Barslund, Mikkel

Working Paper

Censored Demand System Estimation with Endogenous Expenditures in clustered samples: an application to food demand in urban Mozambique

LICOS Discussion Paper, No. 280

Provided in Cooperation with:
LICOS Centre for Institutions and Economic Performance, KU Leuven

Suggested Citation: Barslund, Mikkel (2011) : Censored Demand System Estimation with Endogenous Expenditures in clustered samples: an application to food demand in urban Mozambique, LICOS Discussion Paper, No. 280, Katholieke Universiteit Leuven, LICOS Centre for Institutions and Economic Performance, Leuven

This Version is available at:
https://hdl.handle.net/10419/74834

LICOS Centre for Institutions and Economic Performance

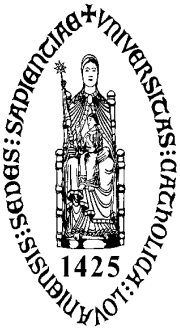
Centre of Excellence

LICOS Discussion Paper Series

Discussion Paper 280/2011

Censored Demand System Estimation with Endogenous Expenditures in Clustered Samples: an Application to Food Demand in Urban Mozambique

Mikkel Barslund



Katholieke Universiteit Leuven

LICOS Centre for Institutions and Economic Performance
Waaistraat 6 – mailbox 3511
3000 Leuven

BELGIUM

TEL:+32-(0)16 32 65 98
FAX:+32-(0)16 32 65 99
http://www.econ.kuleuven.be/licos

Censored demand system estimation with endogenous expenditures in clustered samples: an application to food demand in urban Mozambique

Mikkel Barslund[a]
LICOS Centre for Institutions and Economic Performance, University of Leuven (KUL)

This version: April 2011

Abstract:

We address the issue of endogenous expenditures in the context of a censored demand system by an augmented regression approach estimated with a two-step estimator. An application to food demand by urban households in Mozambique shows that accounting for endogeneity is potentially important in obtaining reliable point estimates of price and, in particular, expenditure elasticities. Furthermore, a bootstrap approach to obtain confidence intervals when data are clustered – as is the case with most household surveys – is devised. Based on a Monte Carlo exercise we speculate that previous studies in failing to account for the clustered nature of the data overstate the precision with which elasticities are estimated.

Keywords – Censored demand system, Endogeneity, Survey data, Elasticities, Mozambique, Food demand.

JEL classification: D12, O12

[a] Address correspondence to: mikkel.barslund@econ.kuleuven.be. I thank Thomas Barnebeck Andersen, Panagiotis Lazaridis, Biing-Hwan Lin, Carol Newman, Finn Tarp and Katleen Van den Broeck for helpful discussions, comments on earlier versions and useful suggestions in the early stages of the research project. All remaining errors are mine.

1. Introduction

Detailed knowledge of households' responses to price changes is a valuable tool for improving policy advice and evaluating the effects of existing policies. Important areas where such knowledge can improve policy advice span a range from tax reform and transfers to public goods provision. In addition, the usefulness of larger scale economy-wide or multi market models in delivering policy relevant quantities hinge, among other things, on having the magnitude of parameters driving consumption behaviour right (see e.g. Jensen & Tarp, 2004; Lin et al., 2010). A particular point in case is the recent attempts at estimating the welfare consequences stemming from the surge in food prices observed during 2008 and more recently (e.g. ul Haq et el., 2008). The proliferation of large household surveys with extensive modules capturing household expenditure has contributed to the increase in the number of studies looking at households' response to changes in prices for a variety of commodities. Household surveys in developing countries often – as in the case of the Mozambican survey utilised in the present study – exhibit spatial price variation which can be used to estimate key income and price response parameters.

Using detailed household survey data for demand analysis has the added advantage that demographic variables at the household level can be included in the analysis. This makes it possible to (partly) control for household heterogeneity in the parameters. However, using survey data at the household level also poses challenges. A recurring problem, known as censoring, is the potentially wide presence of households reporting zero consumption of one or more of the commodities analyzed. Recently, a number of studies have made important contributions to the understanding of how to formulate and estimate demand systems where censoring is non-negligible (see Shonkwiler and Yen, 1999; Perali and Chavas, 2000; Yen, Lin and Smallwood, 2003, Lazaridis, 2003; Dong, Gould and Kaiser, 2004; Meyerhoefer, Ranney and Sahn, 2005; Yen, 2005; Yen and Lin, 2006; Millimet and Tchernis, 2008; Yen, Yuan and Liu, 2009).

The present study builds on this literature by devising a method to control for endogenous total expenditure in a censored demand framework. In the literature on standard (non-censored) demand analysis total expenditure has long been recognized to be potential endogenous because of possible correlation with unobserved characteristics affecting demand behaviour or because of shocks common to total expenditure and expenditure shares (Blundell and Robin, 1999; Robin and Lecoeq, 2006). There is nothing to suggest this should not carry over to the estimation of censored demand systems, although, its severity will clearly be application specific. In the present paper, we account for endogeneity by applying the augmented regression approach of Hausman (1978) and Blundell & Robin (1999) to a system of censored demand equations based on the Almost Ideal Demand (AID) system (Deaton & Muellbauer, 1980). Censoring is taken into account using the two-step approach of Shonkwiler and Yen (1999).[1] While the censoring mechanism used here is of the popular two-step version, the framework for dealing with endogeneity can be generalized to a number of recently suggested methods for accounting for censoring, such as the system of Tobit equations (Yen, Lin and Smallwood, 2003) and the sample selection approach of Yen and Lin (2006). The method is relatively easily implemented and intuitive; in a first stage linear regression, total expenditure is regressed on prices, demographic and other variables included in the system, and the instrument(s). The residual from this regression is then included in the demand system as an additional explanatory variable. A straightforward test for endogeneity for each demand equation in the system is the significance of the included residual. If total expenditure is exogenous (and the instruments are valid) the coefficient on the residual should be insignificant.

The second step of the system is estimated by the iterated least squares method outlined in Blundell and Robin (1999). By doing this, the non-linear nature of the AID system is preserved without the need to rely on maximum likelihood estimation, which in turn allows for the inclusion of a large number of parameters. This approach has not been utilized in a censored system framework even though it offers distinct advantages. Two-step Shonkwiler and Yen-type systems have hitherto either been implemented with a linear version of the

[1] For a recent application of this method see Akbay, Boz and Chern (2007).

