



No. 47-2011

Jing Dai, Stefan Sperlich and Walter Zucchini

**Estimating and Predicting Household Expenditures and
Income Distributions**

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers/index_html%28magks%29

Coordination: Bernd Hayo • Philipps-University Marburg
Faculty of Business Administration and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Estimating and Predicting Household Expenditures and Income Distributions

Jing Dai, **Stefan Sperlich,** **Walter Zucchini**
Universität Kassel, Université de Genève, Universität Göttingen,

November 25, 2011

Abstract

A reliable prediction of unconditional welfare distributions, like income or consumption, is essential for welfare analysis, and in particular for inequality, poverty or development studies. Where observations of expenditures or income are missing, the mean prediction based on available covariates is not just a poor estimator of the unconditional distribution; it fails to predict the required information about tails and quantiles. A new estimation method is introduced which can be combined with any mean prediction model. It is used to calculate the income distribution of a survey based on subsample information, to estimate the unconditional income distribution for the non-responding households, and to predict the household expenditures of a future panel wave. It allows for imputing welfare distributions for a census from a survey or for synthetic populations under specific scenarios. Further inference is straight-forward, including prediction of Lorenz curves, indexes like the Gini, or distribution quantiles, including confidence intervals. ¹

Keywords: household expenditures, income distribution, poverty mapping, project evaluation, data matching.

¹Affiliation of Dr. Jing Dai: Universität Kassel, Fachbereich Wirtschaftswissenschaften, Nora-Platiel-Str. 5, D-34109 Kassel, tel: +49 (0) 561 804 - 7505, fax: +49 (0) 561 - 804 2501, jing.dai@uni-kassel.de; affiliation of Prof. Walter Zucchini: Georg-August Universität Göttingen, Institut für Statistik und Ökonometrie, Platz der Göttinger Sieben 5, D-37073 Göttingen, tel: +49 (0) 551 - 39 7286, fax: +49 (0) 551 - 39 7279, walter.zucchini@wi-wiss.uni-goettingen.de; and affiliation of corresponding author Prof. Stefan Sperlich: Université de Genève, Département des sciences économiques, Bd du Pont d'Arve 40, CH-1204 Genève, tel: +41 (0) 22 379 8223, fax: +41 (0) 22 379 8299, stefan.sperlich@unige.ch

1 Introduction: The problem

For any empirical study on inequality, welfare economics, discrimination or poverty, reliable information about welfare distributions, such as income or expenditure, is fundamental. Especially in underdeveloped areas, having such information is of critical importance for governmental as well as nongovernmental organizations, including research institutions. Decision makers rely on distribution estimates or predictions to assess and monitor social security systems, allocate resources, transfers, etc. Furthermore, those estimations or predictions enable researchers to carry out poverty mapping, to analyze the relationship of poverty or inequality and human development indicators, and to study the pro-poor growth or related issues. For the combination of poverty mapping and policy implications see the recent compendium of Hyman et al. (2005).

There exists many initiatives, e.g. the OPPG², of national and international institutions – the World Bank being probably the most involved – for which this information is an imperative. Ravallion (2001) highlighted the fact that more attention should be given to the micro level and therefore take into account the micro distributions rather than just on the means. However, since the collection of good quality expenditure or income data requires a lot of time and effort, or because the complete information is simply not achievable, researchers and policymakers have a strong interest in approaches which provide good estimates and predictions, allow for inferences, scenario simulations, and comparison over time and space. A most simple question would be, how to study different poverty levels when the information about preferred consumption expenditure or income measures is absent? Certainly, as we propose an econometric method, we assume that some useful information is available. More specifically, we assume the availability of some informative covariates, say x , related to the variables of interest, say y . Additionally, we are provided with a sample containing both x and y , maybe collected from a different population or at a different point in time. We further assume that the conditional distribution of y given x for the population of interest - this can even be a fictitious one - is similar either to the one of our data at hand or approximately known. Our use of the term ‘similar’ will be specified along the presentation of our method.

The literature on the issue of poverty and inequality measurement, as well as policy implications, is abundant. Under www.pep-net.org/programs/pmma, there are more than 1000 “recommended readings” on this topic. Clearly, the huge amount of literature renders a comprehensive discussion impossible and we focus only on the most related contributions. They are still many when concentrating just on imputing income and expenditure. But if it comes down to the estimation or prediction of welfare distributions, we only find a few procedures but always proposed within a specific context, typically when studying poverty levels. In that spirit, Paulin and Ferraro (1994) was an early work on imputing income, Filmer and

²Operationalising Pro-Poor Growth

Prichett (2001), as well as Sahn and Stifel (2003), did welfare studies when expenditure data were missing. Hentschel et al. (2000) imputed the likelihood to be poor from a survey to impute a poverty map with census data.

Most procedures to predict the required micro distribution (thinking of poverty and inequality indexes) or at least some of its parameters, are based on data matching. A mean regression is calculated from a set of available information to then predict the non-reported income or expenditures for the group of interest. For a review on the prediction of expenditure in the context of poverty and inequality analysis see Abeyasekera and Ward (2002). The population of interest can be all individuals for which income or expenditure are missing within the same survey, or a different survey for which this information is not available, maybe a census, a future or past panel wave (or cohort), or simply a fictitious population in case of scenarios. An appropriate data matching technique would therefore allow for comparisons of income, expenditure and related factors over time and space (cf. Sahn and Stifel, 2000). The development of techniques to interpolate from a survey to a more general data set has been well summarized by Davis (2003). Unfortunately, the estimated conditional distribution can be quite different from the required unconditional one, i.e. the distribution of the effective income or expenditure. In particular, the conditional distribution has a substantially smaller spread, and therefore must not be used for welfare, inequality or poverty analysis. In fact, if measures of inequality, poverty or vulnerability are of interest, one has to correct for the shrinkage of the predicted values toward the mean in order to really capture the tails of the distribution. Thus, no matter what kind of regression models or survey types being applied, the problem is that one gets only conditional values, which have a much smaller spread, may differ in the shape, etc. resulting in high misclassification errors. Hentschel et al. (2000) applied numerical “stretching” of the conditional distribution to fit some given percentiles calculated from a sample with complete information.

A more practical and often used remedy, is to add random errors, normally distributed with a given constant variance. As the mean prediction is based on a regression with data sets where full information is available, this variance might be estimated from that data too. In statistical terms, one does a kind of wild bootstrap under homoscedasticity to simulate the welfare distribution for the population of interest. This method, though quite popular, entails several drawbacks like no analytic predictor, random results, no further valid inference, etc. In the context of small area statistics (for a general idea see Ghosh and Rao, 1994) Birkin and Clarke (1989) were probably the first who introduced an approach to simulate micro income distributions. Elbers et al. (2003) proposed a simulation method – though they call it estimation – based on small area modeling to track poverty and inequality issues on the macro-level from micro data, the so-called ELL or World Bank method. Note that these small area based methods are designed to approximate macro-parameters, not the micro-level distribution itself. Moreover, while in statistics a lot of effort is spent on deriving

methods for doing valid inference, applied econometricians mostly rely here on intuitively justified simulations. More recently, Tarozzi and Deaton (2009) and Demombynes et al. (2007) discussed those approaches quite critically. Note finally that Zeller et al. (2004), Zeller et al. (2005) and Azzarri, Carletto, Davis and Zezza (2006) compared various models and methods for poverty assessment.

