MICRO-LEVEL ESTIMATION OF POVERTY AND INEQUALITY

By Chris Elbers, Jean O. Lanjouw, and Peter Lanjouw[1]

1. Introduction

Recent theoretical advances have brought income and wealth distributions back into a prominent position in growth and development theories, and as determinants of specific socio-economic outcomes, such as health or levels of violence. Empirical investigation of the importance of these relationships, however, has been held back by the lack of sufficiently detailed high quality data on distributions. Household surveys that include reasonable measures of income or consumption can be used to calculate distributional measures but at low levels of aggregation these samples are rarely representative or of sufficient size to yield statistically reliable estimates. At the same time, census (or other large sample) data of sufficient size to allow disaggregation either have no information about income or consumption, or measure these variables poorly. This note outlines a statistical procedure to combine these types of data to take advantage of the detail in household sample surveys and the comprehensive coverage of a census. It extends the literature on small area statistics (Ghosh and Rao (1994), Rao (1999)) by developing estimators of population parameters which are non-linear functions of the underlying variable

of interest (here unit level consumption), and by deriving them from the full unit level distribution of that variable.

In examples using Ecuadorian data, our estimates have levels of precision comparable to those of commonly used survey based welfare estimates - but for populations as small as 15,000 households, a 'town'. This is an enormous improvement over survey based estimates, which are typically only consistent for areas encompassing hundreds of thousands, even millions, of households. Experience using the method in South Africa, Brazil, Panama, Madagascar and Nicaragua suggest that Ecuador is not an unusual case (Alderman, et. al. (2002), and Elbers, Lanjouw, Lanjouw, and Leite (2002)).

2. The Basic Idea

The idea is straightforward. Let W be an indicator of poverty or inequality based on the distribution of a household-level variable of interest, y_h. Using the smaller and richer data sample, we estimate the joint distribution of y_h and a vector of covariates, x_h. By restricting the set of explanatory variables to those that can also be linked to households in the larger sample or census, this estimated distribution can be used to generate the distribution of y_h for any sub-population in the larger sample conditional on the sub-population's observed characteristics. This, in turn, allows us to generate the conditional distribution of W, in particular, its point estimate and prediction error.

3. The Consumption Model

The first concern is to develop an accurate empirical model of y_{ch}, the per capita expenditure of household h in sample cluster c. We consider a linear approximation to the conditional distribution of y_{ch},

$$(1) \quad \ln y_{ch} = E[\ln y_{ch}|x_{ch}^T] + u_{ch} = x_{ch}^T\beta + u_{ch},$$

where the vector of disturbances $u \sim \mathcal{F}(0, \Sigma)$.[2] Note that, unlike in much of econometrics, β is not intended to capture only the direct effect of x on y. Because the survey estimates will be used to impute into the census, if there is (unmodelled) variation in the parameters we would prefer to fit most closely the clusters that represent large census populations. This argues for weighting observations by population expansion factors.

To allow for a within cluster correlation in disturbances, we use the following specification:

$$u_{ch} = \eta_c + \varepsilon_{ch},$$

where η and ε are independent of each other and uncorrelated with observables, x_{ch}. Residual location effects can greatly reduce the precision of welfare estimates, so it is important to explain the variation in consumption due to location as far as possible with the choice and construction of x_{ch} variables. We see in the example below that location means of household-level variables are particularly useful. Clusters in survey data

typically correspond to enumeration areas (EA) in the population census. Thus, means can be calculated over all households in an EA and merged into the smaller sample data. Because they include far more households, location means calculated in this way give a considerably less noisy indicator than the same means taken over only the households in a survey cluster.[3]

An initial estimate of β in equation (1) is obtained from OLS or weighted least squares estimation. Denote the residuals of this regression as $\widehat{u}_{ch}$. The number of clusters in a household survey is generally too small to allow for heteroscedasticity in the cluster component of the disturbance. However, the variance of the idiosyncratic part of the disturbance, $\sigma^2_{\varepsilon,ch}$, can be given a flexible form. With consistent estimates of β, the residuals e_{ch} from the decomposition

$$\widehat{u}_{ch} = \widehat{u}_{c.} + (\widehat{u}_{ch} - \widehat{u}_{c.}) = \widehat{\eta}_c + e_{ch},$$