AID model (i.e. Akbay, Boz and Chern, 2007; Lazaridis, 2003) to avoid the use of maximum likelihood estimation in the second step, or via maximum likelihood procedures (i.e. Yen, Kan and Su, 2002; Yen and Lin, 2006). The linear AID model can lead to inconsistent estimates of relevant quantities and maximum likelihood estimation is computationally expensive with large systems, in particular if – as in our case – there is a need to make inference robust to deviation from the standard i.i.d. assumption. In the present case a bootstrapping approach to estimation of standard errors is adopted and its validity together with that of the estimation procedure is assessed through Monte Carlo simulations. This offers two advantages over analytical standard errors. First, confidence intervals for elasticities are obtained without relying on the delta method, since they come as part of the bootstrapping of parameter estimates. Second, and importantly, bootstrapping offers a way to let the confidence intervals reflect the clustered sampling frame used in most household surveys. The literature on censored demand systems is silent on the effect of departures from the standard i.i.d. error assumption on confidence intervals for elasticities and other quantities of interest. We suggest a bootstrap strategy which – as shown in Monte Carlo simulations – is robust to some forms of departures from the standard error distribution assumption. This is likely to be relevant when data is collected as part of a household survey where sampling was based on clusters which is often the case. The resulting standard errors are conservative, lending more credibility to hypothesis testing.[2]

The method is applied to a large demand system for food products in urban Mozambique. In particular, we rely on a nationally representative cross sectional data set for Mozambique (IAF2003) collected in 2002/03 to estimate a large complete demand system for 12 food groups. As with most large household surveys the sampling of households was done with a multistage design where primary sampling units (clusters) were first sampled followed by the selection of households within clusters. The data set has the added advantage that it was

[2] An alternative to the bootstrap methodology is to derive the analytical covariance matrix. However, this becomes rather involved due to the estimation in two steps (see Murphy and Topel (1985) and Blundell and Robin (1999)) and the presence of a generated regressors in the second step. It is not clear how this approach could be made robust to clustering of the errors.

collected throughout a full year, and with around 4,000 observations is relatively large. Thus, it contains ample price variation over and above what exists between locations as a result of lack of market integration. Therefore, price responses are expected to be estimated with better precision than is usually obtainable from cross-sectional samples relying on spatial price variation only. Although, as will be clear, the robust standard error approach we are advocating comes with the cost of relative wide confidence intervals around estimated elasticities. Censoring is rather severe in the sample of urban Mozambican households considered here warranting a censored system approach.

Mozambique is a poor sub-Saharan African country with per capita income of 838 USD (PPP adjusted) in 2008 (World Development Indicators), where the bulk of expenditures are directed towards food consumption. Hence, the focus here is on food demand. We pay particular attention to geographical differences in demand patterns by including indicator variables for the three main regions; south, central and north. Given the geography of Mozambique, a large more than 1,500 km long north-south stretched country along the Indian ocean from South Africa to Tanzania, and the poor infrastructure, there is likely to be differences in food demand due to culture and accessibility of food resources.

As a preview of the findings of the paper, we show that the suggested setup delivers consistent estimates and approximately correct standard errors when the error structure is heteroscedastic and contains correlated errors within clusters. The downside is that confidence intervals around point estimates tend to be wide, which limits our ability to significantly tell if food groups are luxuries or necessities, or whether they are own price elastic or inelastic. This may at first be disappointing from a policy perspective, but if it reflects uncertainty as to how much confidence one can have in price response estimates based on survey data from a common sized sample, we believe it is an importance message to convey. Equally, if not more important, we find that correcting for potential endogeneity of total food expenditure has a large impact on point estimates, which leads to the conjecture that estimates from previous studies might have been contaminated by

endogeneity. Controlling for regional differences in tastes is shown to be important in the case of Mozambique.

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. Section 4 is devoted to a small Monte Carlo study. Results are presented in section 5, while section 6 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).[3] 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 ample price variation to identify price responses relative to what is usually available from surveys spanning a shorter time period. It is natural to divide the 11 provinces of Mozambique into three distinct regions; south, central and north. The south is

[3] See also MPD (2004)

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.

The estimated food demand system 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.

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. 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.[5] Households consuming less than 3 of the 12 food groups were excluded resulting in a final sample of 3,938 households.

[4] 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. Excluding rural household to some extent limits the usefulness of the elasticity estimates obtained here for nationwide policy analysis. However, we exclude them to focus on our main points without additional complications.

[5] The treatment of unit values is a contentious issue in demand system estimation. Using household level unit values is likely to induce endogeneity problems because of quality differences among households' purchases of food products (Deaton, 1988). Using enumeration area mean prices has been shown to perform well when comparing with estimates obtained using market prices (Niimi, 2005). For an alternative see Lazaridis (2003).

[Table 1 about here]

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 23 percent of the budget in the central region is less important in the north and only accounts for around 3 percent of expenditures in the south. Fish and other staples – mostly cassava and potatoes – are the most important food products 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.

The two last columns 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 roughly 90 percent of the households, most food groups have a substantial number of households with zero-purchases.

Apart from dummy variables for location, south and north (central is the base specification), a number of additional explanatory variables are included in the analysis.

[Table 2 about here]

Table 2 lists some summary statistics for the demographic and location dummy variables. The sample is roughly equally divided between the three geographical areas. To control for economies of scale in food preparation household size is measured in household adult

equivalents related to energy requirements (FAO/WHO/UNU, 1985). The adult equivalent household size in the sample ranges from 0.6 (a single woman aged 73) to 16.9. To capture seasonal effects dummies are included for the quarter of the year when the household was interviewed. The urban part of the sample we consider here is not completely balanced between quarters. In addition the gender and age of the household head act as extra control variables together with dummy variables for the educational level of the household head, since food preferences may vary with gender, age and education. Education is controlled for through 4 dummy variables equal to one if the household head has, respectively, no education, first lower primary (EP1), second higher primary (EP2), or secondary or higher education (base specification). The values reflect the relatively low education level prevalent in Mozambique. Around half the household heads in the sample have not completed basic primary education. A further 25 percent have completed 5 years of schooling (lower primary), while the remaining quarter of household heads have 7 or more years of schooling. Finally, we include a dummy for the presence of a woman with completed primary education in the household.

The two last rows of table 2 show summary statistic for the two dummy variables used to instrument total food expenditure. These are whether the household owns at least one bike and whether the household has access to a safe drinking water source. We return to these in the result section.

3. Demand model and estimation

Our starting point is a latent share formulation of the well-known AID system linear in the logarithm of total food expenditure. The basic framework is presented in a number of articles (see e.g. Akbay, Boz and Chern, 2007), hence we only emphasize the main points. Since particular interest is paid to estimation of household survey data with clustered sampling we make an explicit reference to the cluster/group, *g*, the household belongs to. Thus, the latent demand for food group *j* for household *i* in cluster *g* is given by the share equation

$$w_{j,ig} = \alpha_{j0} + \sum_{h=1}^{H} \alpha_{jh} D_{ig,h} + \sum_{k=1}^{J} \gamma_{j,k} lnp_{k,ig} + \beta_j \left(lnx_{ig} - ap_{ig} \right) + u_{j,ig} \text{ with} \quad (1)$$
$$ap_{ig} = \sum_{k=1}^{J} \left(\alpha_{j0} + \sum_{h=1}^{H} \alpha_{jh} D_{ig,h} \right) lnp_{k,ig} + \frac{1}{2} \sum_{j}^{J} \sum_{j}^{J} \gamma_{j,k} lnp_{j,ig} lnp_{k,ig},$$

where $D_{ig,h}$ indicates the *h'*th demographic or location variable for household *ig*. Logarithmic prices and total food expenditure are denoted $lnp_{j,ig}$ and lnx_{ig}, respectively. Note that demographic and location variables enter non-linearly via the price index function, ap_{ig}. The total number of food groups is denoted by *J*.