In this article a method is introduced to reveal the whole unconditional distribution. It is based on distribution theory rather than on simulation methods such as the kind of wild bootstrap or Bayesian small area predictors discussed above. This constitutes a quite different and new way to estimate, or predict, the distribution of interest, although it is based on the same amount of information. We transform the mean prediction by applying an integration-based method to assess the unconditional distribution. This provides analytic calculations instead of simulations; it allows for further valid inference and for a more realistic simulation of scenarios (see Gasparini et al. (2003) for currently used methods). Another advantage is that the modeling of income and expenditure distributions on which our approach is based is a well-studied field in statistics; see the recent compendium of Chotikapanich (2008) or Atkinson and Bourguignon (2000). Note that the new method is applicable independently of the mean regression or model; it can be used for mixed effects (multi level) models as e.g. in the context of small area estimation, nonparametric statistics, any latent variable model (e.g. Tobit regression), simultaneous equation systems, IV methods, etc. It is evident how to extend this method to any other context.

Before introducing the main idea, we should mention two rather different approaches which, in some circumstances, can provide more helpful solutions. First, in the case where only very few but specific quantiles of a particular distribution are of interest, it is recommendable to just stick to these quantiles, i.e. scalars instead of functions, see Koenker (2005) for a recent compendium. The particular interest is directed to quantile regression of conditional distributions and its marginals, see Firpo et al. (2009), cf. also Rothe (2009). While these methods look quite promising, they are not constructed for revealing the whole distribution. Also, they are theoretical rather than practical contributions. Especially the nonparametric approaches are only recommendable for the one dimensional case and can be computationally quite cumbersome. This is in contrast to the second approach we would like to mention. If it is limited to that of imputing some missing values in a large sample, survey or census, we refer to the so-called imputation methods first introduced by Rubin; for a compendium see Littel and Rubin (2002). The provision of the associate software and the description of modules and commands is abundant; see Horton and Lipsitz (2001), Royston (2004), or Su et al. (2010). Note that this method was explicitly developed for the imputation of missings in a survey to subsequently and conveniently carry out statistical data analysis. The algorithms work like “black boxes”; they are not considered as estimators or predictors of (marginal) distributions. Simulations, not shown in this article, revealed that the method

introduced in this article outperforms the publicly available algorithms in this respect - not to mention the problems a black box entails for subsequent inference.

The rest of the paper is organized in the following way: In the next section we introduce our new methodology for estimating the marginal distribution of Y for the population of interest. In Section 3 we consider two different types of problems of estimating the income distribution accounting for possible selectivity biases. In Section 4 we use a panel wave from Indonesia to predict the expenditure distributions for four years later. In two of our applications we are provided with complete information so that we can validate our estimation results. Section 5 concludes.

2 A general methodology for predicting welfare distributions

Both, matching and prediction is based on conditioning variables. We are interested in the distribution of some quantity, say y , where in this article we are thinking of household income or expenditures. Imagine we are provided with a sample \mathcal{S}_1 containing information about y and, additionally about (possibly) related information, say x , for example demographic factors and location. The objective is to estimate the distribution of y for a data set, say \mathcal{S}_2 , where only information on x is available. This can be a different survey or census, a different wave in a panel, or even just an enlarged set containing \mathcal{S}_1 , but with missing responses y for the added records, i.e. households in our case. Alternatively, \mathcal{S}_2 could be a fictitious population with some x changed, e.g. for scenarios typical in forecasting and counterfactual exercises.

There are at least two obvious approaches we would think of; either we estimate directly the joint distribution of (x, y) then extract the marginal one of y for a given set of x (one may also think of a predefined distribution of x), or we concentrate on the conditional moments of $y|x$ which will then allow us to construct the marginal distribution of y for any given set of x . The first idea corresponds directly to the literature we discussed in the context of imputation methods and quantile estimation; the latter to the regression plus simulation methods we mentioned in the context of simulation methods and small area statistics. While for our purpose the first idea looks formally more elegant from a stochastic point of view, the second is more appealing under practical considerations. However, as we will see, depending on the set of prior assumptions, they are even identical and can be converted from one to the other. Without depreciating the former, we therefore follow in the presentation the second approach, starting with the conditional mean.

2.1 From marginalization to local n-fold mixtures

Consider a prior regression setup based on a completely observed sample $\mathcal{S}_1 = \{(y_i, x_i)\}_{i=1}^n$,

$$y_i = g(x_i) + \epsilon_i. \quad (2.1)$$

For another set \mathcal{S}_2 containing $\{x_j\}_{j=1}^m$ the y_j are missing. These data could be: a) in the same survey; b) from another survey or census; c) belong to the same panel as \mathcal{S}_1 but to a different wave; or d) describe a fictitious population. In a first step one estimates the mean prediction $E(Y|X = x) = g(x)$ along its particular model specification of $g(\cdot)$ in (2.1). We could equally well include random or fixed effects if identifiable, as is recommendable for repeated measurements, multilevel or panel models. For specific economic data, $g(\cdot)$ may be estimated via Tobit models, selection bias correction, with weights from strata sampling, etc.. One may even apply non- and semiparametric methods as we are not interested in the interpretation of any parameters in model (2.1). Actually, any consistent estimation of $g(\cdot)$ is valid, and it should be emphasized that the main objective is not identification but estimation and prediction and therefore the minimization of prediction or mean squared errors. From this point of view, even inconsistent estimators would do, especially if they provide the smallest prediction error.

As we mentioned in the introduction, the general problem is that, no matter what kind of prediction models or survey types being applied, one gets only conditional values which have a distribution with density $f(y|x)$ with a smaller spread than the unconditional distribution $f_y(y)$. For welfare analysis, measuring inequality, poverty or discrimination, the conditional distribution alone is of little help. The shrinkage of predictions toward the mean is primarily caused by the fact that the predictions do not explain all the variation in consumption expenditure (or income); therefore some of the existing solutions which do not ignore this ‘shrinkage’ effect, simply add random errors (typically normally distributed) with an appropriate variance to widen the density. The latter method is widely used, often combined with small area estimation and data mapping. However, since the results are generated by simulations and under strong assumption on the model and distribution, the resulting values depend on chance, and any further inference is not statistically justified. Moreover, as discussed in the introduction, many methods are only constructed for simulating particular percentiles, not the whole distribution.

In contrast, we introduce a direct analytic method of the unconditional distribution. Recall that the required marginal distribution $f_{y,2}(y)$ of y in \mathcal{S}_2 can be written in terms of the conditional $f_2(y|x)$ and the unknown $f_{x,2}(x)$, as

$$f_{y,k}(y) = \int f_{(y,x),k}(y, x)dx = \int f_k(y|x)f_{x,k}(x)dx, \text{ for } k = 1, 2 \quad (2.2)$$

by integrating covariates out from the joint distribution, where the k indicates the particular population. A simple numerical approximation of this integration, and to get around the estimation of $f_{x,k}$ is the sample average, i.e.

$$f_{y,k}(y) = \frac{1}{m} \sum_{j=1}^m f_k(y|x_j) + O\left(\frac{1}{m}\right), \text{ for } k = 1, 2. \quad (2.3)$$

So we obtain the required distribution by averaging over all estimated local densities. Certainly, not observing y in \mathcal{S}_2 , we cannot estimate its $f_2(y, x)$ nor $f_2(y|x)$. If it is believed that the conditional distributions are the same for \mathcal{S}_1 and \mathcal{S}_2 , one could use Firpo et al. (2009) or Rothe (2009) to derive parametric or nonparametric estimates for our context. However, in addition to the problems that occur in applying multidimensional nonparametrics, Firpo et al. (2009) found no improvement in their results when looking at some conditional quantiles; they reported several drawbacks instead.

