from the World Banks LSMS surveys (3,000 observations).

There are a number of reasons contributing to the relevance of this exercise. First, it is not evident which estimator is preferable for relatively small sample sizes. Second, even if the SML estimator is superior, the cost of implementation and the computational burden associated with simulating the likelihood function might warrant the use of a sufficiently good second best estimator. Third, the SML estimator has difficulties converging from arbitrary starting values and computation time is reduced substantially by using good starting values possibly obtained from less efficient estimators. Further, the type of QML estimators used in this paper can be applied to more general systems of censored systems, i.e. the system suggested by Yen and Lin (2006).

There exist few application specific comparisons of the SML estimator and the bivariate Tobit QML estimator considered here. Yen, Lin and Smallwood (2003) estimate a large demand system with both the SML and the bivariate Tobit QML and conclude that the QML and SML estimator deliver very similar results. In a similar application Yen and Lin (2002) find QML and SML estimates to be close and similar. Clearly, since the true data generating mechanism and parameters are unknown these studies cannot shed light on the relative performance of the estimators in question.

In the following section the model is outlined together with the three estimators. Section 3 describes the Monte Carlo setup, while section 4 presents results.[2] Some brief concluding remarks are offered at the end.

2 Estimation of a multivariate system of Tobit equations

The point of departure is a multivariate generalization of the Tobit model. Denote the dependent variable by y_i, the matrix of explanatory variables by X and the full set of

[2] While the simulations have been done in the context of estimating an almost ideal demand system with five goods (the last good being determined residually as suggested by Pudney (1989), i.e. four goods/equations estimated), this is not emphasized in the discussion of the results.

parameters to be estimated by θ, then the system of equations ($i = 1, .., M$) can be written (suppressing observation indices)

$$y_i = \max\left(f_i\left(X;\theta\right) + \varepsilon_i, 0\right), \quad i = 1, .., M \tag{1}$$

where ε_i is an equation specific error term. Define the vector of errors $\varepsilon = [\varepsilon_1, .., \varepsilon_M]$ and allow parametric estimation by assuming multivariate normal errors with zero mean and covariance Σ. In the context of demand system estimation, y_i is the expenditure share on good i and the functions f_i are of some flexible form. Note that in applications where there is no need for cross equation restrictions (1) can be estimated consistently using a univariate Tobit model equation by equation. However, even in this case efficiency is gained by estimating all equations jointly as a system.

Simulated Maximum Likelihood estimation

To construct the likelihood function for the system given by (1), let a censoring regime z_c be a $1 \times M-$vector with entries equal to zero for the censored equations and one for the non-censored equations. Each observation belongs to a particular censoring regime. Thus, an observation with the first k equations non-censored and the remaining censored would have ones in the first k entries and zeros for the rest. Call this regime z_c and note that all censoring regimes can be written like this with k equal to the number of non-censored equations and a suitable reorganization of the equations. That is, no generality is lost. To develop the likelihood function for the observations belonging to the censoring regime, z_c, partition the error vector and the covariance matrix such that

$$\varepsilon \equiv [\varepsilon_1, \varepsilon_2] \equiv [\varepsilon_1, \varepsilon_2, .., \varepsilon_k : \varepsilon_{k+1}, \varepsilon_{k+2}, .., \varepsilon_M]$$
$$\Sigma \equiv \begin{bmatrix} \Sigma_{11} & \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where Σ_{11} is a $k \times k$ matrix, Σ_{21} is a $(M-k) \times k$ matrix and Σ_{22} is a $(M-k) \times (M-k)$

matrix. Let $g(\varepsilon_1)$ be the joint marginal probability density function (pdf) for the first k errors. The pdf function for the M errors can be written in terms of $g(\varepsilon_1)$ and the joint marginal pdf of the remaining $(M-k)$ error terms conditional on observing ε_1, $h(\varepsilon_2 \mid \varepsilon_1)$.Thus, the joint marginal pdf, $f(\varepsilon_1, \varepsilon_2)$, can be written as

$$f(\varepsilon_1, \varepsilon_2) \equiv g(\varepsilon_1) \cdot h(\varepsilon_2 \mid \varepsilon_1)$$

It can be shown that $h(\varepsilon_2 \mid \varepsilon_1)$ is distributed multivariate normal with mean and covariance matrix given by (Greene 2000)

$$\mu_{2.1} = \Sigma_{21}\Sigma_{11}^{-1}\varepsilon_1$$

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}^{'}$$

The contribution to the likelihood function for an observation belonging to censoring regime z_c is then given by

$$L^{Z_c} = g(\mathbf{e}_1) \int\limits_{-\infty}^{-f_{k+1}(X;\theta)} \int\limits_{-\infty}^{-f_{k+2}(X;\theta)} .. \int\limits_{-\infty}^{-f_M(X;\theta)} \phi_{(M-k)}\left(u_{k+1}, u_{k+2}, .., u_M\right) \partial u_M .. \partial u_{k+2} \partial u_{k+1} \tag{2}$$

where $g(\mathbf{e}_1)$ is the k-variate normal density with zero mean and covariance matrix Σ_{11} evaluated at $\mathbf{e}_1 = \left[y_1 - f_1\left(X;\theta\right), y_2 - f_2\left(X;\theta\right), .., y_k - f_k\left(X;\theta\right)\right]$. The integration is with respect to the $M-k$-variate normal density with mean and covariance given above. To write the likelihood function for the sample define the indicator function $I_h^{Z_c}$ being one if observation h is in censoring regime z_c and zero otherwise. Since each observation belongs to only one censoring regime the sample likelihood can be written as

$$L = \prod_h \prod_{z_c} \left[L_h^{z_c}\right]^{I_h^{z_c}}$$

The set of censoring regimes includes the two special cases where respectively none and all equations are censored. The contributions to the likelihood function are $L^{Z_c} = g(\mathbf{e}_1)$ and

$L^{Z_c} = \int\limits_{-\infty}^{-f_1(X;\theta)} \int\limits_{-\infty}^{-f_2(X;\theta)} .. \int\limits_{-\infty}^{-f_M(X;\theta)} \phi_{(M)}\left(u_1, u_2, .., u_M\right) \partial u_M .. \partial u_2 \partial u_1$, both M-variate normal density functions having zero mean and covariance Σ.

If for just one observation the number of censored equations exceeds two, a simulation method has to be relied upon to evaluate the integral in (2). I rely on the GHK (Geweke, Hajivassiliou and Keane) simulator to evaluate the integrals[3].