(where a subscript '.' indicates an average over that index) can be used to estimate the variance of ε_{ch}. We propose a logistic form,

$$(2) \quad \sigma^2(z_{ch}, \alpha, A, B) = \left[\frac{A e^{z_{ch}^T \alpha} + B}{1 + e^{z_{ch}^T \alpha}} \right].$$

The upper and lower bounds, A and B, can be estimated along with the parameter vector α using a standard pseudo maximum likelihood procedure.[4] This functional form avoids both negative and extremely high predicted variances.

In what follows we need to simulate the residual terms η and ε. Appropriate distributions can be determined from the cluster residuals $\widehat{\eta}_c$ and standardized household residuals

$$(3) \quad e^*_{ch} = \frac{e_{ch}}{\widehat{\sigma}_{\varepsilon,ch}} - \left[\frac{1}{H}\Sigma_{ch}\frac{e_{ch}}{\widehat{\sigma}_{\varepsilon,ch}}\right],$$

respectively, where H is the number of observations. The second term in e^*_{ch} adjusts for weighting at the first stage. One can avoid making any specific distributional form assumptions by drawing directly from the standardized residuals. Alternatively, percentiles of the empirical distribution of the standardized residuals can be compared to the corresponding percentiles of standardized normal, *t,* or other distributions.

The estimated variance-covariance matrix, weighted by the household expansion factors, is used to obtain GLS estimates of the parameters and their variance.[5]

4. The Welfare Estimator

Although disaggregation may be along any dimension - not necessarily geographic - for convenience we refer to our target populations as 'villages'. There are M_v households in village v and household h has m_h family members. To study the properties of our welfare estimator as a function of population size we assume that the characteristics x_h and the family size m_h of each household are drawn independently from a village-specific constant distribution function $G_v(x, m)$: the super population approach.

While the unit of observation for expenditure in these data is typically the household, we are more often interested in poverty and inequality measures based on individuals. Thus we write $W(m_v,X_v,\beta,u_v)$, where m_v is an M_v-vector of household sizes in village v, X_v is a $M_v \times k$ matrix of observable characteristics and u_v is an M_v-vector of disturbances.

Because the vector of disturbances for the target population, u_v, is unknown, we estimate the expected value of the indicator given the village households' observable characteristics and the model of expenditure. This expectation is denoted $\mu_v = E[W|m_v, X_v, \zeta_v]$, where ζ_v is the vector of model parameters, including those which describe the distribution of the disturbances. For most poverty measures W can be written as an additively separable function of household poverty rates, $w(x_h, \beta, u_h)$, and μ_v can be written

$$(4) \quad \mu_v = \frac{1}{N_v} \sum_{h \in H_v} m_h \int_{u_h} w_h(x_h, \beta, u_h) d\mathcal{F}^{vh}(u_h),$$

where H_v is the set of all households in village v, $N_v = \sum_{h \in H_v} m_h$ is the total number of individuals, and $\mathcal{F}^{vh}$ is the marginal distribution of the disturbance term of household h in village v. When W is an inequality measure, however, the contribution of one household depends on the level of well-being of other households and W is no longer separable. Then we need the more general form,

$$(5) \quad \mu_v = \int_{u_1} ... \int_{u_{M_v}} W(m_v, X_v, \beta, u_v) \, d\mathcal{F}^v(u_{M_v}, ..., u_1),$$

where $u_1 ... u_{M_v}$ are the disturbance terms for the M_v households in village v.

In constructing an estimator of μ_v we replace ζ_v with consistent estimators, $\widehat{\zeta}_v$, from the first stage expenditure regression. This yields $\widehat{\mu}_v = \mathrm{E}[W \mid m_v, X_v, \widehat{\zeta}_v]$. This expectation is often analytically intractable so simulation or numerical integration are used to obtain the estimator $\widetilde{\mu}_v$.