Instead, we give up the strong assumption of having the same conditional distribution of y in both data sets, and we stick to flexible parametric modeling. Our argument is first, that for getting a good approximation of the marginal distribution $f_{y,2}$ in (2.3) it is sufficient to control for a given set of identifiable parameters, in particular the mean and variance, but optionally also the symmetry of $f_2(y|x)$. Second, coming up with a proper conditional a priori for $f_2(y|x)$ is not less justifiable than assuming it to be identical to a nonparametric $f_1(y|x)$. Finally, the estimate of the required unconditional density $\hat{f}_{y,2}(y) = \frac{1}{m} \sum_{j=1}^m \hat{f}_2(y|x_j)$ becomes a kind of a n-mixture of densities. We use here \hat{f}_2 to emphasize that the moments have been estimated before from $\mathcal{S}_1 = \{(y_i, x_i)\}_{i=1}^n$. Mixtures are known to give excellent approximations and are consistent under different sets of typically mild conditions, see for example McLachlan and Peel (2000) for a compendium, or for our context of Bayesian priors and approximations of nonparametric functions, Marin et al. (2005). Recall further that kernel density estimates with second order kernels are local n-fold mixtures. In our case, controlling for the second moment corresponds to local bandwidths in kernel density estimation, and controlling for the third moment corresponds to local kernels - as they are recommended for boundary problems - or asymmetric weighting as in the so-called knn smoothing.

2.2 Modeling, estimation and calibration

Given is a conditional distribution $f_2(y|x)$ up to some unknown parameters, which can be typically expressed in terms of its moments. We concentrate only on distributions with at most three unknown parameters, and the first three moments, namely $E[Y|X]$, $Var[Y|X]$, and $E[(Y - E[Y|X])^3|X]$. Now, the idea is relatively simple: the available data from \mathcal{S}_1 are taken to estimate the necessary moments via mean regression first of Y , then of the squared and, if necessary, also the cubed residuals. For the mean regression it is recommended to use

a model as rich as possible, but to disregard bias reducing methods which may increase the total mean squared error (as e.g., instrumental variable methods do). In our applications we will use all available information x and explore the possible gain of semiparametric models, like the additive partial linear model (APLM). This is a nontrivial extension of the linear model:

$$E[Y|X = (U, T)] = c + U'\beta + \sum_{\alpha=1}^q g_{\alpha}(T_{\alpha}), \quad (2.4)$$

Here, the explanatory variables are separated into two the vectors U and T , where typically, U denotes a vector containing all categorical, especially dummy variables, and vector $T = (T_1, \dots, T_q)$ the vector of continuous variables. The unknown functions $g_{\alpha}(\cdot)$ are estimated in a nonparametric way. Most statistical and econometric software packages offer such a flexible regression model. Where data and model allow for random effect modeling without introducing a bias which leads serious prediction errors, this can be done, too. However, this often renders subsequent statistical inference rather complicated.

In the special case where our method is used to predict income or expenditure for the missing values in the same survey or census, i.e. where $\mathcal{S}_1 \subset \mathcal{S}_2$, one has to control for a possible selection bias. There are several approaches, depending on the economic model and data availability in \mathcal{S}_1 , the Heckman (1976,1979) correction being maybe the oldest but still most popular one. Currently, there also exist different semiparametric approaches as e.g. Ahn and Powell (1993) or Rodríguez-Póo et al. (2005).

In another particular case of predicting a variable y inside a panel structure for a wave, where this information is missing, fixed or random effects models and the inclusion of trends can seriously improve the prediction quality. For panels being large in the time dimension, one can also consider varying coefficients to use time trends or business cycles for improving prediction, i.e.

$$E[Y|X = (U, T), V] = \beta_0(T) + U'\beta(T) + V(T), \quad (2.5)$$

where T can be time and some macroeconomic factors, and V are fixed or random effects, possibly depending on T , too.

Similarly one can proceed with the scedasticity function $\sigma(x)$. Certainly, in a case where homoscedasticity is credible, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \epsilon_i^2$ or its df-adjusted version would do, where $\epsilon_i = y_i - \hat{y}_i$. More recommendable is to use a smoothed version of $\sigma^2(x_i) \approx \epsilon_i^2$ (heteroscedasticity) with the index of the mean function, in case of (generalized) linear models $x'\beta$, as regressor. Our experience for monetary measures is that, in case of heteroscedasticity, a constant coefficient of variation (CoV) often does a very good job for approximating the scedasticity function when the conditional mean is already estimated. With $CoV = \frac{\sigma(x)}{E[Y|x]}$ constant, one gets an appropriate estimator for $Var[Y|x]$ from the simple regression $E[\epsilon^2|x] = c \cdot E^2[Y|x]$ or its simple extension $E[\epsilon^2|x] = c_0 + c_1 E[Y|x] + c_2 E^2[Y|x]$.

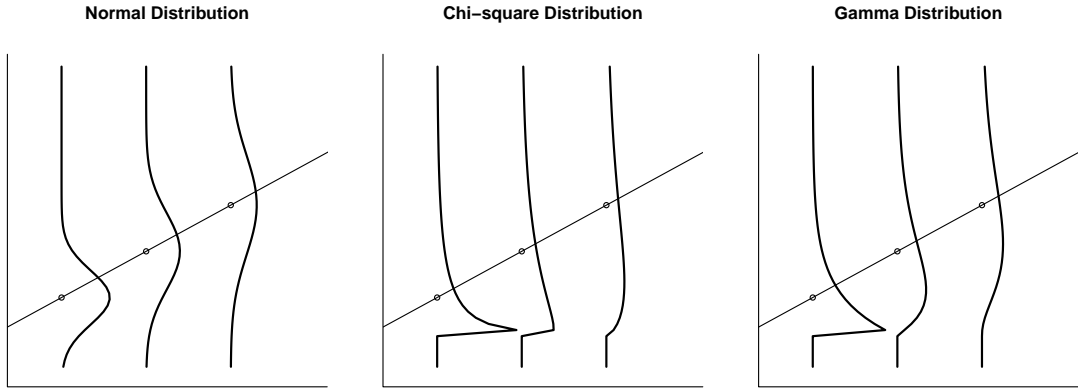


Figure 1: Examples for typical prior conditional distributions with heteroscedasticity when the mean function is a straight line.