Quasi Maximum Likelihood Estimation

Although implementation of the SML estimator is feasible in most statistical packages (such as Stata and Gauss) it is likely to be computationally intensive. In addition without good starting values obtaining convergence can be difficult. Thus, it is of interest to explore simpler estimators which allow for cross-equation restrictions and do not require simulation techniques. Analogous to the literature on multivariate probit models (Avery et al. 1983) a simple alternative is a QML estimator which maximizes the sum of individual equation Tobit models (QML_{T1}). Formally, the QML_{T1} estimator maximizes

$$\ln QML_{T1} = \sum_h \sum_i \ln L_{T1,i} \tag{3}$$

where the subscript $T1$ indicates that the QML is with respect to the univariate Tobit model. $L_{T1,i}$ is the Tobit likelihood function for the i'th equation and as before h indexes observations. The estimator is based on maximizing the sum of marginal densities of the system in (1) and is therefore consistent (Cameron and Trivedi, 2005). However, because the likelihood function in (3) is mis-specified relative to the true likelihood function for the model in (1), Whites robust standard errors should be used for statistical inference (White,

[3]The methodology behind the GHK simulator is explained elsewhere and is beyond the scope of the present paper (see Börsch-Supan and Hajivassiliou, 1993, and Cappellari and Jenkins, 2006). In practice the GHK simulator has been shown to work well (Greene 2000).

1982). The estimator incorporates all equations simultaneously in the procedure so cross equation restrictions can be imposed. The entries in the covariance matrix outside the diagonal in the system in (1) are not identified. Hence, the vector of estimated coefficients has fewer elements than the vector of coefficients from the SML estimator. For demand system applications where cross equation correlations are not of particular value this is of minor importance. On the other hand, if the purpose of the QML estimation is to get starting values for the SML estimator cross equation correlations are valuable in their own right.

An extension to the approach in (3) which yields estimates of cross equation correlation coefficients is to estimate a sequence of pair-wise bivariate Tobit models (QML_{T2}). The QML_{T2} estimator maximizes

$$\ln QML_{T2} = \sum_{h} \sum_{i}^{M-1} \sum_{j=i+1}^{M} \ln L_{T2;(i,j)}$$

$T2$ indicates that the sequence of likelihood functions are bivariate Tobits. The coefficient of correlation between the error terms in the i'th and j'th equations is identified from the contribution of $L_{T2;(i,j)}$ to the sample quasi likelihood. Thus, this approach yields as many estimated coefficients as the SML estimator. This last method has recently been used in a number of studies (see Yen, Lin and Smallwood 2003, Barslund 2006, Lin and Yen 2002, Harris and Shonkwiler 1997).

3 Monte Carlo simulations

The relative performance of each estimator is explored along two dimensions. First, a system of four equations is estimated on three different sample sizes to investigate how closely their performance is related to sample size. The sample sizes of 1,000 and 3,000 observations are chosen to resemble empirically relevant samples from typical cross sectional data sets. The third sample with 200 observations is employed to look at performance in a 'small' sample. Second, the effect of the absolute size of the error correlation coefficients is examined. In particular, since the QML_{T1} estimator ignores cross equation error correlations its performance should deteriorate as the absolute size of the correlation coefficients increases. The QML_{T2} estimator identifies the correlation coefficient via the bivariate Tobit formulation, but unlike the SML estimator it does not take into account the complete correlation structure when estimating the pair-wise correlation coefficients. Overall, the SML estimator should improve relative to the other two estimators when the correlation between equations increases. I compare the estimators for each sample size and for two error correlation structures; namely an empirically relevant correlation matrix ('base' correlations) and a matrix where the base correlations are doubled ('large' correlations). For comparison, and using the 'base' correlation matrix, the system of latent shares is estimated ignoring the issue of censoring and the errors are assumed multivariate normally distributed. Each scenario consists of 500 simulations.

Monte Carlo setup

The system of equations is based on a censored almost ideal demand system. In the context of an empirical application this corresponds to a five good system where the last good is residually determined as suggested by Pudney (1989). Although adding up (expenditure shares sum to one) is accommodated in this way, parameter restrictions designed to facilitate adding up in the latent system of expenditure shares are still imposed. In addition, in order to see how the estimators perform in the presence of cross equation restrictions, slutsky symmetry is imposed on latent shares even if the theoretical justification for this is blurred in censored systems.[4] The latent almost ideal demand system has the form

[4] In any case, imposing slutsky symmetry in censored demand systems is standard practice in empirical applications.

(observation indices suppressed)

$$w_i^* = \alpha_i + \sum_{j=1}^{M+1} \gamma_{ij} \log p_j + \beta_i \log\left(x/a(p)\right) \tag{4}$$
$$\text{with } \log a(p) = \alpha_0 + \sum_{j=1}^{M+1} \alpha_j \log p_j + 1/2 \sum_{i=1}^{M+1} \sum_{j=1}^{M+1} \gamma_{ij} \log p_i \log p_j$$

Where w_i^* is the latent expenditure share on commodity i. α_i, β_i and γ_{ij} are parameters to be estimated. Exogenous variables are prices, p_i and income/expenditure x. As is often done in empirical applications, the unidentified parameter α_0 is set equal to zero (Moschini, 1998). The indices i, j denote commodities, thus $i, j \in \{1, .., M+1\}$. Adding-up, slutsky symmetry and homogeneity of the latent shares are ensured by the parameter restrictions: $\sum_{i=1}^{M+1} \alpha_i = 1$, $\gamma_{ij} = \gamma_{ji}, \forall i, j$ and $\sum_{i=1}^{M+1} \beta_i = 0$.Denoting the full parameter vector by θ the observed shares are given by (equivalent to the system in (1))

$$w_i = \max\left(w_i^*\left(\theta\right) + \varepsilon_i, 0\right), \quad i = 1, .., M$$

For each scenario the simulations are done in the following steps:

1) Exogenous variables (logarithmic prices and incomes) are drawn from a standard normal distribution.

2) Errors are drawn from the specified multivariate normal distribution (cf. below).

3) Latent shares are calculated, errors added, and the observed share is determined from the censoring rule.

4) Each estimator is estimated using the observed shares and exogenous variables.

5) Estimates are saved.

Step 2 to 5 are carried out 500 times for each scenario with fixed exogenous variables.

Parameters are chosen such that they are within a range often found in empirical applications.

	Alpha	Beta	Gamma:	Eq. 1	Eq. 2	Eq. 3	Eq. 4	Eq. 5
Equation 1	0.3	-0.025		-0.06	-0.03	0.05	0.20	0.02
Equation 2	0.25	0.03		-0.03	-0.01	0.02	0.01	0.01
Equation 3	0.05	-0.01		0.05	0.02	-0.03	-0.02	-0.02
Equation 4	0.1	0.02		0.20	0.01	-0.02	0.01	-0.02
Equation 5	0.3	-0.015		0.02	0.01	-0.02	-0.02	0.01

The values ensure both adding up and slutsky symmetry of the latent shares. The errors are drawn from a multivariate normal distribution with the base error correlation structure given by:

	Standard deviation (σ).	Probability censored (%)	Correlation matrix (base corr. coef.): Eq. 1	Eq. 2	Eq. 3	Eq. 4
Equation 1	0.6	30.9	1.00	-0.20	-0.15	-0.08
Equation 2	0.5	30.9	-0.20	1.00	-0.15	-0.07
Equation 3	0.4	45.0	-0.15	-0.15	1.00	-0.10
Equation 4	0.3	36.9	-0.08	-0.07	-0.10	1.00