5. Properties and Precision of the Estimator

The difference between $\widetilde{\mu}$, our estimator of the expected value of W for the village, and the actual level may be written

$$(6) \quad W - \widetilde{\mu} = (W - \mu) + (\mu - \widehat{\mu}) + (\widehat{\mu} - \widetilde{\mu}).$$

(The index v is suppressed here and below). Thus the prediction error has three components:[6]

Idiosyncratic Error - $(W - \mu)$

The actual value of the welfare indicator for a village deviates from its expected value, μ, as a result of the realizations of the unobserved component of expenditure. When W is separable, this error is a weighted sum of household contributions:

$$(7) \quad (W - \mu) = \frac{1}{\overline{m}_M} \frac{1}{M} \sum_{h \in H_v} m_h \left[w(x_h, \beta, u_h) - \int_{u_h} w(x_h, \beta, u_h) d\mathcal{F}(u_h) \right],$$

where $\overline{m}_M = N/M$ is the mean household size among M village households. As the village population size increases, new values of x, and m are drawn from the constant

distribution function $G_v(x, m)$. To draw new error terms in accordance with the model $u_{ch} = \eta_c + \varepsilon_{ch}$ complete enumeration areas are added, independently of previous EAs. Since $\overline{m}_M$ converges in probability to $\mathrm{E}[m]$,

$$(8) \quad \sqrt{M}(\mu - W) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{I}}) \quad \text{as } M \to \infty,$$

where

$$(9) \quad \Sigma_{\mathrm{I}} = \frac{1}{(\mathrm{E}[m])^2} \mathrm{E}[m_h^2 \mathrm{Var}(w|x_h, \beta)].$$

When W is a non-separable inequality measure there usually is some pair of functions f and g, such that W may be written $W = f(\overline{y}, \overline{g})$,where $\overline{y} = \frac{1}{N} \sum_{h \in H_v} m_h y_h$ and $\overline{g} = \frac{1}{N} \sum_{h \in H_v} m_h g(y_h)$ are means of independent random variables.[7] The latter may be written

$$(10) \quad \overline{g} = \frac{1}{\overline{m}_M} \frac{1}{M} \sum_{h \in H_v} m_h g(y_h),$$

which is the ratio of means of M *iid* random variables $g_h = m_h g(y_h)$ and m_h. Assuming that the second moments of g_h exist, $\overline{g}$ converges to its expectation and is asymptotically normal. The same remark holds for $\overline{y}$. Thus, non-separable measures of welfare also converge as in (8) for some covariance matrix Σ_{I}.

The idiosyncratic component, $\mathrm{V}_{\mathrm{I}} = \Sigma_{\mathrm{I}}/M$, falls approximately proportionately in M. Said conversely, this component of the error in our estimator increases as one focuses on smaller target populations, which limits the degree of disaggregation possible.[8]

Model Error - $(\mu - \widehat{\mu})$

This is the second term in the error decomposition of equation (6). The expected welfare estimator $\widehat{\mu} = \mathrm{E}[W \mid m_v, X_v, \widehat{\zeta}_v]$ is a continuous and differentiable function of $\widehat{\zeta}$, which are consistent estimators of the parameters. Thus $\widehat{\mu}$ is a consistent estimator of μ and:

$$(11) \quad \sqrt{s}(\mu - \widehat{\mu}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{M}}) \quad \text{as } s \to \infty,$$

where s is the number of survey households used in estimation.[9] We use the delta method to calculate the variance Σ_{M}, taking advantage of the fact that μ admits of continuous first-order partial derivatives with respect to ζ. Let $\bigtriangledown = [\partial\mu \ / \partial\zeta]|_{\widehat{\zeta}}$ be a consistent estimator of the derivative vector. Then $\mathrm{V}_{\mathrm{M}} = \Sigma_{\mathrm{M}}/s \approx \bigtriangledown^T \mathrm{V}(\widehat{\zeta}) \bigtriangledown$, where $\mathrm{V}(\widehat{\zeta})$ is the asymptotic variance-covariance matrix of the first stage parameter estimators.