In the latent share equation, (1), total expenditure is likely to be endogenous due to total food expenditure being correlated with unobserved characteristics affecting demand behaviour, or because of shocks common to total food expenditure and some of the expenditure shares. In the demand system literature where censoring is not central, total expenditure is often found to be endogenous (e.g. Blundell and Robin, 1999). In this case estimated parameters will be biased.

To address this issue we extend the augmented regression framework used by Blundell and Robin (1999) (building on Hausman (1978)) to the censored demand system case. Assume the error terms $u_{j,ig}$ have the orthogonal decomposition

$$u_{j,ig} = \rho_j r_{ig} + \varepsilon_{j,ig} \tag{2}$$

where r_{ig} are the residuals from the regression of total expenditure on the set of instruments and $\varepsilon_{j,ig}$ is normally distributed with zero mean and covariance matrix Σ_{ig}. This allows for heteroscedasticity as well as correlated errors within clusters. Errors are assumed independent across clusters, g. The parameters, ρ_j, provide a test of exogeneity of total expenditure for each consumption share, since under the null hypothesis of total expenditure being exogenous, they should equal zero.

The observed expenditure shares result from a Shonkwiler and Yen (1999) type specification. Specifically, let the dichotomous variable $d_{j,ig}$ take the form

$$d_{j,ig} = \begin{cases} 1 \; if \; \gamma_j z_{ig} + v_{j,ig} > 0 \\ 0 \; otherwise \end{cases} \tag{3}$$

where γ_j is a vector of coefficients and z_{ig} a vector of explanatory variables. The equation specific error term, $v_{j,ig}$, is distributed normally with zero mean and unit variance, but with errors allowed to be correlated within but not across clusters. The observed expenditure shares are then given by

$$w_{j,ig}^{obs} = (w_{j,ig} + \rho_j r_{ig}) \cdot d_{j,ig}. \tag{4}$$

To arrive at the popular two-step estimating equations assume that $(v_{j,ig}, \varepsilon_{j,ig})$ are distributed bivariate normally with covariance $Cov(v_{j,ig}, \varepsilon_{j,ig}) = \delta_j$. Consistent parameters in the latent share equation can then be recovered by estimating the observed share equations

$$w_{j,ig}^{obs} = \Phi(\hat{\gamma}_j z_{ig})(\mathrm{w_{j,ig}} + \rho_j r_{ig}) + \delta_j \phi(\hat{\gamma}_j z_{ig}) + \xi_{j,ig} \tag{5}$$

where $\hat{\gamma}_j z_{ig}$ are predicted indices from the first step probit estimation of the equations in (3) and ϕ and Φ are, respectively, the standard normal density and cumulative density functions. As pointed out by Shonkwiler and Yen (1999) the composite error terms in (5), $\xi_{j,ig}$, are conventionally heteroscedastic since they depend on the terms $\Phi(\hat{\gamma}_j z_{ig})$, $w_{j,ig}$ and $\phi(\hat{\gamma}_j z_{ig})$ on the right hand side of (5). Furthermore, in the less restrictive setup here, $\xi_{j,ig}$ are correlated within sampling clusters, since $\varepsilon_{j,ig}$ are allowed to be correlated within clusters.

In the literature the system of equations given by (5) has been estimated with either Seemingly Unrelated Regression (SUR) or ML (see i.e. Akbay, Boz and Chern, 2007 and Yen, Kan and Su, 2002). To invoke SUR one has to approximate the price index such that it is exogenous to the system of equations (so that it does not depend on system parameters). The approximation will bias the estimated coefficients. An alternative is ML estimation of the second step and subsequent correction of the covariance matrix following Murphy and Topel (1985) (see Yen, Kan and Su, 2002) – or, as in Yen and Lin (2006), specifying the complete distribution of errors and do the full system estimation in one step – however, when the system is large with many demographic variables, this becomes computationally expensive. In addition, it is unclear how these estimators would fare when errors are allowed to be correlated within clusters and when additional generated regressors are introduced to control for endogeneity.

To estimate the system given by (5) we use the Iterated Linear Least squares Estimator (ILLE) proposed by Blundell and Robin (1999) for the second step estimation. The complete estimation procedure then follows the following pattern: in the first step, estimate the first stage regression of the logarithm of total expenditure on instruments and other explanatory variables and form the residuals to include in the term for the latent expenditure share (equation (1) and (2)). Estimate the system (3) by univariate or multivariate probits. In the second step, give initial parameter values to form the price

index, ap_{ig}. Given the initial values of the price index, the residuals and the predicted linear indices from the probit regressions, form the system (5) which is now linear in parameters and estimate each equation by OLS. This results in a new set of parameter estimates which are used to update the price index ap_{ig}. Inserting the updated price index in equation (5) and estimate by OLS yields another set of parameters which are used to update the price index. This process continues until parameter estimates do not change. In practice this is most often achieved within a few iterations, even for quite strict levels of tolerance for convergence.[6]

As showed by Blundell and Robin (1999) the ILLE estimator is consistent for the class of conditional linear systems to which the system in (5) belongs. Thus, the same procedure can be utilized if the translog demand system (Christensen, Jorgensen and Lau, 1975) is used instead of the AID system.

Before we turn to the discussion of inference and estimation of standard errors a couple of observations are noteworthy. First, the augmented regression framework considered here to account for endogeneity is equally applicable to demand systems formulated as a system of Tobit equations (Yen, Fang and Su, 2004; Yen, Lin and Smallwood, 2003) – as an extension to the results for the univariate tobit equation in Smith and Blundell (1986) – and to the sample selection approach of Yen and Lin (2006). It is not obvious, however, to what extent it can be modified to work in the copula approach recently introduced in Yen and Lin (2008) and Yen, Yuan and Liu (2009).

The second issue relates to the conditions of homogeneity of degree zero of prices and total expenditure, Slutsky symmetry and adding up of expenditure shares, which a theoretically consistent demand system satisfies. In censored demand system applications, these conditions are routinely imposed on the systems of latent shares in (1). One immediate benefit is the saving in terms of number of parameters to estimate. However, this does not

[6] In the application the tolerance criterion used is that the sum of absolute changes of the parameters is less than 10^{-6}.

ensure that the conditions are satisfied for the system of observed shares in (5), and imposing them on the latent shares is therefore strictly not warranted on theoretical grounds. Consequently, in our application we do not impose any restrictions on the parameters. While it is possible to ensure homogeneity and adding-up of the latent system by parametric restrictions in the ILLE framework, symmetry would have to be accommodated with a minimum distance estimator applied to the unconstrained consistent set of estimates coming out of the procedure described here (Blundell and Robin, 2009; Wooldridge, 2003). The adding-up property of the system of observed shares (4) can be accommodated by estimating only the system of *J-1* food groups and obtain parameters for the *J'th* food group from the adding-up restriction (Yen, Lin and Smallwood, 2003). However, this has the drawback that estimated parameters are not invariant to which food group is determined by the restriction, and the residual expenditure share is not guaranteed to be positive.[7]

Finally, consider the issue of estimating standard errors which allow for correct inference for the estimated parameters. Data for estimating demand systems often originate from household surveys where households are sampled in clusters as part of a multistage sampling design, where first clusters/primary sampling units are selected and then household within each cluster.[8] Therefore, errors are – if not likely to be – potentially correlated within clusters. This issue has been ignored in the literature on censored demand system estimation. We address it by estimating standard errors from a bootstrap approach which samples clusters of households, thus, making it robust to arbitrary correlation structures within clusters. The next section presents a Monte Carlo analysis which shows that the method works well. It also alleges that ignoring error correlation within clusters causes confidence intervals to be misleadingly narrow.