The residual distributions of income and expenditure regressions are bounded from below and can be quite skewed for income and expenditure regressions. For log-income and log-expenditures they are only somewhat skewed for higher means. In any case it is worth considering distributions not restricted to symmetry. One can either work then with parametric families containing (at least) three parameters, or make the skewness depend on the mean–variance proportion. The latter is especially recommended if the residual distribution is bounded from below or above. In Figure 1 we give three typical examples for appropriate priors of conditional distributions where the mean is a simply a straight line. A simple example for an according parametrization is the application of the gamma distribution for f_2 , i.e. if one uses

$$\hat{f}_{y,2}(y) = \frac{1}{m} \sum_{j=1}^m \Gamma(y | \hat{k}(x_j), \hat{s}(x_j)) \quad (2.6)$$

with $E[Y|x] = k(x)s(x)$ and $Var[Y|x] = k(x)s^2(x)$, $s(x), k(x) > 0$. It is easy to see that we get $k(x) = \frac{E^2[Y|x]}{Var[Y|x]}$ and $s(x) = \sqrt{\frac{Var[Y|x]}{k(x)}}$; analogously its estimates. For homoscedasticity one obtains the restriction $s(x) = \sqrt{\frac{1}{k(x)}}$, and for heteroscedasticity with constant CoV we obtain $k(x) = k$.

When the conditional moments have been estimated and the estimation of the marginal distribution has been done via (2.3), a final calibration is still recommended as long as information about the variable of interest or its distribution is available. The most evident case is when for \mathcal{S}_2 the mean of Y is known; then the mean of the estimated distribution is adjusted accordingly.

3 Estimating the income distributions

In the first application we use a continuing longitudinal household-level data set from the Indonesia Family Life Survey (IFLS). It provides data at the individual and household level on consumption, income, health, education, housing and employment. Following Alisjahbana et al. (2003) the IFLS sample is representative for about 83% of the Indonesian population living in 13 of the 26 provinces in the country. In 1997, 2000 and 2008 the IFLS contains about 6500 to 10000 households which are partly cross section cohorts and partly a panel, also because the questionnaires changed over time. The available data also contain some sensitive information, including the household expenditures and income - though with 20% to almost 50% missing values. The consumption expenditures and income are expressed in logarithms of Rupiah.

3.1 A first exercise: sample split

The first application is an easy, artificial exercise to study the functioning of our procedure. We take the 5567 households in 2008 for which income has been recorded and split them into half, 2783 for \mathcal{S}_1 and 2784 for \mathcal{S}_2 . We will apply our procedure with different estimators and priors to finally compare the resulting predictions with the actually recorded income distribution. In our study, household income per capita is a summation of five income sources: (1) income from wage and salary in both cash and in-kind transfers; (2) income from agricultural business; (3) income from non-agricultural business; (4) household non labor income, i.e. income outside wage/salary and business e.g. estimated house rent, pension, scholarship, transfer received, etc; (5) household assets income.

Besides the classic references to the Mincer model, the data availability is a main consideration when choosing the set of explanatory variables. Human capital theory suggests that education (measured as average years of schooling) and experience of working household members (measured here as average age) are chosen as explanatory variables. Other socio-economic variables available are the share of working household members, household size, household's female labor ratio, and whether there is a farmer in the family. The latter, together with regional dummies (provinces and urban-rural location) accounts also for the considerably high discrepancy between urban and rural areas. We neglected the possibility of area-varying returns to assets or to human capital. The asset ownership variable enters the model separately by log of assets per capita and share of those assets devoted to household business activity. We end up with 22 predictor variables.

The model we have outlined at first is the simple linear one used in most poverty assessments that rely on regression methods. By assuming that income, consumption expenditures, and many other monetary welfare indicators are conditionally approximately log-normally

distributed (i.e. $\ln(y)|x$ is normally distributed), we constructed an income prediction model with log household total annual income per capita as response and our set of non-income regressors referring also to the empirical studies in Alisjahbana et al. (2003). We simply applied OLS. Today a much more flexible alternative is the additive partial linear model introduced above, cf. equation (2.4). The resulting coefficients can be seen in Table 1. For the additive partial linear model (APLM) we only give the coefficients for the parametric part without standard deviations.

	Application 1		Application 2	
	Linear Model	APLM	2-step-est.	Heckman-2-step
Constant	10.957 (.2529)			11.55 (.0650)
Average age	.0415 (.0102)		.0494	.0438 (.0001)
Average age squared	-.0006 (.0001)		-.0007	-.0007 (.0000)
Average year of schooling	.0400 (.0053)		.0358	.0310 (.0000)
Log of assets per capita	.2261 (.0122)		.2294	.2254 (.0001)
Share of asset to business	.4672 (.0808)		.5027	.5602 (.0054)
Farmer in family	-.2351 (.0489)	-.2102	-.2102	-.1331 (.0024)
Share of working hhm	1.5472 (.0977)		1.203	.6681 (.0360)
Share of female hhm	-.6385 (.1037)		-.5233	-.5826 (.0081)
HH size	-.0685 (.0099)		-.1441	-.2607 (.0016)
Located in urban area	.2871 (.0430)	.2717	.2610	.2651 (.0014)
North Sumatera	-.2365 (.0894)	-.2162	-.0747	-.0074 (.0062)
West Sumatera	-.0194 (.1102)	.0056	-.0433	-.1396 (.0094)
South Sumatera	-.1287 (.0942)	-.0738	-.0513	.0250 (.0076)
Lampung	-.4017 (.0956)	-.3702	-.3721	-.2591 (.0081)
West Java	-.2611 (.0678)	-.2206	-.2196	-.2300 (.0034)
Central Java	-.6629 (.0739)	-.6095	-.6076	-.5800 (.0041)
Yogyakarta	-.6850 (.1022)	-.6261	-.6692	-.7929 (.0089)
East Java	-.5239 (.0723)	-.4890	-.5010	-.5872 (.0042)
Bali	-.4119 (.0931)	-.4343	-.3488	-.3065 (.0069)
West Nusa Tenggara	-.5801 (.0846)	-.5296	-.5084	-.4556 (.0056)
South Kalimantan	-.0314 (.0965)	.0295	-.0234	-.0917 (.0075)
South Sulawesi	-.6058 (.1068)	-.5632	-.6022	-.6826 (.0086)
Number of observations	2783	2783	5567	5567

Table 1: Coefficients of the mean income models with standard deviations in parentheses.