The probability of an observation being censored is calculated using that for any given observation the expected latent expenditure share w_i^* is equal to α_i since expected logarithmic prices and income are zero (drawn from a standard normal distribution). The base correlation matrix is chosen to resemble the range of values found in empirical studies. The average absolute value over the error correlation coefficients is 0.125 with a maximum absolute value of 0.20. This compares well with the average of 0.083 over absolute correlation coefficients found in Yen, Lin and Smallwood (2003) with only one coefficient out of 66 being significantly larger than 0.20. Yen, Fang and Su (2004) report slightly larger coefficients. The absolute average is 0.118 and 6 out of 45 coefficients are significantly larger than 0.20 with a maximum of 0.288. Similarly, Barslund (2006) finds an absolute average of 0.116 with 5 out of 55 correlation coefficients being significantly larger than 0.20. The maximum value reported is 0.327. Lastly, Yen and Lin (2002) estimate a three equation system with the largest correlation coefficient not significantly larger than 0.20

and with an average value of 0.136 in absolute terms. Although the absolute size of the correlation coefficients is application specific, the scenarios with 'large' correlations should provide an upper bound for differences in the estimators likely to be found in empirical applications.

A final issue relates to the evaluation of the SML log-likelihood using the GHK simulator. The GHK simulator relies on a specific number, R, of random draws from the unit interval. Because the accuracy of the GHK simulator relies on the size of R, formally, the efficiency of the SML estimator hinges on $\sqrt{N}/R \longrightarrow 0$, where N is the number of observations (see Train 2003). The pitfall to avoid in relation to the Monte Carlo simulation is that the relative performance of the SML versus the QML_{T1} and QML_{T2} is not confounded with poor accuracy of the SML estimator due to an inadequate number of draws when using the GHK simulator. The random draws were generated by Statas `mdraws` command (Cappellari and Jenkins, 2006). In practice, the number of draws were determined following a suggestion by Haan and Uhlendorff (2006). They propose to start by maximizing a simulated log-likelihood function using R equal to $N^{0.55}$ random draws and then increase the number of draws until the maximized log-likelihood function stabilizes on a value. For all three sample sizes Halton sequences with $R = 84$ and antithetic draws were used (Cappellari and Jenkins, 2006). For the sample of 3,000 observations the change in the log-likelihood value at $R = 84$ was below 1/100 of a percentage point. The change in parameter values was on average less than 1/25 of the difference between the SML and the QML_{T2} estimator.[5]

[5] All estimations were done in Stata with 'seeding' of the random generator used for drawing errors so as to facilitate replicability. Files are available from the author.

4 Results

To manage the amount of output the discussion of the results will concentrate on differences in the mean squared error (MSE) between the estimators. The performance of the QML estimators, QML_{T1} and QML_{T2}, is measured relative to the asymptotically more efficient SML estimator. Table 1 shows the results for the base correlation specification with a sample of 200 observations.

Columns numbered 2 through 10 show the percentage deviation of the mean over the 500 simulations from the true value and the MSE for each of the estimated parameters for respectively, the SML, QML_{T1}, QML_{T2}, and the non-censored estimator. The deviation from the mean is reported in order to gorge the biasness in finite samples. As expected - given the degree of censoring - the non-censored estimator shows a large bias for all coefficients (column 8). Turning to the three estimators of primary interest, the most interesting thing coming out of Table 1 is how similar the results are. Looking across the rows it is clear that the differences between the estimators for both measures are small. When one estimator performs particularly well with respect to a point estimate of a coefficient the other two also do well. And similar when coefficients are less precisely estimated. For an illustration look at the estimated standard deviations for the error term of equation 1 (σ_1) and 4 (σ_4), respectively. In terms of their MSE, σ_4 performs well over all three estimators whereas the opposite is true for σ_1. Column 10, 11 and 12 summarize the differences between the coefficient MSEs over the three estimators. Column 10 shows a comparison of the SML and QML_{T2} estimator, where a plus indicates that the SML has the lowest MSE. Similar for column 11 where the SML is compared to the QML_{T1} estimator. Lastly, the QML_{T2} and QML_{T1} estimators are compared in column 12. Thus, for all three columns a plus signifies that the estimator using the most sample information performed better. Reflecting the resemblance of column 2 to 7 none of the estimators

Table 1: Simulation results. Base correlations, 200 observations (500 simulations).

Parameter	True value	SML Deviation mean (%)	SML MSE (x10000)	QML-T2 Deviation mean (%)	QML-T2 MSE (x10000)	QML-T1 Deviation mean (%)	QML-T1 MSE (x10000)	Non-censored Deviation mean (%)	Non-censored MSE (x10000)	Comparison $(5)-(3)$	Comparison $(7)-(3)$	Comparison $(7)-(5)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\gamma 11$	-0.06	-2.77	15.70	-2.75	15.87	-2.75	16.01	-31.86	11.20	$+$	$+$	$+^{**}$
$\gamma 12$	-0.03	5.72	7.42	5.58	7.43	5.47	7.46	-35.01	4.71	$+$	$+$	$+^{*}$
$\gamma 13$	0.05	-0.22	5.57	-0.09	5.61	-0.02	5.64	-45.79	7.24	$+$	$+$	$+^{**}$
$\gamma 14$	0.02	2.43	3.38	2.84	3.32	3.08	3.31	-26.90	1.78	$-^{**}$	$-^{*}$	$-$
$\gamma 22$	-0.01	-10.19	10.80	-9.07	10.89	-8.63	10.98	-78.08	6.48	$+$	$+$	$+^{*}$
$\gamma 23$	0.02	4.82	4.74	4.38	4.77	4.24	4.81	-34.82	2.24	$+$	$+$	$+^{**}$
$\gamma 24$	0.01	-9.88	2.91	-9.61	2.90	-9.53	2.91	-22.01	1.40	$-$	$+$	$+$
$\gamma 33$	-0.03	3.59	7.44	3.09	7.43	2.83	7.48	-37.77	3.80	$-$	$+$	$+^{*}$
$\gamma 34$	-0.02	0.60	2.93	0.53	2.89	0.61	2.88	-43.33	1.89	$-^{*}$	$-^{*}$	$-$
$\gamma 44$	0.01	1.35	3.26	1.40	3.24	1.50	3.24	-24.16	1.45	$-$	$-$	$+$
$\sigma 1$	0.6	-0.89	16.75	-0.93	16.65	-0.94	16.62	-26.34	255.86	$-$	$-$	$-$
$\sigma 2$	0.5	-1.32	11.10	-1.34	11.10	-1.34	11.11	-26.30	177.26	$-$	$+$	$+$
$\sigma 3$	0.4	-1.64	8.97	-1.65	9.00	-1.65	9.02	-38.66	241.84	$+$	$+$	$+$
$\sigma 4$	0.3	-1.45	4.11	-1.45	4.08	-1.44	4.07	-31.59	91.33	$-^{*}$	$-^{*}$	$-^{*}$
$\rho 12$	-0.2	-0.93	56.10	-1.50	55.90	.	.	-17.03	54.22	$-$	.	.
$\rho 13$	-0.15	-2.28	66.39	-2.95	65.21	.	.	-25.61	58.59	$-^{**}$	.	.
$\rho 14$	-0.08	2.92	69.92	2.37	68.95	.	.	-15.14	53.43	$-^{**}$	.	.
$\rho 23$	-0.15	0.58	64.63	-0.23	63.93	.	.	-21.61	50.37	$-$	.	.
$\rho 24$	-0.07	-1.99	58.73	-2.26	58.37	.	.	-17.42	46.81	$-$	.	.
$\rho 34$	-0.1	-3.81	69.18	-4.45	68.52	.	.	-25.92	56.72	$-$	.	.
$\alpha 01$	0.3	-0.82	21.60	-0.80	21.54	-0.79	21.51	40.07	154.52	$-$	$-$	$-$
$\alpha 02$	0.25	0.81	14.47	0.80	14.46	0.80	14.47	40.75	110.84	$-$	$+$	$+$
$\alpha 03$	0.05	-1.64	10.67	-1.59	10.73	-1.58	10.76	274.73	191.48	$+$	$+$	$+$
$\alpha 04$	0.1	1.45	5.89	1.46	5.87	1.46	5.87	78.76	64.36	$-$	$-$	$-$
$\beta 1$	-0.025	-9.35	17.37	-9.02	17.52	-8.87	17.57	-37.63	9.20	$+^{**}$	$+^{**}$	$+$
$\beta 2$	0.03	0.85	13.00	0.62	13.01	0.56	13.01	-31.85	7.59	$+$	$+$	$+$
$\beta 3$	-0.01	9.63	8.80	10.50	8.80	10.86	8.81	-47.66	2.98	$+$	$+$	$+$
$\beta 4$	0.02	6.68	4.41	6.47	4.41	6.40	4.42	-31.14	2.24	$+$	$+$	$+$