Because this component of the prediction error is determined by the properties of the first stage estimators, it does not increase or fall systematically as the size of the target population changes.

Computation Error - $(\widehat{\mu} - \widetilde{\mu})$

The distribution of this component of the prediction error depends on the method of computation used. When simulation is used this error has the asymptotic distribution given below in (14). It can be made as small as computational resources allow.

The computation error is uncorrelated with the model and idiosyncratic errors. There may be some correlation between the model error, caused by disturbances in the sample survey data, and the idiosyncratic error, caused by disturbances in the census, because of overlap in the samples. However, the approach described here is necessary precisely *because* the number of sampled households that are also part of the target population is very small. Thus, we can safely neglect such correlation.

6. Computation

We use Monte Carlo simulation to calculate: $\widehat{\mu}$, the expected value of the welfare measure given the first stage model of expenditure; $\mathrm{V_I}$, the variance in W due to the idiosyncratic component of household expenditures; and the gradient vector $\bigtriangledown = [\partial\mu/\partial\zeta]|_{\widehat{\zeta}}$.

Let the vector $\widehat{u}^r$ be the r^{th} simulated disturbance vector. Treated parametrically, $\widehat{u}^r$ is constructed by taking a random draw from an M_v-variate standardized distribution and pre-multiplying this vector by a matrix T, defined such that $TT^T = \widehat{\Sigma}$. Treated semi-parametrically, $\widehat{u}^r$ is drawn from the residuals with an adjustment for heteroscedasticity. We consider two approaches. First, a location effect, $\widehat{\eta}_c^r$, is drawn randomly, and with replacement, from the set of all sample $\widehat{\eta}_c$. Then an idiosyncratic component, e_{ch}^{*r}, is drawn for each household κ with replacement from the set of all standardized residuals and $e_{c\kappa}^r = \widehat{\sigma}_{\varepsilon,c\kappa}(e_{ch}^{*r})$. The second approach differs in that this component is drawn only

from the standardized residuals e^*_{ch} that correspond to the cluster from which household κ's location effect was derived. Although $\widehat{\eta}_c$ and e_{ch} are uncorrelated, the second approach allows for non-linear relationships between location and household unobservables.

With each vector of simulated disturbances we construct a value for the indicator, $\widehat{W}_r = W(m, \widehat{t}, \widehat{u}^r)$, where $\widehat{t} = \mathrm{X}\widehat{\beta}$, the predicted part of log per-capita expenditure. The simulated expected value for the indicator is the mean over R replications,

$$(12) \quad \widetilde{\mu} = \frac{1}{R}\sum_{r=1}^{R}\widehat{W}_r.$$

The variance of W around its expected value μ due to the idiosyncratic component of expenditures can be estimated in a straightforward manner using the same simulated values,

$$(13) \quad \widetilde{\mathrm{V}}_{\mathrm{I}} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{W}_r - \widetilde{\mu})^2.$$

Simulated numerical gradient estimators are constructed as follows: We make a positive perturbation to a parameter estimate, say $\widehat{\beta}_k$, by adding $\delta|\widehat{\beta}_k|$, and then calculate $\widehat{t}^+$, followed by $\widehat{W}^+_r = W(m, \widehat{t}^+, \widehat{u}^r)$, and $\widetilde{\mu}^+$. A negative perturbation of the same size is used to obtain $\widetilde{\mu}^-$. The simulated central distance estimator of the derivative $\partial\mu/\partial\beta_k|_{\widehat{\zeta}}$ is $(\widetilde{\mu}^+ - \widetilde{\mu}^-)/(2\delta|\widehat{\beta}_k|)$. As we use the same simulation draws in the calculation of $\widetilde{\mu}$, $\widetilde{\mu}^+$and $\widetilde{\mu}^-$, these gradient estimators are consistent as long as δ is specified to fall sufficiently rapidly as $R \to \infty$ (Pakes and Pollard (1989)). Having thus derived an estimate of the

gradient vector $\bigtriangledown = [\partial\mu/\partial\zeta]|_{\widehat{\zeta}}$, we can calculate $\widetilde{\mathrm{V}}_{\mathrm{M}} = \bigtriangledown^T \mathrm{V}(\widehat{\zeta})\bigtriangledown$.