As parametric prior distributions for the conditional density of $\ln(y)|x$ in \mathcal{S}_2 we tried the (log) normal distribution and, to account for some asymmetry, the gamma distribution. Then, for the second moment we compared different estimates for the scedasticity function but, for the sake of brevity, we present only results under homoscedasticity, and results under constant CoV; compare Section 2.2. The resulting estimates for the income distribution in \mathcal{S}_2 are given in Figure 2 for the linear regression model, and in Figure 3 for the additive partial linear model. The density plots for the real income distribution and the distribution of conditional

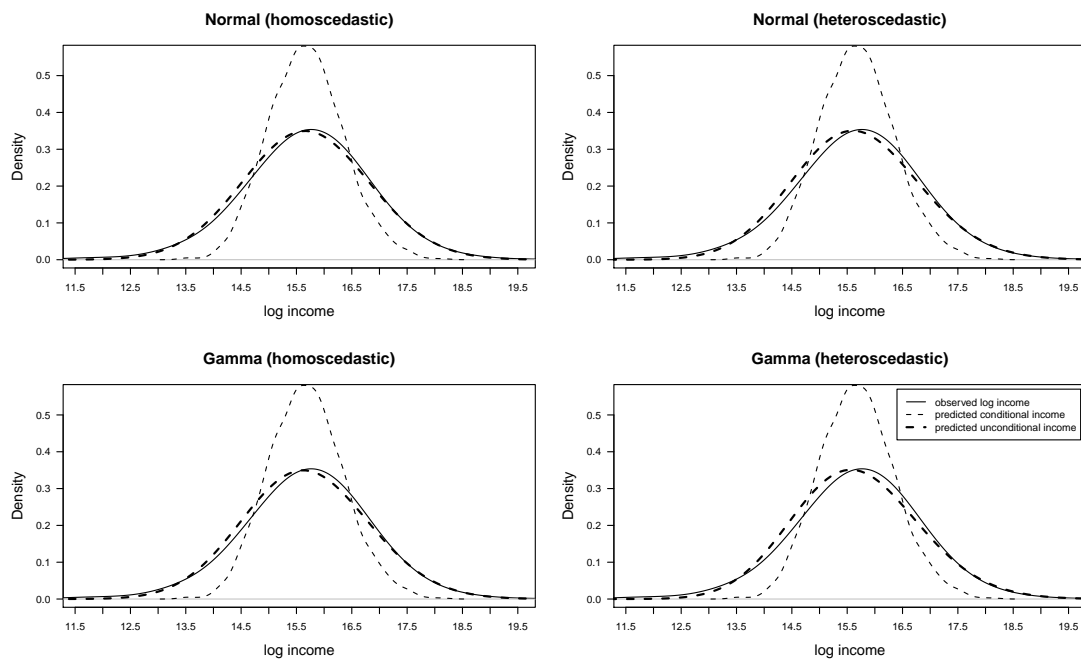


Figure 2: Probability density curves based on a linear regression mean for the conditional (grey dashed) and the unconditional (dark dashed) predicted income, compared to a kernel density estimate based on the real incomes (black line). Different modeling approaches from the upper left to the lower right.

incomes, were created by kernel methods with Gaussian kernel and two times Silvermans rule-of-thumb bandwidths, as the default still gave wiggly outcomes. What can be seen first is that there is an enormous difference between the distributions of the conditional and the unconditional log income respectively. This is not surprising given an R^2 of slightly above 30% for both regressions. Even though the APLM does slightly better, the improvement is hardly visible. The choice between homo- and heteroscedasticity, and also the choice of the prior conditional distribution, seem to have a little bit more impact than the regression model. The differences are nevertheless marginal when looking at the integrated squared error, which can only be estimated because the real income distribution has to be calculated via smoothing methods. Repeating this exercise several times, i.e. splitting the original 5567 observations into two sets and estimating one from the other shows that a representative sampling from the provinces and the urban area is responsible for the shift of the mode (in our example to the left) of the estimate. Apart from such sampling biases, the prediction methods seems to work quite well. The outcome is robust and does not depend much on our prior assumptions. Again, recall that our final estimator can be considered as an n -fold mixture. For samples \mathcal{S}_2 larger than $n = 100$ the differences due to the prior modeling diminish rapidly, except for extremely different models. In practice one does not really know which of the models (linear, APLM, homoscedastic, heteroscedastic, normal or gamma

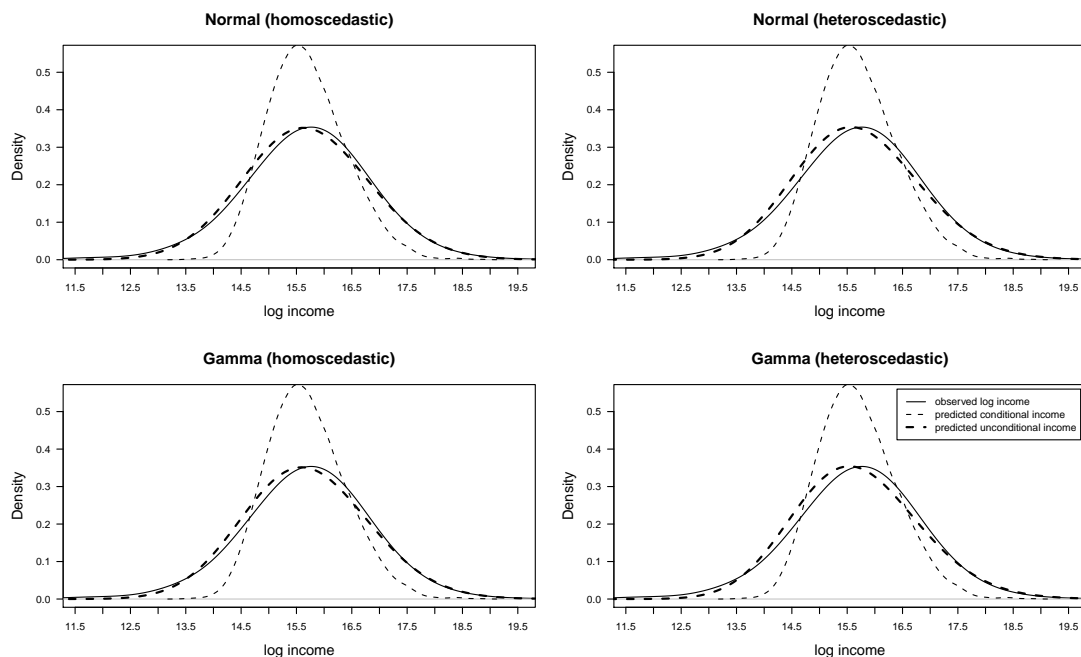


Figure 3: Probability density curves based on an additive partial linear regression mean for the conditional (grey dashed) and the unconditional (dark dashed) predicted income, compared to a kernel density estimate based on the real incomes (black line). Different modeling approaches from the upper left to the lower right.

distribution) is closest to the real data generating process, and so it is always recommendable to try more than one for such a robustness check.

3.2 Predicting the income distribution with missing values

In the first application we looked for an artificial problem that allowed us to study and illustrate the performance of the introduced method. We therefore considered an - admittedly, less interesting - situation where it is quite likely that the moment regressions and the unknown distribution in \mathcal{S}_1 and \mathcal{S}_2 are similar, i.e. come from the same population when disregarding selectivity biases.

In our second application we now turn to a problem where both data sets again come from the same population but present the outcomes of a selection that is most likely endogenous. Furthermore, we will not be able to check our results, simply due to the lack of complete information. More specific, we again take the IFLS data from 2008 where 5567 households reported their income but 4894 did not. Even though it is improbable that the same selectivity mechanism applied to almost 50% of the total survey, to assume them to be missing at random would be rather optimistic. We therefore applied a two step estimator that accounts

for the selection. The idea is as follows. We face two equations,

$$y^* = x^T \beta + u, \text{ income} \tag{3.1}$$

$$s = \mathbb{1}\{z^T \theta + \epsilon\}, \text{ reports income or not} \tag{3.2}$$

with the typical assumptions on u and ϵ . In our case z contains x and the additional dummy variable “respondent was household head” which turned out to be significant in the selectivity equation (3.2). Let y be the reported income (else $y = 0$), then we have

$$\begin{aligned} E(y|x, y > 0) &= x^T \beta + E(u|x, y > 0) \\ &= x^T \beta + \alpha \cdot \lambda(z^T \theta) \end{aligned} \tag{3.3}$$

where $\lambda(\cdot)$ is parametrically specified if the joint distribution of (u, ϵ) from equations (3.1) and (3.2) is. Therefore, the first step is the estimation of equation (3.2) to obtain θ , and the second step is the estimation of equation

$$y = x^T \beta + \alpha \cdot \lambda(z^T \hat{\theta}) + v \tag{3.4}$$

where $E[v] = E[v|x, z^T \theta] = 0$. Note that for the prediction of the means of the missing values one refers again to the original equation (3.1).