[7] Adding-up is almost satisfied in our application with the sum of expenditure shares being above 0.96 for a reference household.

[8] For instance, Akbay, Boz and Chern (2007:212) mention that data comes from a ".. stratified multistage systematic cluster sampling method ..".

4. Monte Carlo simulation

The principal aim of this section is to verify that the proposed estimator delivers expected results in a sample size resembling the one available for our empirical application. The performance of the clustered bootstrap approach to the estimation of standard errors is also evaluated. In addition, it is shown that not accounting for endogneity biases the estimated coefficients.

A three equation system of the form (1)-(5) with 335 clusters of 12 observations is simulated 1,000 times. All variables are redrawn in each simulation. Full details of the setup are given in Appendix A. The exogenous logarithmic prices are drawn from independent normal distributions. Total expenditure is defined as a function of prices and an error term which is correlated within clusters, with the error term of the (constructed) instrument, and the error term of the latent share equation in (1). The probit equations corresponding to the system (4) have two explanatory variables. Error terms for each probit equation are correlated within clusters, and bivariate normally distributed with the error terms of each of the latent share equations, which are again correlated within clusters and heteroscedastic. The covariance between the probit error terms and the latent share error terms is, respectively, -0.13, -0.25, 0. This mimics the results from the first 3 equations in our application. The design reflects our focus on the impact of endogeneity on coefficient estimates and obtaining correct inference when data come from a clustered survey design.

[Table 3 about here]

Table 3 summarizes the main findings from the Monte Carlo analysis.[9] Only results related to the coefficients on total expenditure are discussed since they will be most affected by endogenous total expenditure. Also presented are results related to the parameters α_1 and $\gamma_{1,1}$ (equation (1)). Results for the other parameters are similar to those. In panel A the mean value of each coefficient for the 1,000 simulations is presented together with its mean squared error (MSE). The endogeneity corrected estimates perform well with the mean

[9] A full set of results and replication code for Stata 10 are available upon request.

values very close to the true values. This confirms that the estimation procedure is well-functioning. Not taking endogeneity into account biases the estimated coefficients as illustrated in column five and six. Of course, the extent of the bias depends on the degree of endogeneity, so the results serve only to demonstrate that the suggested procedure works when endogeneity is present. The bottom three rows of panel A show how the bias in the estimates of the parameters on total expenditure carries over to expenditure elasticities.

Panel B of table 3 presents rejection rates for the hypothesis that the estimated value is equal to the true value at the 5 percent significance level when standard errors are obtained in different ways. The correct interval defined by the estimated standard errors should reject the true hypothesis in around 5 percent of the simulation. Cluster bootstrap refers to standard errors obtained by bootstrapping the standard errors of each of the 1,000 parameter estimates by drawing clusters of observations with replacement (200 replications). This is our preferred method. IID bootstrap is identical to the cluster bootstrap except that individual observations are drawn instead of clusters of observations. This should account for the heteroscedasticity but not within cluster correlation of error terms. Clustered SEs come from using cluster robust standard errors in the last iteration of the ILLE estimator. This to some extent accounts for clustering and heteroscedasticity but ignores the inclusion of generated regressors from the first stage linear regression of instruments on total expenditure and the first step probit estimation. Robust SEs are similar except that the robust estimator is applied in the last iteration of the estimation procedure.

Column three shows that the proposed cluster bootstrap method performs well with rejection rates close to 5 percent. The IID bootstrap on the other hand provides confidence intervals which are too narrow leading to a too high rate of rejections. The Cluster SE approach performs somewhat better but still has a sizeable over-rejection rate. In this case with intra-cluster correlation the method of Robust SEs performs poorly as expected.

Overall, the Monte Carlo simulation suggests that endogeneity is potentially a severe problem in a censored demand system framework and, that it can be effectively dealt with

by the method outlined above. In addition, when data comes from a clustered sample framework, confidence intervals should be obtained by a method robust to both correlation of errors within clusters as well as the multiple step estimation procedure.

5. Results

We now apply the method outlined above to the sample of urban Mozambican households discussed previously. All estimations are carried out using household weights provided by the National Statistical Office in Mozambique. In the first step 12 univariate probit equations are estimated (model equation (3)) using the explanatory variables from table 3: location dummy variables for north and south, adult equivalent household size, dummy variables for season of interview, gender, age and education of household head, and a dummy for the level of female education available within the household. In total 12 coefficients are estimated for each of the 12 equations. Of the 144 estimated coefficients around half (68) are significant at the 5 percent level with cluster robust standard errors. All the explanatory variables are highly significant in at least two equations and pseudo R^2 range from 0.02 to 0.20 with the probability associated with the hypothesis of no joint explanatory power comfortably below $1/100^{th}$ of a percent.[10]

All variables from the first step are included as additional explanatory variables in the second step. To control for endogeneity the residuals from the first stage regression of the logarithm of total food expenditure on explanatory variables and instruments are also included. Including the residuals, the system of equations in the second step has a total of 360 parameters; 30 for each of the 12 equations.[11]

We turn first to the issue of endogeneity of total food expenditure. Total income is often used as an instrument for total expenditure (Blundell and Robin, 1999; Lecocq and Robin, 2006). Unfortunately, we do not have a reliable income measure at our disposal. Instead the two indicator variables reported in table 3 – indicating if the household is in possession of a

[10] Results not shown but are available upon request.
[11] Only selected results related to elasticities are reported. A full set of coefficient estimates are available upon request.

bike, and if it has access to a safe water source – are used as instruments. Both instruments are meant to capture wealth.[12] The instruments are shown to be relevant with good explanatory power and a joint cluster robust F-test for significance of 9.41 with a p-value of 0.0001. The results are reported in table 4.

[table 4 about here]

The extent to which endogeneity is a problem is reported in Table 5. The second column shows the system coefficient estimates on the residual from the first stage with an indication of their significance level based on the bootstrapped standard errors reported in column 3. For 5 of the 12 food groups – vegetables, maize flour, fish, beans, and other foods – exogeneity is rejected at the 5 percent level, while another 3 food groups, bread, meat, and fruits, reject exogeneity at the 10 percent level, thus suggesting that potential endogeneity of total food expenditure is a well-founded concern.