Notes: Parameters refer to the main text (eq. 4). σ and ρ are respectively the standard error and correlation of the error term. Deviation mean (column 2, 4, 6 and 8) shows the percentage deviation of the mean over the 500 simualtions from the true value. Column 10, 11 and 12 show the sign of difference between column 5 and 3, 7 and 3, and 7 and 5, respectively. *, ** and *** indicate that the compared columns are significantly different at 10, 5 and 1 percent based on two sided t-tests.

perform better than the two others for all coefficients. However, the QML_{T2} seems to have on average slightly lower MSEs than both the SML (minus in column 10) and the QML_{T1} (plus in column 12).

It is of interest to test if the small differences in performance between the estimators are statistically significant. For this purpose I perform two-tailed t-tests of equality of MSEs based on the sample of 500 replications. Significance levels are indicated in the three last columns by one, two or three asterisks equivalent to significance at 10, 5 and 1 percent, respectively. For only one coefficient (β_1) does the SML estimator perform significantly better than the two others, while the QML_{T2} does significantly better than the SML for five coefficients and better than the QML_{T1} for seven coefficients. In sum, for small samples with error correlations of empirical relevant size both the QML_{T1} and QML_{T2} perform very well.

Table 2 is similar to Table 1, but the sample size is increased to 1,000 observations. First, note that for the three estimators of primary interest the deviations of the mean for most estimated coefficients are smaller than in Table 1. This is to be expected from consistent estimators. Contrast that with the biased non-censored estimator, where deviations from the mean are more or less unchanged between Table 1 and 2. Also the MSEs are reduced substantially. Regarding the comparison in the last three columns, the SML estimator has lower MSEs for a majority of coefficients than both the QML_{T1} and QML_{T2} estimators. However, this better performance is not statistically significant (except for one coefficient in the comparison between the SML and QML_{T1} estimators), again reflecting that the coefficient estimates coming from the three estimators are very close for each simulation. The QML_{T2} estimator has lower MSEs for all but two parameters (five are significantly lower) compared with the QML_{T1} estimator. In Table 3 the number of observations is further increased to 3,000 while keeping the same base error correlation structure. Except for the non-censored estimator the effect on the deviation of the mean

Table 2: Simulation results. Base correlations, 1000 observations (500 simulations).

Parameter	True value	SML Deviation mean (%)	SML MSE (x10000)	QML-T2 Deviation mean (%)	QML-T2 MSE (x10000)	QML-T1 Deviation mean (%)	QML-T1 MSE (x10000)	Non-censored Deviation mean (%)	Non-censored MSE (x10000)	Comparison (5) − (3)	Comparison (7) − (3)	Comparison (7) − (5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\gamma 11$	-0.06	-2.480	2.737	-2.483	2.718	-2.467	2.715	-29.71	4.53	$-$	$-$	$-$
$\gamma 12$	-0.03	1.124	1.284	1.095	1.293	1.093	1.300	-37.29	1.89	$+$	$+^{*}$	$+^{***}$
$\gamma 13$	0.05	-1.251	0.979	-1.251	0.982	-1.252	0.985	-45.74	5.61	$+$	$+$	$+$
$\gamma 14$	0.02	0.703	0.591	0.801	0.591	0.843	0.593	-30.16	0.64	$+$	$+$	$+$
$\gamma 22$	-0.01	-8.771	2.171	-8.451	2.186	-8.383	2.205	-72.66	1.66	$+$	$+$	$+^{***}$
$\gamma 23$	0.02	-1.747	0.898	-1.782	0.906	-1.793	0.912	-39.12	0.98	$+$	$+$	$+^{**}$
$\gamma 24$	0.01	-0.674	0.573	-0.875	0.572	-0.973	0.574	-18.70	0.29	$-$	$+$	$+$
$\gamma 33$	-0.03	-4.146	1.392	-4.184	1.395	-4.181	1.401	-42.23	2.08	$+$	$+$	$+$
$\gamma 34$	-0.02	2.443	0.506	2.472	0.508	2.477	0.510	-41.47	0.89	$+$	$+$	$+^{*}$
$\gamma 44$	0.01	3.529	0.779	3.238	0.779	3.157	0.780	-25.82	0.39	$+$	$+$	$+$
$\sigma 1$	0.6	-0.164	3.016	-0.163	3.022	-0.161	3.027	-25.95	243.61	$+$	$+$	$+$
$\sigma 2$	0.5	-0.190	2.182	-0.202	2.187	-0.206	2.191	-25.80	167.25	$+$	$+$	$+$
$\sigma 3$	0.4	-0.363	1.660	-0.359	1.667	-0.358	1.673	-37.16	221.60	$+$	$+$	$+^{**}$
$\sigma 4$	0.3	-0.164	0.825	-0.159	0.822	-0.156	0.823	-30.78	85.55	$-$	$-$	$+$
$\rho 12$	-0.2	0.532	11.451	0.421	11.494	.	.	-15.55	18.28	$+$	.	.
$\rho 13$	-0.15	-0.685	11.999	-0.770	12.013	.	.	-22.55	18.82	$+$	.	.
$\rho 14$	-0.08	0.701	14.008	0.592	13.913	.	.	-17.17	12.03	$-$	.	.
$\rho 23$	-0.15	0.866	12.489	0.818	12.442	.	.	-20.93	17.86	$-$	.	.
$\rho 24$	-0.07	-0.573	11.692	-0.757	11.769	.	.	-17.59	10.39	$+$	.	.
$\rho 34$	-0.1	-2.027	13.230	-2.049	13.162	.	.	-22.59	14.37	$-$	.	.
$\alpha 01$	0.3	-0.369	4.104	-0.376	4.104	-0.380	4.109	39.91	145.28	$-$	$+$	$+$
$\alpha 02$	0.25	0.212	2.894	0.228	2.874	0.233	2.870	40.13	101.93	$-^{***}$	$-^{**}$	$-^{*}$
$\alpha 03$	0.05	3.536	2.152	3.511	2.145	3.508	2.146	278.07	193.97	$-$	$-$	$+$
$\alpha 04$	0.1	-0.610	1.111	-0.615	1.110	-0.617	1.110	77.51	60.51	$-$	$-$	$+$
$\beta 1$	-0.025	-3.961	3.113	-3.985	3.115	-4.007	3.117	-36.44	2.37	$+$	$+$	$+$
$\beta 2$	0.03	0.829	1.939	0.998	1.945	1.066	1.947	-33.14	2.03	$+$	$+$	$+$
$\beta 3$	-0.01	8.575	1.536	8.675	1.538	8.681	1.539	-49.14	0.73	$+$	$+$	$+$
$\beta 4$	0.02	0.411	0.994	0.383	0.994	0.383	0.995	-37.86	0.97	$+$	$+$	$+$