We tried several parametric and semiparametric estimation methods; see references in Section 2.1. We started with the fully parameterized version of Heckman where, as a result from assuming joint normality for (u, ϵ) , $\lambda(\cdot)$ is the inverse Mill’s ratio; see Figure 4. Then we tried to use a semiparametric single index estimator for equation (3.2), and a partial linear model estimator for the second step. As all implementations for the single index estimation we tried turned out to be quite unstable, we finally estimated the selectivity equation with a probit and applied its $\hat{\theta}$ in a smoothing-spline based partial linear model in (3.4); see the next to last column of Table 1. Similar to what we found in the first exercise, Section 3.1, this semiparametric estimation procedure had hardly an impact on the final results for the unconditional income distribution of \mathcal{S}_2 .

In Figure 4 we compare, once again, the different predictions based on either normality or gamma for the prior conditional distribution for homo- and heteroscedasticity, respectively. Again we show only results where the heteroscedasticity is constraint to a constant coefficient of variance CoV. Contrary to what we often observe in rich, industrialized countries, our estimates suggest that the households not reporting their income tend to have smaller incomes, on average, compared to households with the same characteristics but reporting their income. Though it would be interesting to study this finding in more depth, this is clearly beyond the scope, and is not the motivation, of this paper. As it is about half of the households that did not report their income, this could have a notable impact on the total

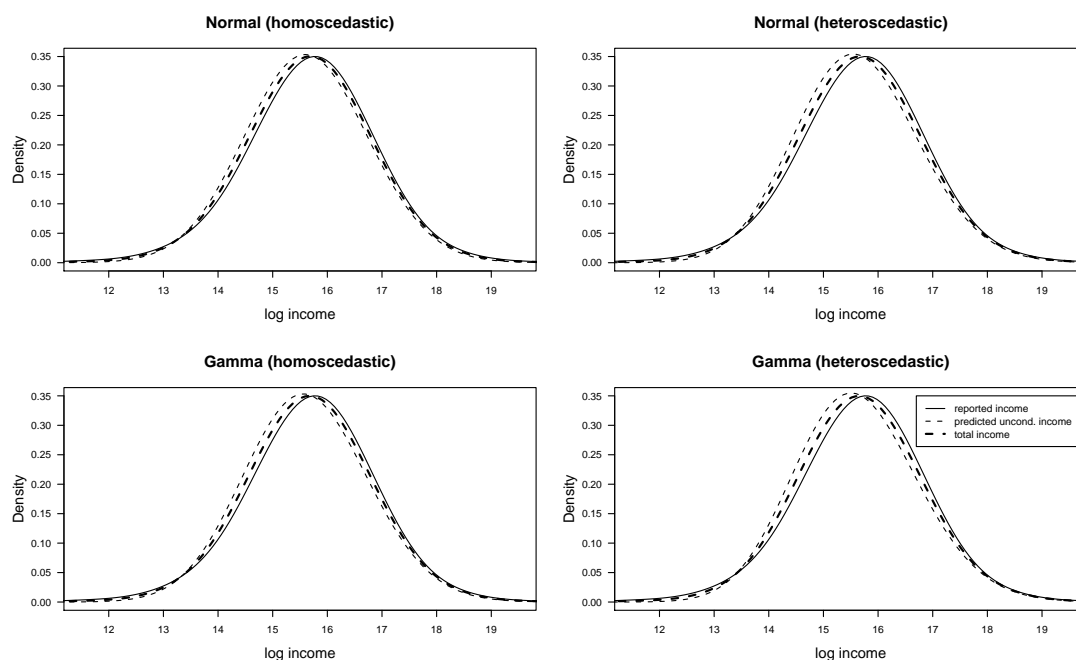


Figure 4: Estimated and predicted density curves of unconditional income for households with not reported income (grey dashed), households with reported income (solid line), and for the whole sample (dark dashed) in 2008, based on different prior assumptions from the upper left to the lower right.

income distribution which is also shown in Figure 4.

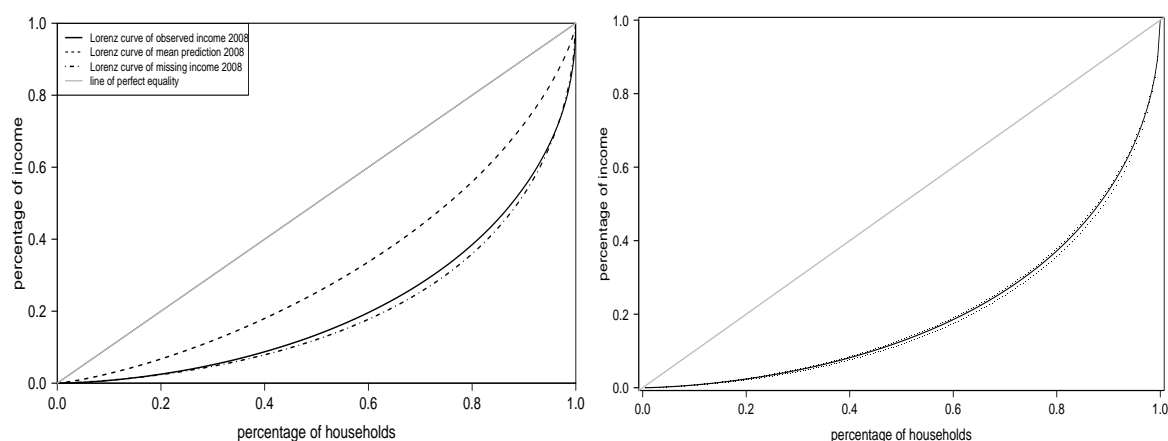


Figure 5: Left figure: The Lorenz curves for the observed (solid) income, the conditional income (thick dashed) and the predicted income (dotted-dashed). Right figure: The Lorenz curve for the total survey, i.e. observed plus predicted with 99% point-wise confidence intervals.