[Table 5 about here]

With two instruments it is possible to assess their validity through an overidentification test. In this context, this is done by regressing the system residuals (equation-by-equation) on the two instruments. The parameter vector from this regression is asymptotically normally distributed (Blundell and Robin, 1999) with a covariance matrix which is obtained from the bootstrapped sample of coefficients. The joint significance of the parameter vector can then be evaluated by a chi2-test of the null-hypothesis that the parameters are jointly zero. Column 4 in Table 5 reports p-values for the null-hypothesis of the instruments having no explanatory power on the system residuals, and therefore being valid as instruments, for those food groups where exogeneity is rejected. For most equations the proposed instrument set does well, however, for meat equation their validity

[12] We did initially experiment with an instrument set additionally including a dummy variable for access to sanitation facilities, a dummy variable for a radio being present in the household and the number of rooms in the house. The results (available upon request) were qualitatively similar to those presented here, but the validity of the instruments was rejected in most equations.

is rejected at the 1 percent. The p-value of 0.13 for the maize equation is also somewhat low, although the null-hypothesis is not rejected at conventional levels. Although, this warrants caution in interpreting the results, we emphasize that the overidentification test firmly rejects endogenous instruments in 6 out of the 8 equations where endogeneity is detected.

Before discussing the implications of this finding in terms of the magnitude of the estimated elasticities we discuss a number of other aspects of the results.

In total 57 (17 percent) of the estimated parameters in the second step are significant at the five percent level and a further 22 at the 10 percent significance level. Of the 144 price response parameters 24 are significant. This includes only one of the own price response parameters which are particular important when calculating own price elasticities. Thus, we cannot hope to recover own price responses with great precision. The same holds for expenditure elasticities since none of the 12 coefficients on total food expenditures are significant. We return to this issue below. Inclusion of the demographic and location variables is warranted from the results; of 144 estimated marginal effects 42 are significant at 5 percent. Table 6 shows the size and significance of the demographic and location variables.

[Table 6 about here]

Each entry in table 6 shows the marginal effect from the demographic and location variables on the observed share evaluated at the sample mean of household size and the age of the household head, and with all dummy variables equal to zero.[13] The reference household is therefore located in the central part of Mozambique, headed by a male with sample average age, who has not completed primary school, and without any female members with a completed primary school education. The reference household has the

[13] Marginal effects and elasticities (as noted by Lazaridis, 2004) are all calculated with respect to observed shares given in equation (4).

sample average number of family members and the expenditure share refers to food consumption in the first quarter of the year.

Looking at each column it is clear that the inclusion of all the demographic variables is warranted. Each one is significant in at least three food demand equations, except for household size and the age of the household head, which are only significant in the maize flour, respectively, the meat equation. Most explanatory variables have small marginal effects as one would hope to find after having controlled for income and prices. In particular the seasonal dummy variables are quite small in magnitude. This lends credibility to the supposition that the demand system manages to pick up changes in relative prices. Seasonal scarcity of the supply of some food products should result in higher relative prices for these products and subsequently lower demand. Were seasonal dummy variables large in magnitude, we would question the reliability of the price response mechanism that we hope to identify.

Contrary to the other explanatory variables the location dummy variables for residing in the south or the north of the country have sizable marginal effects for several of the food demand equations. Only expenditure shares for rice, sugar and the residual food group are not affected by the location of the household given prices and income. The marginal effects mimic the different food consumption patterns for the different regions of Mozambique reported in table 1 quite well. For instance, take maize flour where the observed differences in consumption patterns among south, central and north are the strongest, cf. table 1. The marginal effects of both the north and south dummy variables are large and negative as suggested by the observed shares in table 1. Similarly, for vegetables and fish where both the north and south dummy marginal effects 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 and income differences have been accounted for. In sum, the results listed in Table 6 point to the importance of controlling for location differences in a country as geographically diverse as Mozambique.

The limited precision with which the price response and expenditure coefficients are estimated is expected to carry over to the confidence intervals around the estimated elasticities. Table 7, which shows estimated own price and expenditure elasticities and their standard errors evaluated at the sample mean, confirms this.

[Table 7 about here]

Ten of the 11 own-price elasticties are significantly different from zero and have reasonable values. The one exception is the positive own-price elasticity for maize flour suggesting that it is a Giffen good. Since one would think that maize flour has fairly close substitutes it is difficult to attribute much faith to this particular result. The confidence interval around the estimate is large, so the coefficient is not statistically different from zero. The result could be due to the large differences in consumption of maize flour among the regions, cf. table 1, combined with interaction effects with other explanatory variables which are not accounted for in the estimation. As will be clear below the outlying point estimate for the own-price elasticity for maize flour is not due to the procedure used to control for endogeneity. Based on the point estimates only vegetables, meat and other staples are price inelastic, with bread, rice and beans having a price elasticity close to minus one. Price sensitive food groups consist of fish, oil and fats, fruits, sugar and the residual food group.

While a vast majority of own-price elasticities are significantly different from zero, it is important to note that based on the bootstrapped confidence intervals it is only possible to significantly label meat and other staples as price inelastic. None of the food groups with price elastic point estimates has confidence intervals narrow enough to reject the hypothesis of a coefficient equal to minus one. If the estimates are to be used for policy analysis and modelling of changes in food consumption under different price scenarios, it is important that the true uncertainty surrounding the estimates is revealed, hence the need for reliable confidence intervals accompanying the point estimates.

Columns 4 and 5 of table 7 show estimated expenditure elasticities and associated bootstrapped standard errors.[14] All expenditure elasticities have sensible point estimates, however as with the own-price elasticities, confidence intervals are wide. Although all are clearly significantly different from zero, none of them are significantly different from one at any conventional levels. Vegetables, maize flour, oil and fats, fruits, other staples and the residual food group are necessities, while fish, bread, rice, meat, sugar and beans are found to be luxury goods. Even if this conforms reasonably well with prior expectations, the distinction between necessities and luxury goods is less meaningful when confidence intervals are large. This is particularly so for bread and beans which have point estimates of expenditure elasticities around one.

Before turning to our main interest of differences in elasticities between endogeneity adjusted estimates and unadjusted estimates we briefly touch upon cross-price effects. Unsurprisingly, it is difficult to pick up significant cross-price effects. Of the 132 estimated uncompensated cross-price elasticities 34 (44) are significantly different from zero at 5 (10) percent (not reported). 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, fish, beans and other staples, which all have relatively large cross price effects although few of these are significant. 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).