Because $\widetilde{\mu}$ is a sample mean of R independent random draws from the distribution of $(W \mid m,\widehat{t},\widehat{\Sigma})$, the central limit theorem implies that

$$(14) \quad \sqrt{R}(\widetilde{\mu} - \widehat{\mu}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{C}}) \quad \text{as } R \to \infty,$$

where $\Sigma_{\mathrm{C}} = \mathrm{Var}(W|m,\widehat{t},\widehat{\Sigma})$.

When the decomposition of the prediction error into its component parts is not important, a far more efficient computational strategy is available. Write

$$\ln y_{ch} = x_{ch}^T\beta + \eta_c(\zeta) + \varepsilon_{ch}(\zeta),$$

where we have stressed that the distribution of η and ε depend on the parameter vector ζ. By simulating ζ from the sampling distribution of $\widehat{\zeta}$, and $\{\eta_c^r\}$ and $\{\varepsilon_{ch}^r\}$ conditional on the simulated value ζ^r, we obtain simulated values $\{y_{ch}^r\}$, consistent with the model's distributional characteristics, from which welfare estimates W^r can be derived (Mackay (1998)). Estimates of expected welfare, μ, and its variance are calculated as in equations (12) and (13). Drawing from the sampling distribution of the parameters replaces the delta method as a way to incorporate model error into the total prediction error. Equation (13) now gives a sum of the variance components $\widetilde{\mathrm{V}}_{\mathrm{I}} + \widetilde{\mathrm{V}}_{\mathrm{M}}$, while Σ_{C} in equation (14) becomes $\Sigma_{\mathrm{C}} = \mathrm{Var}(W|m, X, \widehat{\zeta},\mathrm{V}(\widehat{\zeta}))$.

7. Results

We apply the approach using household per capita expenditure as our measure of well-being, y_h, but others could be used, such as assets, income, or health status. Our smaller detailed sample is the 1994 Ecuadorian *Encuesta Sobre Las Condiciones de Vida,* a household survey following the general format of a World Bank Living Standards Measurement Survey. It is stratified by 8 regions and is representative only at that level. Our larger sample is the 1990 Ecuadorian census.

Models are estimated for each stratum. Hausman tests indicate that expansion factors have a statistically significant effect on our coefficients, so we weight accordingly (see Deaton (1997)). Subsequent analysis of the resulting estimates of welfare for localities in rural Costa indicates that this choice has a substantial effect on estimated welfare rankings. (See Elbers, Lanjouw, and Lanjouw (2002) for a fuller discussion of all results.)

Most of the effect of location on consumption is captured with available explanatory variables. In the rural Costa stratum, for example, the estimated share of the location component in the total residual variance, $\widehat{\sigma}_\eta^2/\widehat{\sigma}_u^2$, falls from 14% to 5% with the inclusion of location means (but no infrastructure variables) and to just 2% with the addition of information about household access to sewage infrastructure.[10] Using the latter model, in that stratum we cannot reject the null hypothesis that location effects are jointly zero in a fixed effects specification.

Heteroscedasticity models are selected from all potential explanatory variables, their squares, cubes and interactions.[11] In all strata, chi-square tests of the null that estimated parameters are jointly zero reject homoscedasticity (with p-values < 0.001). As with weighting, subsequent analysis for rural Costa indicates that allowing this flexibility has a substantial effect on estimated welfare rankings of localities.

For some strata in Ecuador the standarized residual distribution appears to be approximately normal, even if formally rejected by tests based on skewness and kurtosis. Elsewhere, we find a $t(5)$ distribution to be the better approximation. Relaxing the distributional form restrictions on the disturbance term and taking either of the semi-parametric approaches outlined above makes very little difference in the results for our Ecuadorian example.