Notes: Parameters refer to the main text (eq. 4). σ and ρ are respectively the standard error and correlation of the error term. Deviation mean (column 2, 4, 6 and 8) shows the percentage deviation of the mean over the 500 simualtions from the true value. Column 10, 11 and 12 show the sign of difference between column 5 and 3, 7 and 3, and 7 and 5, respectively. *, ** and *** indicate that the compared columns are significantly different at 10, 5 and 1 percent based on two sided t-tests.

Table 3: Simulation results. Base correlations, 3000 observations (500 simulations).

Parameter	True value	SML Deviation mean (%)	SML MSE (x10000)	QML-T2 Deviation mean (%)	QML-T2 MSE (x10000)	QML-T1 Deviation mean (%)	QML-T1 MSE (x10000)	Non-censored Deviation mean (%)	Non-censored MSE (x10000)	Comparison (5) − (3)	Comparison (7) − (3)	Comparison (7) − (5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\gamma11$	-0.06	-0.775	0.916	-0.838	0.924	-0.865	0.930	-28.76	3.45	$+$	$+^{*}$	$+^{**}$
$\gamma12$	-0.03	-1.078	0.428	-1.064	0.426	-1.059	0.425	-38.01	1.52	$-$	$-$	$-$
$\gamma13$	0.05	-0.163	0.292	-0.154	0.294	-0.150	0.295	-45.59	5.31	$+$	$+$	$+^{**}$
$\gamma14$	0.02	-0.138	0.199	-0.103	0.201	-0.096	0.201	-30.31	0.45	$+$	$+$	$+$
$\gamma22$	-0.01	-2.218	0.688	-1.680	0.692	-1.471	0.695	-71.25	0.88	$+$	$+$	$+^{**}$
$\gamma23$	0.02	-2.307	0.324	-2.391	0.324	-2.410	0.325	-39.38	0.75	$-$	$+$	$+$
$\gamma24$	0.01	-0.099	0.201	-0.278	0.202	-0.333	0.202	-17.71	0.12	$+$	$+$	$+$
$\gamma33$	-0.03	-1.212	0.493	-1.286	0.490	-1.297	0.490	-41.52	1.72	$-$	$-$	$+$
$\gamma34$	-0.02	-0.810	0.187	-0.833	0.188	-0.838	0.188	-43.20	0.82	$+$	$+$	$+$
$\gamma44$	0.01	-0.374	0.272	-0.343	0.273	-0.330	0.273	-27.92	0.19	$+$	$+$	$+$
$\sigma1$	0.6	-0.077	0.986	-0.075	0.990	-0.074	0.993	-25.57	235.79	$+$	$+$	$+^{**}$
$\sigma2$	0.5	-0.107	0.684	-0.107	0.681	-0.106	0.680	-25.58	163.88	$-$	$-$	$-$
$\sigma3$	0.4	-0.067	0.522	-0.067	0.527	-0.067	0.529	-37.51	225.31	$+^{**}$	$+^{**}$	$+^{***}$
$\sigma4$	0.3	-0.065	0.269	-0.064	0.270	-0.064	0.270	-30.73	85.10	$+$	$+$	$+$
$\rho12$	-0.2	-0.173	3.596	-0.209	3.592	.	.	-15.98	12.93	$-$	.	.
$\rho13$	-0.15	-0.067	4.164	-0.131	4.228	.	.	-22.20	13.91	$+^{**}$	.	.
$\rho14$	-0.08	2.393	4.109	2.330	4.106	.	.	-15.06	4.74	$-$	.	.
$\rho23$	-0.15	0.756	4.211	0.666	4.215	.	.	-20.86	12.56	$+$	.	.
$\rho24$	-0.07	-1.203	4.073	-1.242	4.084	.	.	-17.35	4.72	$+$	.	.
$\rho34$	-0.1	-0.183	4.531	-0.227	4.592	.	.	-22.78	8.44	$+^{**}$	.	.
$\alpha01$	0.3	0.189	1.330	0.186	1.331	0.186	1.332	40.36	147.24	$+$	$+$	$+$
$\alpha02$	0.25	-0.069	0.959	-0.069	0.959	-0.069	0.959	40.07	100.82	$-$	$-$	$+$
$\alpha03$	0.05	-1.180	0.776	-1.171	0.781	-1.170	0.783	275.76	190.32	$+^{**}$	$+^{**}$	$+^{***}$
$\alpha04$	0.1	0.299	0.393	0.301	0.393	0.301	0.393	77.79	60.67	$-$	$-$	$+$
$\beta1$	-0.025	0.445	1.007	0.564	1.006	0.615	1.006	-33.30	1.20	$-$	$-$	$+$
$\beta2$	0.03	0.560	0.770	0.560	0.772	0.565	0.774	-31.76	1.31	$+$	$+$	$+$
$\beta3$	-0.01	3.056	0.530	3.139	0.527	3.171	0.526	-55.25	0.48	$-$	$-$	$-$
$\beta4$	0.02	-0.024	0.273	-0.019	0.273	-0.017	0.273	-36.71	0.66	$+$	$+$	$+$

Notes: Parameters refer to the main text (eq. 4). σ and ρ are respectively the standard error and correlation of the error term. Deviation mean (column 2, 4, 6 and 8) shows the percentage deviation of the mean over the 500 simualtions from the true value. Column 10, 11 and 12 show the sign of difference between column 5 and 3, 7 and 3, and 7 and 5, respectively. *, ** and *** indicate that the compared columns are significantly different at 10, 5 and 1 percent based on two sided t-tests.

and the MSEs for all estimators are as expected. The great majority of parameters have means within one percent of the true value and the MSEs have decreased compared to Table 2. However, the SML estimator is now superior to both QML estimators. Not only does it perform better for the great majority of parameters (lower MSEs) as indicated in column 10 and 11 in Table 3, but it is also significantly better for a small number of parameters. The QML_{T2} does a better job than the QML_{T1} estimator (column 12).