In view of this potential source of bias, one should study the consequences e.g. for the Lorenz curve and Gini coefficient. In Figure 5, left column, we see the resulting Lorenz curves for the

conditional and the unconditional predicted incomes and for comparison the Lorenz curve for the observed incomes. This once more demonstrates that missing values must not be replaced by mean predictions even if mean prediction might be the best one can do for the prediction of individual household incomes. Concerning the observed versus the predicted income distribution we see the main difference for the mean of households. Nonetheless we see also, that the income distribution for households which did not report income does not substantially deviate from the one of reported incomes. Moreover, one should have in mind that our predictions are based on estimation, so they are random although they are not based on simulations. One would therefore like to have an idea of this randomness and construct confidence intervals. We could do this for densities but, equally well, we can do this for the Lorenz curve. In the literature one can find confidence intervals for the simulation based predictions (where normal random errors were added to the individual income predictions). However, they were constructed from repeated simulations, which shows the uncertainty of the simulation method - and therefore proves why an explicit analytic method like ours might be preferable, but it does not reflect the uncertainty due to the estimation based prediction. We recommend to construct confidence intervals or bands based on bootstrap or subsampling from the very first step. For parametric bootstrap or the alternative subsampling we refer to Politis et al. (1999). For bootstrap inference in semiparametric additive models to Härdle et al. (2004), and for mixed effects or small area models to Lombardía and Sperlich (2008). For the purely parametric model, a trivial bootstrap that draws random samples of size n from the original sample and then simply repeats the whole procedure, is sufficient. In Figure 5, right column, we see the 99% confidence interval for the Lorenz curve.

As we already mentioned in the introduction, predicted income values typically tend to be too high for the poorest households and too low for the richest. Measures of inequality in an income or expenditure distribution such as the Gini coefficient are certainly very sensitive to that. Therefore we study also the performance of our method to estimate the Gini coefficient. This coefficient is a specific indicator, which ranges from 0 to 1, where 0 indicates perfect equality and 1 total inequality. It corresponds to twice the area between the Lorenz curve and the diagonal. In our application now, the Gini for the observed income is 0.579, for the income of non-reporting households it is 0.581 with our method but just 0.368 for the conditionally predicted incomes. Putting together observed and predicted unconditional income for the missing values respectively, the total Gini for the population is 0.582 with a 90% bootstrap confidence interval of $[0.578, 0.590]$. Note that the Gini of the observed is right the upper bound of this interval.

4 Predicting the expenditure distribution

Already the first case studies gave us some evidence of the importance of a method for poverty, inequality and vulnerability analysis. In this Section, we perform a further study, but now for prediction rather than for estimation. The problem is to predict the distribution of consumption expenditures of a cohort from the past. Certainly, it is also possible to change the role of \mathcal{S}_1 and \mathcal{S}_2 for historical studies to get an idea for past distributions thanks to extrapolation from earlier but complete data. A prediction from the 2000 cohort to the 2008 cohort is maybe a little bit too adventurous as the returns have probably changed over that time period, especially in Indonesia. Therefore, either the mean prediction or the scedasticity prediction will fail. Instead, we tried to predict the expenditure distribution of the 2000 cohort with the aid of the 1997 cohort. For evaluation issues we will predict the expenditures in 2000 only for that part of the population (4585 households) for which we had actually observed the expenditures. In practice one predicts correctly for the households and cohorts where there is a lack of information. From 1997 we can use 5406 observations having reported their expenditures and all predictor variables x , compared to only 439 incomplete records.

Given our experiences from above, for brevity we limit the presentation to the results based on a linear regression model for the mean. The coefficients with its standard deviations are given in Table 5 in the Appendix. We calculated the real per capita consumption for each household by dividing nominal per capita consumption by the inflation rate of the respondent household's province. We used a provincial price deflator based on the Badan Pusat Statistic consumer price indexes (CPI) reported for 45 cities in Indonesia and matched to the provinces included in the sample. For provinces with more than one city we use the simple average of the price index; cf. Chaudhuri et al. (2002). This gave us the regional inflation rates shown in Table 6 in the Appendix. This makes expenditures more comparable and meaningful over time and regions. Then, assuming that the expenditure behavior reflected by these coefficients is relatively stable over the considered time period, we applied the four different priors on \mathcal{S}_2 , i.e. conditional normality and gamma under homo- and heteroscedasticity with constant CoV. The final step is the in Section 2 mentioned calibration. Referring to the measurement of the real GDP per capita provided by the WDI in 2003 we notice that there is a decline of nearly 11.97% from 259 in 1997 to 228 in 2000. Given the assumption that the economy of average household income is mirrored in the national real GDP per capita, we expect a decrease in household income of around 11.97% from 1997 to 2000. The resulting unconditional predictions of expenditure distribution become comparable to the - in our illustration - observed one. The results are given in Figure 6.

To better quantify the differences of the performance among different settings, we estimated

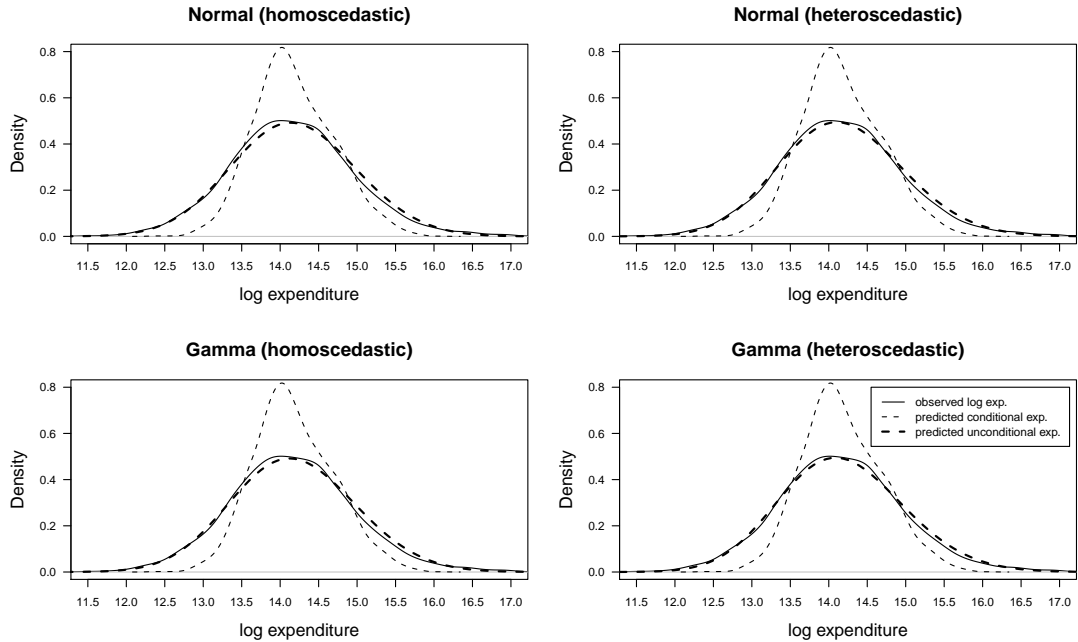


Figure 6: Density curves for the conditional expenditures (grey dashed) the predicted unconditional expenditures (thick dashed) for 2000 based on a 1997 cohort, and a kernel density estimates of the observed expenditures (solid line) in 2000.

the integrated squared error

$$ISE = \int_{-\infty}^{\infty} [\hat{f}(y) - f(y)]^2 dy , \quad (4.1)$$

where $f(\cdot)$ indicates the true expenditure density and $\hat{f}(\cdot)$ our predictor. As we do not really know the true f , this was replaced in our calculations by a kernel density estimate with Gaussian kernel, Silverman's rule-of-thumb bandwidths and using the in 2000 actually reported 4585 household expenditures. Under the assumption of homoscedasticity we got 0.0012 for normal and 0.0010 for gamma priors, but only 0.0008 and 0.0007 for heteroscedastic normal and gamma priors. Not that surprising for people familiar with mixture methods, and because one maybe does not expect important asymmetries in the conditional density, the difference between normal and gamma priors is less accentuated than the somewhat remarkable difference between homo- and heteroscedasticity. For the predicted distributions under heteroscedasticity for the prior, the corresponding Lorenz curves hardly differ from the one based on the actually reported expenditures. Similar to the preceding application, we again calculate the measurement index of inequality in the expenditure distributions, here the Gini coefficient. The results were a predicted value of 0.447 with $[0.429; 0.455]$ as its 90% bootstrap confidence interval, and a value of 0.451 for the observed expenditures. These results, as well as the following ones refer to the gamma prior under heteroscedasticity

but hardly differ from those obtained when substituting by the normal. Overall, the results are very promising so far.