Simulation results for the headcount measure of poverty and the general entropy (0.5) measure of inequality are in Table I. (For other measures see Elbers, et. al. (2002).) We construct populations of increasing size from a constant distribution $G_v(x, m)$ by drawing households randomly from all census households in the rural Costa region. They are allocated in groups of 100 to pseudo enumeration areas, with '*parroquias*' of a thousand households created out of groups of ten EAs. We continue aggregating to obtain nested populations with 100 to 100,000 households. For each population, the table shows estimates of the expected value of the welfare indicator, the standard error of the prediction,

and the share of the total variance due to the idiosyncratic component. The location effect estimated at the cluster level in the survey data is applied to EAs in the census. In all cases the standard error due to computation is less than 0.001.

Looking across columns one sees how the variance of the estimator falls as the size of the target population increases. For both measures the total standard error of the prediction falls to about five percent of the point estimate with a population of just 15,000 households. At this point, the share of the total variance due to the idiosyncratic component of expenditure is already small, so there is little to gain from moving to higher levels of aggregation. The table also shows that estimates for populations of 100 have large errors Clearly it would be ill advised to use this approach to determine the poverty of yet smaller groups or single households.

Most users of welfare indicators rely, by necessity, on sample survey based estimates. Table II demonstrates how much is gained by combining data sources. The second column gives the sampling errors on headcount measures estimated for each stratum using the survey data alone (taking account of sample design). There is only one estimate per region as this is the lowest level at which the sample is representative. The population of each region is in the third column. When combining census and survey data it becomes possible to disaggregate to sub-regions and estimate poverty for specific localities. Here we choose as sub-regions *parroquia*s or, in the cities of Quito and Guayaquil, *zona*s, because

our prediction errors for these administrative units are similar in magnitude to the survey based sampling error on the region level estimates. (See the median standard error among sub-regions in the fourth column.) The final column gives the median population among these sub-regions. Comparing the third and final columns it is clear that, for the same prediction error commonly encountered in sample data, one can estimate poverty using combined data for sub-populations of a hundredth the size. This becomes increasingly useful the more there is spatial variation in well-being that can be identified using this approach. Considering this question, Demombynes, et. al. (2002) find, for several countries, that most sub-region headcount estimates do differ significantly from their region's average level.

We can also answer questions about the level and heterogeneity of welfare at different levels of governmental adminstration. Decomposing inequality in rural Ecuador into between and within group components, we find that even at the level of *parroquias* 85% of overall rural inequality can still be attributed to differences within groups. Thus, as often suggested by anecdotal evidence, even within local communities there exists a considerable heterogeneity of living standards. This may affect the likelihood of political capture (Bardhan and Mookherjee (1999)), the functioning of local institutions, the feasibility of raising revenues locally, and other issues of importance in political economy and public policy. We expect that the empirical analysis of these issues will be strengthened by the

micro-level information on distribution that the method described here can offer.

Department of Economics, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, N.L.; celbers@feweb.vu.nl,

and

Department of Economics, Yale University and the Brookings Institution, 1775 Massachusetts Avenue NW, Washington, DC, 20036, U.S.A.; jlanjouw@brook.edu,

and

The World Bank, 1818 H. Street, Washington, DC, 20433, U.S.A.; planjouw@worldbank.org.

REFERENCES

Alderman, H., M. Babita, G. Demombynes, N. Makhatha, and B. Özler (2002): "How Low Can You Go?: Combining Census and Survey Data for Mapping Poverty in South Africa," *Journal of African Economics,* forthcoming.

Bardhan, P., and D. Mookherjee (1999): "Relative Capture of Local and Central Governments: An Essay in the Political Economy of Decentralization," CIDER Working Paper no. C99-109, University of California at Berkeley.

Chesher, A., and C. Schluter (2002): "Welfare Measurement and Measurement Error," *Review of Economic Studies,* forthcoming.