The main message from Table 1 to 3 where simulations are done with the base error correlation structure is that it takes a relatively large sample size before the theoretical better performance of the SML estimator shows up. Even then the gains from employing the SML estimator are small. In particular, it is clear that both QML estimators provide very accurate approximations of the SML estimator for the sample sizes examined here, although only the QML_{T2} estimator yields error correlation estimates. To illustrate the last point consider the difference in individual point estimates between the SML and QML_{T1} estimators for the 500 simulations with 3,000 observations. The two estimators have 22 parameters in common since the QML_{T1} estimator does not identify error correlations. For 13 of the 22 parameters the QML_{T1} estimator is never more than 5 percent worse than the SML estimator. For the remaining parameters, more than 85 of the 500 point estimates are not more than 5 percent further from the true value than the SML estimator. The only exception being γ_{22} where only 71 percent lies within this criterion.

Table 4 to 6 are analogous to table 1, 2 and 3, but with the correlation matrix multiplied by two ('large' correlations). The non-censored estimator is not included since the results above showed it to be clearly biased. Although all three tables are presented for completeness, table 4 with 200 observations provides a clear case of how the results differ between the two sets of tables. With the absolute larger correlation coefficients the SML estimator is superior to both the QML_{T2} and the QML_{T1} estimators with very few

Table 4: Simulation results. Large correlations, 200 observations (500 simulations).

Parameter	True value	SML Deviation mean (%)	SML MSE (x10000)	QML-T2 Deviation mean (%)	QML-T2 MSE (x10000)	QML-T1 Deviation mean (%)	QML-T1 MSE (x10000)	Comparison (5) − (3)	Comparison (7) − (3)	Comparison (7) − (5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\gamma 11$	-0.06	-2.626	16.185	-2.902	16.923	-3.068	17.348	$+^{**}$	$+^{***}$	$+^{***}$
$\gamma 12$	-0.03	5.202	7.938	5.446	8.029	5.342	8.119	$+$	$+$	$+^{***}$
$\gamma 13$	0.05	-0.847	5.784	-0.753	5.964	-0.688	6.036	$+^{**}$	$+^{**}$	$+^{***}$
$\gamma 14$	0.02	2.229	3.384	2.305	3.365	2.538	3.386	$-$	$+$	$+$
$\gamma 22$	-0.01	-16.906	11.324	-13.258	11.826	-11.972	12.203	$+^{**}$	$+^{***}$	$+^{***}$
$\gamma 23$	0.02	5.599	5.004	5.381	5.119	5.273	5.222	$+$	$+^{*}$	$+^{***}$
$\gamma 24$	0.01	-10.918	3.022	-10.530	3.096	-10.348	3.127	$+$	$+^{*}$	$+^{**}$
$\gamma 33$	-0.03	3.444	7.858	2.715	8.069	2.477	8.227	$+$	$+^{**}$	$+^{***}$
$\gamma 34$	-0.02	-1.043	2.961	-1.160	2.983	-0.929	2.988	$+$	$+$	$+$
$\gamma 44$	0.01	3.549	3.414	4.721	3.403	5.032	3.408	$-$	$-$	$+$
$\sigma 1$	0.6	-0.798	16.560	-0.980	16.587	-0.999	16.647	$+$	$+$	$+$
$\sigma 2$	0.5	-1.522	11.383	-1.546	11.412	-1.535	11.456	$+$	$+$	$+$
$\sigma 3$	0.4	-1.787	8.933	-1.841	9.297	-1.825	9.398	$+^{**}$	$+^{**}$	$+^{***}$
$\sigma 4$	0.3	-1.726	4.044	-1.722	4.071	-1.715	4.079	$+$	$+$	$+$
$\rho 12$	-0.4	0.018	42.495	-0.545	43.286	.	.	$+$	.	.
$\rho 13$	-0.3	-0.485	54.258	-1.272	55.339	.	.	$+$	.	.
$\rho 14$	-0.16	3.140	66.555	2.427	66.150	.	.	$-$	.	.
$\rho 23$	-0.3	0.420	56.633	-0.538	57.726	.	.	$+$	.	.
$\rho 24$	-0.14	-1.745	55.820	-1.287	57.044	.	.	$+$	.	.
$\rho 34$	-0.2	-1.683	62.616	-2.445	63.271	.	.	$+$	.	.
$\alpha 01$	0.3	-0.859	21.747	-0.755	21.653	-0.743	21.654	$-$	$-$	$+$
$\alpha 02$	0.25	1.280	14.350	1.181	14.481	1.157	14.513	$+$	$+$	$+^{*}$
$\alpha 03$	0.05	-1.166	10.672	-1.220	10.889	-1.307	10.962	$+$	$+^{*}$	$+$
$\alpha 04$	0.1	1.599	5.453	1.584	5.489	1.580	5.499	$+$	$+$	$+$
$\beta 1$	-0.025	-8.928	17.212	-8.708	17.594	-8.450	17.727	$+^{**}$	$+^{**}$	$+$
$\beta 2$	0.03	1.123	12.583	0.290	12.782	0.208	12.873	$+$	$+^{**}$	$+$
$\beta 3$	-0.01	12.967	9.272	15.359	9.332	16.089	9.391	$+$	$+$	$+^{**}$
$\beta 4$	0.02	6.409	4.480	5.921	4.522	5.905	4.538	$+$	$+$	$+$

Notes: Parameters refer to the main text (eq. 4). σ and ρ are respectively the standard error and correlation of the error term. Deviation mean (column 2, 4, 6 and 8) shows the percentage deviation of the mean over the 500 simualtions from the true value. Column 10, 11 and 12 show the sign of difference between column 5 and 3, 7 and 3, and 7 and 5, respectively. *, ** and *** indicate that the compared columns are significantly different at 10, 5 and 1 percent based on two sided t-tests.

Table 5: Simulation results. Large correlations, 1,000 observations (500 simulations).