[Table 8 about here]

Table 8 presents the own price and expenditure elasticities obtained from an estimation of the demand system without correcting for potential endogeneity of total food expenditure together with the differences and their bootstrapped standard errors from the estimates when the correction for endogeneity is introduced (cf. table 7). Recall from table 5 that

[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.

endogeneity of total food expenditure could not be rejected for the following eight food groups: vegetables, maize flour, fish, bread, meat, fruits, beans and the residual food group. This pattern is evident in the third column of table 8. The difference in estimated expenditure elasticities is large in absolute value and significant for seven of the eight food groups where endogeneity is a concern. For meat the difference is substantial but it is only borderline significant at the 10 percent level, which is consistent with the large confidence interval around the estimate reported in table 5. The differences in expenditure elasticities are smaller for the food groups where endogeneity was not detected. The magnitude of the differences suggests that policy prescriptions or modelling results based on estimated elasticities would change significantly if endogeneity is not taken into account. As an illustration consider maize flour with an estimated expenditure elasticity of 0.77 when the correction for endogeneity is applied and 1.26 without the correction, or beans with a difference of similar size.

Considering estimated own price effects, the difference between estimates with and without the endogeneity correction is much less stark as can be seen from columns five to seven in table 8. In fact only the own price of maize flour is seriously affected by the correction. However, as discussed above the point estimate on the own price effect for maize flour is not in line with what we would expect. However, note that even without controlling for endogeneity maize flour continues to have a positive own price effect. This suggests that the unusual result is not generated by the procedure whereby we control for endogeneity. The fact that the own price effects are not affected by the endogeneity correction further reinforces the belief that the suggested procedure takes care of endogeneity of total food expenditure which – without correction – biases the estimated parameters, β_j, on total food expenditure.

To sum up, the application shows that correcting for endogenous total food expenditure is important in obtaining unbiased estimates of expenditure elasticities. It also highlights that not only may parameters be biased, the size of the bias can also substantially affect the

estimation of expenditure elasticities limiting their value to policy analysis and as input in models of economic behaviour.

6. Conclusion

The literature on how to appropriately estimate censored demand systems and its applications to policy analysis has moved forward in recent years. Building on this literature two novelties are introduced in this paper. First, we devise an augmented regression method to account for potential endogeneity of total expenditures in the share equations. The method is based on the literature on estimation of non-censored demand systems. It can be generalized to several of the censored demand system estimators suggested in the literature. Second, building on earlier literature a new implementation of the popular two-step estimator first suggested by Shonkwiler and Yen is set up. In doing this we explicitly account for the clustered nature of the data sample in the estimation of parameters and elasticities through bootstrapping of the standard errors.

Our application to food demand of urban Mozambican households indicates the importance of controlling for potential endogeneity. Large differences in expenditure elasticities are found between estimates with and without accounting for endogeneity. Unfortunately, we also find that the cluster-robust standard errors lead to quite large confidence intervals surrounding the estimated elasticities.

This leads us to suspect that results from previous studies might have been contaminated by endogeneity and – in those cases where data have come from clustered samples – the precision with which parameters and elasticities have been estimated has been overstated.

References

Akbay, C., Boz, I. & Chern, W.S., 2007. Household food consumption in Turkey. European Review of Agricultural Economics 34(2), pp. 209-231.

Blundell, R. & Robin, J-M. 1999. Estimation in large and disaggregated demand systems: an estimator for conditionally linear systems. *Journal of Applied Econometrics* 14, pp. 209-232.

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

Deaton, A. & Muellbauer, J., 1980, 'An almost ideal demand system', *American Economic Review* 70, pp. 312-26.

Deaton, A., 1988, 'Quality, quantity, and spatial variation of price', *American Economic Review* 78, pp. 418-430.

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.

FAO/WHO/UNU. (1985). *Protein and Energy Requirements*, Food and Agriculture Organization, World Health Organization, United Nations University, Rome.

Hausman, J.A., 1978. 'Specification tests in econometrics'. *Econometrica* 46, pp. 931-959.

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.

Lazaridis, P., 2003, 'Household meat demand in Greece: A demand system approach using micro data', *AgriBusinesss* 19(1), pp. 43-59.

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

Lin, B-H., Yen, S.T., Dong, D. & Smallwood, D., 2010, Economic incentives for dietary improvement among food stamp recipients. *Contemporary Economic Policy*, Vol. 28(4), pp. 534-536.

Meyerhoefer, C.D., Ranney, C.K. & Sahn, D.E., 2005, Consistent estimation of censored demand systems using panel data. *American Journal of Agricultural Economics* 87, Iss. 3, pp. 660-672.

Millimet, D.L. & Tchernis, R., 2008, Estimating high-dimensional demand systems in the presence of many binding non-negativity constraints. *Journal of Econometrics* 147, pp. 384-395.

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.

Niimi, Y., 2005, An analysis of household responses to price shocks in Vietnam: Can unit values substitute for market prices? PRUS Working Paper no. 30, Poverty Research Unit of Sussex, Department of Economics, University of Sussex.

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.

Robin, J-M. & Lecoeq, S., 2006. Estimating demand response with panel data. *Empirical Economics 31*, pp. 1043-1060.

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.

Smith, R.J. & Blundell, R.W. 1986. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica* 54(3), PP. 679-685.

ul Haq, Z., Nazli, H. & Meilke, K., 2008, Implications of high food prices for poverty in Pakistan, *Agricultural Economics 39 (2008) supplement*, pp. 477–484.

Wooldridge, J.M., 2003. *Econometric analysis of cross section and panel data*. Cambridge, MA. MIT Press.

Yen, S.T., 2005. A multivariate sample-selection model: Estimating cigarette and alcohol demands with zero observations. *American Journal of Agricultural Economics* 87, Iss. 2, pp. 453-466.

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.

Yen, S.T., 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., 2006. 'A Sample Selection Approach to Censored Demand Systems.' *American Journal of Agricultural Economics* 88(3), pp. 742-749.

Yen, S.T., and B-H. Lin. 2008. Quasi-Maximum Likelihood Estimation of a Censored Equation System with a Copula Approach: Meat Consumption by US Individuals. *Agricultural Economics* 39(2):207-217

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.

Yen, S.T., Yuan, Y. & Liu, X. 2009. Alcohol consumption by men in China: A non-Gaussian censored system approach. *China Economic Review* 20, pp. 162-173.

Appendix A

The three equation system in the Monte Carlo exercise has in total 4,020 observations which are distributed with 12 observations in 335 clusters. All variables are redrawn in each of the 1,000 simulations. Logarithmic prices are drawn from independent normal distributions with zero mean and standard deviations of, respectively, 0.41, 0.36 and 0.41, reflecting the variation found in our application.[15] Total expenditure is generated as a function of prices and standard normal errors, (subscript i is suppressed for brevity and subscript g denotes cluster/group)

$$lnx = .2lnp_1 + .36lnp_2 + .14lnp_3 + e_{lnx} + e_{lnx,g} + .5\left(e_{in} + e_{in,g}\right)$$

where the combined error term $e_{lnx} + e_{lnx,g} + .5\left(e_{in} + e_{in,g}\right)$ ensures correlation within clusters, $e_{lnx} + e_{lnx,g}$, and with the instrument, $.5\left(e_{in} + e_{in,g}\right)$, since the instrument is defined as $inc = 1 + .5\left(e_{in} + e_{in,g}\right) + e_{inc}$. All e's with a subscript refer to standard normal variables.