Deaton, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Washington, D.C.: The Johns Hopkins University Press for the World Bank.

Demombynes, G., C. Elbers, J. O. Lanjouw, P. Lanjouw, J. Mistiaen, and B. Őzler (2002): "Producing an Improved Geographic Profile of Poverty: Methodology and Evidence from Three Developing Countries," WIDER Discussion Paper no. 2002/39, The United Nations.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2000): "Welfare in Villages and Towns: Micro-Measurement of Poverty and Inequality," Tinbergen Institute Working Paper no. 2000-029/2.

____(2002): "Micro-Level Estimation of Welfare," Policy Research Department Working Paper, The World Bank, forthcoming.

Elbers, C., J. O. Lanjouw, P. Lanjouw, and P. G. Leite (2001): "Poverty and Inequality in Brazil: New Estimates from Combined PPV-PNAD Data," Unpublished Manuscript, The World Bank.

Ghosh, M., and J. N. K. Rao (1994): "Small Area Estimation: An Appraisal," *Statistical Science*, 9, 55-93.

Greene, W. H. (2000): *Econometric Analysis.* Fourth Edition. New Jersey: Prentice-Hall Inc.

Keyzer, M., and Y. Ermoliev (2000): "Reweighting Survey Observations by Monte Carlo Integration on a Census," Stichting Onderzoek Wereldvoedselvoorziening, Staff Working Paper no. 00.04, the Vrije Universiteit, Amsterdam.

Mackay, D. J. C. (1998): "Introduction to Monte Carlo Methods," in *Learning in Graphical Models; Proceedings of the NATO Advanced Study Institute,* ed. by M. I.

Jordan. Kluwer Academic Publishers Group.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica,* 57, 1027-58.

Rao, J. N. K. (1999): "Some Recent Advances in Model-Based Small Area Estimation," *Survey Methodology*, 25, 175-86.

Tarozzi, A. (2001): "Estimating Comparable Poverty Counts from Incomparable Surveys: Measuring Poverty in India," Unpublished Manuscript, Princeton University.

FOOTNOTES

[1] We are very grateful to Ecuador's Instituto Nacional de Estadistica y Censo (INEC) for making its 1990 unit-record census data available to us. Much of this research was done while the authors were at the Vrije Universiteit, Amsterdam, and we appreciate the hospitality and input from colleagues there. We also thank Don Andrews, François Bourguignon, Andrew Chesher, Denis Cogneau, Angus Deaton, Jean-Yves Duclos, Francisco Ferreira, Jesko Hentschel, Michiel Keyzer, Steven Ludlow, Berk Őzler, Giovanna Prennushi, Martin Ravallion, Piet Rietveld, John Rust and Chris Udry for comments and useful discussions, as well as seminar participants at the Vrije Universiteit, ENRA (Paris), U.C. Berkeley, Georgetown University, the World Bank and the Brookings Institution. Financial support was received from the Bank Netherlands Partnership Program. However, the views presented here should not be taken to reflect those of the World Bank or any of its affiliates. All errors are our own.

[2] One could consider estimating $\mathrm{E}(y|x)$ or the conditional density $\mathrm{p}(y|x)$ non-parametrically. In estimating expenditure for each household in the populations of interest (perhaps totalling millions) conditioning on, say, thirty observed characteristics, a major difficulty is to find a method of weighting that lowers the computational burden. See Keyzer and Ermoliev (2000) and Tarozzi (2001) for examples and discussion.

[3] Other sources of information could be merged with both census and survey datasets

to explain location effects as needed. Geographic information system databases, for example, allow a multitude of environmental and community characteristics to be geographically defined both comprehensively and with great precision.