Parameter	True value	SML Deviation mean (%)	SML MSE (x10000)	QML-T2 Deviation mean (%)	QML-T2 MSE (x10000)	QML-T1 Deviation mean (%)	QML-T1 MSE (x10000)	Comparison (5) − (3)	Comparison (7) − (3)	Comparison (7) − (5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\gamma11$	-0.06	-2.253	2.965	-2.459	2.975	-2.397	2.995	$+$	$+$	$+^{*}$
$\gamma12$	-0.03	1.594	1.376	1.407	1.411	1.185	1.450	$+^{*}$	$+^{**}$	$+^{***}$
$\gamma13$	0.05	-1.285	1.016	-1.201	1.057	-1.182	1.073	$+^{**}$	$+^{***}$	$+^{***}$
$\gamma14$	0.02	1.096	0.583	1.304	0.593	1.249	0.598	$+$	$+$	$+^{**}$
$\gamma22$	-0.01	-9.747	2.259	-8.951	2.350	-8.396	2.435	$+^{*}$	$+^{**}$	$+^{***}$
$\gamma23$	0.02	-2.811	0.930	-2.730	0.986	-2.741	1.008	$+^{***}$	$+^{***}$	$+^{***}$
$\gamma24$	0.01	-1.001	0.605	-1.395	0.613	-1.589	0.618	$+$	$+$	$+^{**}$
$\gamma33$	-0.03	-4.361	1.354	-4.240	1.456	-4.236	1.492	$+^{***}$	$+^{***}$	$+^{***}$
$\gamma34$	-0.02	2.006	0.522	2.410	0.534	2.436	0.539	$+$	$+^{*}$	$+^{***}$
$\gamma44$	0.01	4.491	0.761	3.452	0.773	3.678	0.780	$+$	$+$	$+^{*}$
$\sigma1$	0.6	-0.166	2.920	-0.174	3.009	-0.171	3.030	$+^{*}$	$+^{**}$	$+^{***}$
$\sigma2$	0.5	-0.130	2.061	-0.153	2.129	-0.162	2.148	$+^{***}$	$+^{***}$	$+^{***}$
$\sigma3$	0.4	-0.307	1.579	-0.283	1.663	-0.279	1.681	$+^{***}$	$+^{***}$	$+^{***}$
$\sigma4$	0.3	-0.295	0.822	-0.283	0.819	-0.283	0.820	$-$	$-$	$+$
$\rho12$	-0.4	0.170	8.536	0.029	8.922	.	.	$+^{***}$	.	.
$\rho13$	-0.3	-0.632	9.654	-0.786	10.316	.	.	$+^{***}$	.	.
$\rho14$	-0.16	0.232	13.439	0.219	13.460	.	.	$+$	.	.
$\rho23$	-0.3	0.447	10.124	0.439	10.440	.	.	$+$	.	.
$\rho24$	-0.14	0.166	11.079	-0.042	11.511	.	.	$+^{***}$	.	.
$\rho34$	-0.2	-0.514	12.119	-0.512	12.586	.	.	$+^{**}$	.	.
$\alpha01$	0.3	-0.352	4.113	-0.365	4.113	-0.376	4.124	$+$	$+$	$+$
$\alpha02$	0.25	0.160	2.856	0.183	2.821	0.198	2.823	$-$	$-$	$+^{*}$
$\alpha03$	0.05	2.892	2.160	2.628	2.198	2.590	2.210	$+$	$+^{**}$	$+$
$\alpha04$	0.1	-0.723	1.094	-0.759	1.095	-0.788	1.097	$+$	$+$	$+$
$\beta1$	-0.025	-3.892	3.094	-3.931	3.119	-3.915	3.127	$+$	$+$	$+$
$\beta2$	0.03	-0.887	1.826	-0.544	1.862	-0.435	1.865	$+^{**}$	$+^{**}$	$+^{**}$
$\beta3$	-0.01	9.496	1.465	9.814	1.519	9.259	1.542	$+^{***}$	$+^{***}$	$+$
$\beta4$	0.02	-0.740	0.994	-0.427	0.998	-0.355	0.997	$+$	$+$	$-$

Notes: Parameters refer to the main text (eq. 4). σ and ρ are respectively the standard error and correlation of the error term. Deviation mean (column 2, 4, 6 and 8) shows the percentage deviation of the mean over the 500 simualtions from the true value. Column 10, 11 and 12 show the sign of difference between column 5 and 3, 7 and 3, and 7 and 5, respectively. *, ** and *** indicate that the compared columns are significantly different at 10, 5 and 1 percent based on two sided t-tests.

Table 6: Simulation results. Large correlations, 3,000 observations (500 simulations).

Parameter	True value	SML Deviation mean (%)	SML MSE (x10000)	QML-T2 Deviation mean (%)	QML-T2 MSE (x10000)	QML-T1 Deviation mean (%)	QML-T1 MSE (x10000)	Comparison (5) − (3)	Comparison (7) − (3)	Comparison (7) − (5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\gamma 11$	-0.06	-0.614	0.948	-0.824	0.996	-0.897	1.018	$+^{***}$	$+^{***}$	$+^{***}$
$\gamma 12$	-0.03	-1.202	0.460	-1.158	0.459	-1.160	0.461	$-$	$+$	$+$
$\gamma 13$	0.05	-0.454	0.324	-0.366	0.334	-0.352	0.339	$+$	$+^{**}$	$+^{***}$
$\gamma 14$	0.02	-0.129	0.212	-0.001	0.219	0.008	0.220	$+^{**}$	$+^{**}$	$+^{*}$
$\gamma 22$	-0.01	-2.396	0.714	-0.887	0.740	-0.408	0.756	$+^{*}$	$+^{**}$	$+^{***}$
$\gamma 23$	0.02	-1.967	0.334	-2.269	0.342	-2.302	0.347	$+$	$+^{*}$	$+^{***}$
$\gamma 24$	0.01	-0.241	0.211	-0.792	0.218	-0.886	0.221	$+^{**}$	$+^{**}$	$+^{***}$
$\gamma 33$	-0.03	-1.113	0.498	-1.407	0.519	-1.440	0.527	$+^{*}$	$+^{**}$	$+^{***}$
$\gamma 34$	-0.02	-0.722	0.179	-0.776	0.187	-0.777	0.189	$+^{***}$	$+^{***}$	$+^{***}$
$\gamma 44$	0.01	-0.403	0.275	-0.415	0.281	-0.427	0.282	$+^{**}$	$+^{*}$	$+$
$\sigma 1$	0.6	-0.083	0.945	-0.080	0.982	-0.078	0.992	$+$	$+^{***}$	$+^{***}$
$\sigma 2$	0.5	-0.143	0.652	-0.146	0.651	-0.144	0.655	$-$	$+$	$+^{**}$
$\sigma 3$	0.4	-0.013	0.496	0.004	0.525	0.006	0.532	$+^{***}$	$+^{***}$	$+^{***}$
$\sigma 4$	0.3	-0.072	0.272	-0.063	0.276	-0.062	0.276	$+$	$+$	$+^{*}$
$\rho 12$	-0.4	0.032	2.660	-0.011	2.703	.	.	$+$	.	.
$\rho 13$	-0.3	-0.131	3.681	-0.198	3.893	.	.	$+^{***}$	.	.
$\rho 14$	-0.16	0.988	3.934	0.899	3.956	.	.	$+$	.	.
$\rho 23$	-0.3	0.044	3.332	-0.029	3.500	.	.	$+^{***}$	.	.
$\rho 24$	-0.14	-0.695	3.658	-0.698	3.766	.	.	$+^{**}$	.	.
$\rho 34$	-0.2	-0.291	3.959	-0.135	4.148	.	.	$+^{***}$	.	.
$\alpha 01$	0.3	0.202	1.310	0.191	1.326	0.190	1.333	$+$	$+^{*}$	$+$
$\alpha 02$	0.25	-0.043	0.935	-0.044	0.938	-0.046	0.941	$+$	$+$	$+$
$\alpha 03$	0.05	-1.589	0.762	-1.701	0.768	-1.710	0.771	$+$	$+$	$+$
$\alpha 04$	0.1	0.315	0.404	0.295	0.403	0.295	0.403	$-$	$-$	$+$
$\beta 1$	-0.025	0.094	0.994	0.458	1.000	0.562	1.004	$+$	$+$	$+^{**}$
$\beta 2$	0.03	0.621	0.773	0.669	0.782	0.698	0.786	$+$	$+$	$+$
$\beta 3$	-0.01	2.300	0.530	2.348	0.536	2.388	0.539	$+$	$+$	$+$
$\beta 4$	0.02	0.591	0.263	0.675	0.263	0.695	0.263	$+$	$+$	$+$