The censoring mechanism in (3) is constructed as follows:

$$d_j = I(c_j + z_1 + z_2 + v_j + \sqrt{.25}v_{j,g})$$

Here *I()* is an indicator function, c_j is a constant controlling censoring proportions with values equal to 2.5, -.5 and 1, respectively for the three equations. The explanatory variables z_1 and z_2 are, respectively, independently normally distributed, and an independent dummy variable taking the value 1 with probability .5 for each observation. Each v_j is drawn from a bivariate normal distribution together with ε_j (in system equations (1)-(2)) such that $(v_j, \varepsilon_j) \sim N\left(0,0, \begin{matrix} .75 & cv_j \\ cv_j & \sigma_j^2 \end{matrix}\right)$. Thus the combined error term above, $v_j + \sqrt{.25}v_{j,g}$, is correlated within clusters and has unit variance. The covariance between the errors varies with j with values in turn equal to -.13, -.25 and 0. The heteroscedastic ε_j's (σ_j^2's are uniformly distributed on the interval 0.4 to 0.6) form the first part of the combined error $\varepsilon_j + \sqrt{.5}\varepsilon_{j,g} + .5\left(e_{lnx} + e_{lnx,g}\right)$ in (1)-(2). $\varepsilon_{j,g}$ is standard normally

[15] This is the case for all coefficients that differ from unity unless otherwise mentioned.

distributed and correlated within clusters. The second part, $.5\left(e_{lnx} + e_{lnx,g}\right)$, ensures that total expenditure is endogenous in the system of equations. Finally, we need to specify the values of the parameters of interest in (1), the vectors α_0 (the system is not augmented with additional explanatory variables) and β, and the matrix γ. These are

$$\alpha_0 = (.3, .3, .3), \beta = (-.1, .1, .2) \text{ and } \gamma = (.1, -.2, .1 \setminus -.2, .3, -.1 \setminus .2, -1, -1).$$

For each of the 1,000 estimations of the parameter a bootstrap with 200 replications is made to obtain standard errors for, respectively, the cluster bootstrap and the IID bootstrap case. In the cluster bootstrap case 335 clusters are drawn with replacement for each of the 200 replications and the standard deviation of the resulting sample of estimates is used to form confidence intervals for the given draw out of the 1,000 used in the Monte Carlo. A similar procedure is used in the IID bootstrap case except here observations, rather than clusters, are drawn with replacement.

Tables

Table 1. Food consumption for urban Mozambican households.

	South	Central	North	Full Sample	
	Share in total food expenditure (%)			Households consuming (%)	Mean expenditure share (%)
Food group					
Vegetables	19.6	10.9	5.1	90.0	11.9
Maize flour	2.9	23.3	15.3	51.8	13.0
Fish	11.2	13.3	20.9	89.4	15.4
Bread	13.3	5.3	3.5	64.5	7.5
Rice	8.6	9.9	6.5	49.2	8.2
Meat	9.0	6.7	4.3	32.3	6.6
Oil & fats	3.5	5.4	2.5	57.0	3.7
Fruits	12.1	4.0	5.1	82.9	7.3
Sugar	3.5	3.3	3.1	50.2	3.3
Beans	3.5	5.7	4.3	53.3	4.4
Other staples	4.8	5.9	23.8	66.1	12.1
Other food	7.9	6.4	5.6	70.6	6.7
No. Obs. (N)	1,964	1,168	806	3,938	3,938

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.

Table 2. Sample means of demographic and location variables.

Explanatory variables	Description	Mean	Min	Max
South	Household located in south (=1)	0.36	0	1
Central	Household located in the centre (omitted category)	0.27	0	1
North	Household located in north (=1)	0.37	0	1
Household size	Household size in adult equivalents	3.7	0.6	16.9
Quarter 1	Interview in the first quarter of the year	0.34	0	1
Quarter 2	Interview in the first quarter of the year	0.16	0	1
Quarter 3	Interview in the first quarter of the year	0.26	0	1
Quarter 4	Interview in the first quarter of the year	0.24	0	1
Gender (head)	Gender of household head (female=1)	0.26	0	1
Age (head)	Age of household head	42.2	16	99
Education 1 (head)	No education - lower primary (EP1) not completed	0.47	0	1
Education 2	Lower primary (EP1) completed	0.25	0	1
Education 3	Upper primary (EP2) completed	0.14	0	1
Education 4	Secondary or higher education completed (base)	0.13	0	1
Woman (EP1)	Dummy for woman with at least EP1 education residing in the household (EP1=1)	0.44	0	1
Variables used as instruments				
Bike	Household has at least one bike (=1)	0.20	0	1
Safe water	Household has access to safe water (=1)	0.58	0	1

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

Table 3. Monte Carlo results

Panel A. Parameters		**Endogeneity corrected**		**No correction for endogeneity**
Parameter[a)]	**True Value**	**Mean**	**MSE**	**Mean**
β_1	-0.1	-0.098	0.004	0.299
β_2	0.1	0.100	0.007	0.499
β_3	0.2	0.198	0.005	0.599
α_1	0.3	0.300	0.002	0.310
γ_{11}	0.1	0.096	0.016	-0.157
$E\beta_1$	0.67	0.66	0.058	2.04
$E\beta_2$	1.59	1.66	1.195	5.53
$E\beta_3$	1.66	1.69	0.124	3.13

Panel B. Rejection rates (test of nominal size 0.05)

		Standard errors obtained by[b)]**:**			
Parameter	**True (theoretical)**	**Cluster bootstrap**	**IID bootstrap**	**Clustered SE**	**Robust SE**
β_1	0.05	0.046	0.162	0.091	0.245
β_2	0.05	0.046	0.095	0.066	0.125
β_3	0.05	0.056	0.152	0.092	0.220

Source: Authors calculation based on Monte Carlo simulation described in the main text.

[a)] The parameters β_1,β_2,β_3,α_1, γ_{11} refer to equation (1) in the main text, $E\beta_1$, $E\beta_2$,$E\beta_3$ refer to the marginal effect of total expenditure on the observed share given in (5).

[b)] Cluster bootstrap standard errors are obtained by bootstrapping the standard errors sampling clusters of observations (200 replications). IID bootstrap refers to standard errors generated by bootstrapping individual observations (200 repications). Clustered SEs are generated by taking specifying clustered standard errors in the last iteration of the ILLE estimator. Similarly, Robust SEs come from specifying the robust estimator in the last iteration of the ILLE estimator.

Table 4. First stage regression of total food expenditure on instruments.