[4] An estimate of the variance of the estimators can be derived from the information matrix and used to construct a Wald test for homoscedasticity (Greene (2000), Section 12.5.3). Allowing the bounds to be freely estimated generates a standardized distribution for predicted disturbances which is well behaved in our experience. This is particularly important when using the standardized residuals directly in a semi-parametric approach to simulation (see Section 6 below.) However, we have also found that imposing a minimum bound of zero and a maximum bound $A^* = (1.05)\max\{e_{ch}^2\}$ yields similar estimates of the parameters α . These restrictions allow one to estimate the simpler form $\ln\left[\frac{e_{ch}^2}{A^* - e_{ch}^2}\right] = z_{ch}^T\alpha + r_{ch}$. Use of this form would be a practical approach for initial model selection.

[5] In our experience, model estimates have been very robust to estimation strategy, with weighted GLS estimates not significantly different from the results of OLS or quantile regressions weighted by expansion factors.

[6] Our target is the level of welfare that could be calculated if we were fortunate enough to have *observations* on expenditure for all households in a population. Clearly because expenditures are measured with error this may differ from a measure based on true expenditures. See Chesher and Schluter (2002) for methods to estimate the sensitivity

of welfare measures to mismeasurement in y.

[7] The Gini coefficient is an exception but it can be handled effectively with a separable approximation. See Elbers, et. al. (2000).

[8] The above discussion concerns the asymptotic properties of the welfare estimator, in particular consistency. In practice we simulate the idiosyncratic variance for an actual sub-population rather than calculate the asymptotic variance.

[9] Although $\widehat{\mu}$ is a consistent estimator, it is biased. Our own experiments and analysis by Saul Morris (IFPRI) for Honduras indicate that the degree of bias is extremely small. We thank him for his communication on this point. Below we suggest using simulation to integrate over the model parameter estimates, $\widehat{\zeta}$, which yields an unbiased estimator.

[10] To choose which variable means to include we estimate the model with only household-level variables. We then estimate residual cluster effects, and regress them on variable means to determine those that best identify the effect of location. We limit the chosen number so as to avoid over-fitting. The variance, σ_η^2, of the remaining (weighted) cluster random effect is estimated non-parametrically, allowing for heteroscedasticity in ε_{ch}. This is a straightforward application of random effect modelling (e.g., Greene (2000), Section 14.4.2). An alternative approach based on moment conditions gives similar results.

[11] In the results presented here, the constrained logistic model in footnote 3 was used to model heteroscedasticity.

TABLE I

SIMULATION RESULTS

Measure	**Estimated Values**	**Number of Households**			
		100	**1,000**	**15,000**	**100,000**
Headcount	μ	0.46	0.50	0.51	0.51
	Total Standard Error	0.067	0.039	0.024	0.024
	V_I / Total Variance	0.75	0.24	0.04	0.02
General Entropy (0.5)	μ	0.26	0.28	0.28	0.28
	Total Standard Error	0.044	0.020	0.012	0.011
	V_I / Total Variance.	0.91	0.56	0.11	0.03

TABLE II

IMPROVEMENT USING COMBINED DATA

	Sample Data Only (region)		**Combined Data (sub-regions)**	
	(2)	(3)	(4)	(5)
Region	**S.E. of Estimate**	**Population** (1000s)	**S.E. of Estimate** Median	**Population** Median (1000s)
Rural Sierra	.027	2,509	.038	3.3
Rural Costa	.042	1,985	.046	4.6
Rural Oriente	.054	298	.043	1.2
Urban Sierra	.026	1,139	.026	10.0
Urban Costa	.030	1,895	.031	11.0
Urban Oriente	.050	55	.027	8.0
Quito	.033	1,193	.048	5.8
Guayaquil	.027	1,718	.039	6.5

TYPESCRIPT SYMBOLS LIST

x_h (ex)

$\times$ (multiplication)

β (beta)

$\mathcal{F}$ (calligraphic ef)

Σ (capital sigma)

η (eta)

ε (epsilon)

σ (sigma)

α (alpha)

ζ (zeta)

μ (mu)

$\int$ (integral)

$\sum$ (summation)

v (nu)

$\in$ (element of)

∞ (infinity)

$\rightarrow$ (arrow right)

κ (kappa)

δ (delta)

$\bigtriangledown$ (nabla)