Notes: Parameters refer to the main text (eq. 4). σ and ρ are respectively the standard error and correlation of the error term. Deviation mean (column 2, 4, 6 and 8) shows the percentage deviation of the mean over the 500 simualtions from the true value. Column 10, 11 and 12 show the sign of difference between column 5 and 3, 7 and 3, and 7 and 5, respectively. *, ** and *** indicate that the compared columns are significantly different at 10, 5 and 1 percent based on two sided t-tests.

(five) MSEs larger than those for the two QML estimators. Further, significant differences show up for a substantial number of coefficients. Similarly, the QML_{T2} estimator, which takes error correlations into account, performs much better than the QML_{T1} estimator that does not. In table 5 with 1,000 observations this is even more evident. The majority of parameters show the SML estimator to be significantly better performing than the two other estimators, whereas the same is the case for the QML_{T2} versus the QML_{T1} estimator. The picture is the same in table 6 where the simulations are done on 3,000 observations.

Table 4, 5 and 6 show, that with larger correlation coefficients there are significant gains from using more sophisticated estimators and that the gains are apparent at all sample sizes analyzed here. Since it is often difficult to have a prior opinion on the size of the correlation coefficients for a given application, one recommendation would be to first apply the QML_{T2} estimator to assess the size of the correlation coefficients before considering to go on with the SML estimator. In that case, the QML_{T2} provides some very accurate starting values.

5 Concluding remarks

The results in this paper indicate that there is very little to gain from using a SML estimator compared to the two simpler QML estimators investigated here, if the absolute size of the error correlation coefficients is of the same magnitude as usually found in empirical studies. However, the error correlation structure can not be known prior to an application, and if these are large in absolute value there will be gains from using the asymptotically better SML estimator. In this case both QML estimators provide good starting values for the SML estimator; something which is useful in the cause of achieving convergence of the maximum likelihood routine.

Even if Monte Carlo simulations are subject to the problem of specificity which makes broad generalizations of the results difficult, this study has shown that for moderate sample sizes most commonly found in empirical applications simple QML estimators perform surprisingly well. The results herein also suggest that QML estimators of a similar type to those presented here might be useful in more general systems of censored equations.

References

[1] Avery, R.B., Hansen, L.P. & Hotz, V.J., 1983, 'Multi-period probit models and orthogonality condition estimation', *International Economic Review 24*, pp. 21-35.

[2] Barslund, M., 2007, 'Food consumption in urban Mozambique: A censored demand system approach', Discussion Paper, Department of Economics, University of Copenhagen.

[3] Börsch-Supan, A. & Hajivassiliou, V.A., 1993, 'Smooth Unbiased Multivariate Probability Simulatorfor Maximum Likelihood Estimation of Limited Dependent Variable Models', *Journal of Econometrics 58*, pp. 347-368.

[4] Cameron, A.C. & Trivedi, P.K., 2005, '*Microeconometrics - methods and applications*', Cambridge University Press, US.

[5] Cappellari, L & Jenkins, S.P., 2006, 'Calculation of multivariate normal probabilities by simulation, with applications to maximum simulated likelihood estimation', *The Stata Journal 6*, pp. 156-189.

[6] Christensen L.R., Jorgenson, D.W. & Lau, L.J., 1975, 'Transcendental Logarithmic Utility Functions', *American Economic Review 65*, pp. 367-383.

[7] Deaton, A. & Muellbauer, J., 1980, 'An almost ideal demand system', *American Economic Review 70*, pp. 312-326.

[8] Greene, W.H., 2000, *Econometric analysis*, 4. ed., Prentice Hall Int., US.

[9] Haan, P., and Uhlendorff, A., 2006, 'Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood', *Stata Journal 6*, pp. 229-245.

[10] Harris, T.R. & Shonkwiler, J.S., 1997, 'Interdependence of Retail Businesses', *Growth and Change 28*, pp. 520-533.

[11] Lee, F. & Pitt, M.M., 1986, 'Microeconometric demand systems with binding nonnegativity constraints: the dual approach', *Econometrica 54*, No. 5, pp. 1237-42.

[12] Moschini, G., 1998, 'The semiflexible almost ideal demand system', *European Economic Review 42*, pp. 349-364.

[13] Pudney, S., 1989, '*Modelling Individual Choice: Econometrics of Corners, Kinks and Holes*', Cambridge: Blackwell Publishers.

[14] Shonkwiler, J.S. & Yen, S.T., 1999, 'Two-step etsimation of a censored system of equations', *American Journal of Agricultural Economics 81*, pp. 972-82.

[15] Train, K., 2003, 'Discrete Choice Methods with Simulation', Cambridge, Cambridge University Press.

[16] Yen, S.T., Fang, C. & Su, S-J., 2004, 'Household food demand in urban China: a censored system approach', *Journal of Comparative Economics 32*, pp. 564-585.

[17] Yen, S.T. & Lin, B-H., 2002, 'Beverage consumption among US children and adolesscents: full-information and quasi maximum-likelihood estimation of a censored system', *European Review of Agricultural Economics 29*, pp. 85-103.

[18] Yen, S.T. & Lin, B-H., 2006, 'A sample selection approach to censored demand systems', *American Journal of Agricultural Economics 88*, pp. 742-749.

[19] Yen, S.T., Lin, B-H. & Smallwood, D.M., 2003, 'Quasi- and simulated-likelihood approaches to censored demand systems: food consumption by food stamp recipients in the United States', *American Journal of Agricultural Economics 85*(2), pp. 458-478.

[20] Wales, T.J. & Woodland A.D., 1983, 'Estimation of consumer demand systems with binding non-negativity constraints', *Journal of Econometrics 21*, pp. 262-285.

[21] White, H., 1982, 'Maximum likelihood estimation of misspecified models', *Econometrica 50*, pp.1-25.