Dependent variable: total food expenditure

Variables (first stage probit)	Coefficient	SE[a)]	Variables (logarithmic prices)	Coefficient	SE[a)]
South	0.029***	0.064	Vegetables	0.205***	0.068
North	0.245**	0.068	Maize flour	0.433***	0.086
Household size	0.162***	0.057	Fish	0.039	0.066
Quarter 2	0.131	0.010	Bread	0.042	0.084
Quarter 3	-0.243***	0.077	Rice	0.271**	0.135
Quarter 4	-0.166***	0.076	Meat	0.129**	0.051
Gender (head)	-0.116**	0.051	Oil & fats	-0.026	0.074
Age (head)	-0.002**	0.001	Fruits	0.027	0.039
Education 1 (head)	-0.264***	0.063	Sugar	0.153	0.080
Education 2	-0.262***	0.055	Beans	-0.094	0.073
Education 3	-0.209***	0.059	Other staples	0.087**	0.042
Woman (EP1)	0.202***	0.032	Other foods	-0.029	0.035
Instruments					
Bike	0.176***	0.042			
Safe water	0.110**	0.047			
Observations	3938 (335 clusters)		R^2	0.46	

Cluster robust F-test for instrument relevance: F(2,334) = 9.41 with p-value equal to 0.0001.

,* denote significance at 5 and 1 percent, respectively.

a) Standard errors are obtained by bootstrapping clusters of observations as explained in the main text.

Table 5. Test for exogeneity and validity of instruments.

	Test for exogeneity		Test for validity
Equation	**Coefficient residual**	**SE**[a)]	**p-value of Chi2(2)-test**[b)]
Vegetables	-0.051**	0.020	0.78
Maize flour	0.125**	0.061	0.13
Fish	-0.062**	0.027	0.92
Bread	-0.035*	0.019	0.99
Rice	-0.034	0.038	..
Meat	0.093*	0.056	0.00**
Oil & fats	0.001	0.010	..
Fruits	-0.025*	0.014	0.18
Sugar	-0.016	0.018	..
Beans	-0.037**	0.016	0.87
Other staples	0.010	0.045	..
Other foods	0.067**	0.032	0.99
Observations	3938 (335 clusters)		

*,** denote significance at 10 and 5 percent, respectively.

a) Standard errors are obtained by bootstrapping clusters of observations as explained in the main text.

b) p-values refer to a Chi2(2)-test of the null hypothesis that instruments have no explanatory power in a regression of systems residuals on all instruments (see Blundell and Robin (1999)).

Table 6. Marginal effect of demographic and location variables

	Q2	Q3	Q4	HH size	South	North	Gender Head	HH Age	Education (None)	Education (EP1)	Education (EP2)	Female Primary
Vegetables	0.01	-0.00	-0.00	-0.00	0.08**	-0.06**	0.01**	0.00	-0.00	-0.00	-0.00	-0.00
Maize flour	-0.02	-0.01	-0.01	0.02**	-0.11**	-0.10**	-0.00	-0.00	0.11**	0.08**	0.05**	-0.03**
Fish	-0.01	-0.04**	-0.03	-0.00	-0.04**	0.09**	-0.02**	-0.00	-0.00	-0.00	0.01	-0.00
Bread	-0.00	-0.00	-0.00	-0.00	0.04**	-0.01	-0.00	-0.00	-0.04**	-0.03**	-0.02**	0.02**
Rice	-0.01	0.02	0.01	-0.00	-0.02	-0.00	-0.00	0.00	0.03*	0.02	-0.00	-0.01
Meat	0.01	0.01	0.01	-0.00	0.01	-0.02**	-0.00	-0.00	-0.03**	-0.02**	-0.02*	0.02**
Oil & fats	0.01**	0.01*	0.02**	-0.00	-0.02**	-0.02**	-0.00	-0.00**	-0.00	-0.00	0.01	-0.00
Fruits	-0.00	-0.04**	-0.03**	-0.00	0.09**	0.03**	0.01**	0.00	0.01	0.01	0.01	-0.00
Sugar	-0.00	-0.00	-0.00	-0.00	-0.01	0.01	-0.00	0.00	0.01*	0.01*	0.01*	-0.00
Beans	0.02**	-0.00	-0.00	-0.00	-0.01**	-0.02**	-0.00	-0.00	0.01	0.01*	0.01	-0.00
Other staples	-0.03*	-0.04	-0.02	-0.00	0.01	0.14**	0.02	0.00	-0.01	-0.01	-0.01	-0.01
Other foods	0.01	0.05**	0.02**	-0.00	-0.00	-0.01	-0.00	0.00	-0.04**	-0.03**	-0.03**	-0.00

*,** denote significance at 10 and 5 percent, respectively.

Note: Entries show the marginal effect on the observed expenditure share of demographic and location variables relative to the reference household: a household interviewed in quarter 1, with mean household size, located in the central part of the country, with mean age of household head and secondary or higher level of education and without female household members, who have completed lower primary education (EP1).

Table 7. Estimated uncompensated own price and expenditure elasticities for a reference household.[a)]

	Own price elasticities		Expenditure elasticities	
	Elasticity	Standard error	Elasticity	Standard error
Vegetables	-0.91**	0.082	0.95**	0.149
Maize flour	0.36	0.296	0.77**	0.286
Fish	-1.13**	0.086	1.13**	0.148
Bread	-1.05**	0.163	1.08**	0.194
Rice	-1.02**	0.264	1.24**	0.187
Meat	-0.75**	0.124	1.34**	0.295
Oil & fats	-1.13**	0.126	0.96**	0.181
Fruits	-1.12**	0.064	0.88**	0.146
Sugar	-1.28**	0.172	1.31**	0.257
Beans	-1.00**	0.137	1.08**	0.193
Other staples	-0.69**	0.142	0.86**	0.254
Other foods	-1.22**	0.129	0.80**	0.398

[a)] The reference household is defined as: a household interviewed in quarter 1, with mean household size, located in the central part of the country, with mean age of household head and secondary or higher level of education and without female household members, who have completed lower primary education (EP1).Reference household: 5 household members, located in central Mozambique.

Notes: The table show percentage points change in demand for the row food group when the price of the food group or total expenditure on food changes by 1 percent.

*, ** denotes significance at 10 and 5 percent level.

All standard errors are calculated by a clustered boot strap as explained in the main text.

Table 8. The effect of endogeneity on own price and expenditure elasticities

Equation	Expenditure			Own price		
	Elasticity w/o endogeneity	Difference	SEs of difference	Elasticity w/o endogeneity	Difference	SEs of difference
Vegetables	0.60	0.36***	0.13	-0.86	-0.05	0.10
Maize flour	1.26	-0.49*	0.28	0.06	0.31	0.32
Fish	0.83	0.31**	0.14	-1.07	-0.06	0.10
Bread	0.74	0.33*	0.18	-1.09	0.04	0.18
Rice	1.09	0.15	0.17	-0.92	-0.09	0.30
Meat	1.75	-0.41	0.25	-0.82	0.07	0.97
Oil & fats	0.97	-0.01	0.16	-1.13	0.00	0.14
Fruits	0.62	0.25*	0.14	-1.12	0.00	0.08
Sugar	1.11	0.20	0.22	-1.28	0.00	0.20
Beans	0.65	0.43**	0.19	-1.03	0.03	0.17
Other staples	0.91	-0.05	0.25	-0.70	0.01	0.19
Other foods	1.55	-0.75**	0.36	-1.20	-0.02	0.48

*,**,*** denote significance at 10, 5 and 1 percent level, respectively.
All standard errors are calculated by a clustered bootstrap as explained in the main text.