A question of central interest is to trace the development of poverty in the underdeveloped and the developing countries. Certainly, there exist many different definitions of poverty lines. The hardest ones to predict in our context are probably the absolute ones as any slight shift of the mean e.g. by calibration can easily have a fundamental impact on the prediction of the number of households being classified as poor. Therefore, if prediction methods for other cohorts or years have to be applied or for scenario studies, it is more reasonable to consider relative poverty measures. Hence, we used the poverty line defined as 40 percent of the country's median consumption. The poverty line was then at 13.21458 log Rps per year along the reported, and 13.20469 log Rps along the predicted income distribution. Once the poverty line is fixed, one can see from the predicted density the percentile lying below this line. For the case of a particular small or moderate set of households it might be even interesting to look directly at the individuals. In that case we need to assign each household a position inside the unconditional distribution, based on his characteristics x . Based on the probability densities obtained above, one could approximate the distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ e.g. by linear interpolation using the cumulated distribution value. Then, for a household with given x and predicted mean \hat{y} one may construct a projection into the unconditional distribution along

$$\hat{y}_{uncond} = F^{-1}(F_{\hat{y}}(\hat{y})), \quad (4.2)$$

where $F_{\hat{y}}$ indicates the cumulated distribution function of the conditional income. We emphasize that this must not be considered as optimal prediction of the household income, which is still the mean prediction with an accuracy depending for example on the R^2 of the mean regression. We are simply assigning each individual a place according to its x inside the predicted unconditional distribution. In contrast, this can be very helpful for the analysis of vulnerability to poverty.

Now, the approximated expenditures generated from the inverse distribution function 4.2 give an estimate for how many people will fall below the poverty line. The accuracy of the predicted unconditional consumption expenditures can then be examined by cross tabulating the predicted with the observed consumption expenditures, see Tables 2 to 4. In Table 2 are compared the number and percentages of actual non-poor and poor compared to the predicted values and its confidence intervals. One is tempted to speak of an almost perfect prediction thanks to our new method. In Tables 3 is shown what a purely mean prediction would tell us about poverty. Finally, in Table 4 we analyzed the prediction quality of our method for the individual household level. While, not surprising, most of the non-poor are classified correctly, this is not the case for the poor. The outcomes of Table 3 and 4 are not surprising insofar that the mean regression had an R^2 of about 39% for 1997. The tails of

the marginal expenditure distributions are therefore mainly determined by the households' unobserved heterogeneity. This is why we said these methods are helpful for vulnerability but not for tail predictions of the individual level.

	Observed	Predicted	90% Prediction Interval
Not Poor	4079 (88.96%)	4063 (88.62%)	[4056; 4079] (88.46% – 88.96%)
Poor	506 (11.04%)	522 (11.38%)	[506; 529] (11.04% – 11.54%)

Table 2: Number of Households below the relative poverty line according to the unconditional distribution prediction

	Observed	Predicted	90% Conf.Int.	Predicted	
				<i>NotPoor</i>	<i>Poor</i>
<i>NotPoor</i>	4079	4495	[4476; 4509]	3711	368
<i>Poor</i>	506	90	[76; 109]	352	154

Table 3: Number of Households below the relative poverty line according to the mean prediction

Table 4: Individual classification of households, predicted versus reported.

5 Conclusions

Our aim is to estimate or predict a monetary distribution, like income or consumption expenditures. The mean regression gives only the conditional distribution which is only poor estimator of the unconditional (marginal) distribution. For certain welfare studies one could use quantile regression instead but again fails to predict the marginal distribution as a whole. If one uses quantile regressions for each percentile to afterward (re)construct the unconditional distribution, one lacks of a common model and estimator, probably suffers estimation problems at the (most interesting) tails, and further inference is hardly possible. In the literature many different models and methods were proposed, compared and rejected; many of them being simulation methods.

We propose a simple method based on mild assumptions to get an analytic and unique estimator for the whole required marginal distribution. The calculus of derivatives, Lorenz curve, or any index, poverty or inequality measure is straight forward. Furthermore, the explicit analytic form of our estimate makes inference possible and similar to, for example, the construction of confidence and prediction intervals.

There exist mainly two or three ways to understand and interpret our method, in particular the integration or averaging step; see equation (2.3). The Bayesian approach is to think of

the required distribution as a random function which can be described via estimated moments and appropriate conditional prior distributions. A more frequentist, but still modeling approach, is to rely on n-fold mixture models working with estimated but (via common regression models) linked parameters. As a special case we can even think of the nonparametric approach via kernel density estimation. Here now, the conditional prior distribution is our kernel, and the scedasticity function is the data-adaptive local bandwidth. Homoscedasticity then resembles the use of a common global bandwidth. The use of asymmetric priors corresponds to the case of applying special kernels typically used for boundary correction or asymmetric information (like the knn estimators do in nonparametric regression). They are therefore recommendable if prior knowledge on boundaries or skewness is available. A common conclusion of each of these three interpretations is that the choice of the prior distribution plays a minor role, the scedasticity function is indeed more important, and the quality of the mean regression has mainly an impact on the variability of the final estimate.

For the regression estimations necessary for the required moments, our method is not at all restricted to particular methods or models; parametric, nonparametric, semiparametric, selectivity correction or mixed effects models for cross section, panel or times series; the here proposed method can straightforwardly be combined with each of them. Inference can most easily be based on bootstrap or subsampling methods.

We have shown the use and the practical usefulness of our method in three different contexts: data matching from one sample to another, the completing of surveys with many missing values (probably endogenous), and the prediction to the future. One could add survey-to-census, cross-survey or cross-country data matching or scenarios for the prior evaluation of treatment and policy effects. Our motivation, however, was the illustration and the study of the performance of this method that can only be done if a reference distribution based on real observations is available. As the implementation and use of our method is relatively simple in any of the typically applied software packages like, for example, gretl, R, SAS, S-plus or Stata, this presents a rather powerful though handy tool for practitioners and empirical researchers.

References

- Abeyasekera, S. and Ward, P., 2002. Models for Predicting Expenditure per Adult Equivalent (for AMMP surveillance sentinel sites). Tanzanian Ministry of Health, UK DFID, University of Newcastle upon Tyne, and Districts of Hai, Ilala, Morogoro, Rufiji, Temeke, Adult Morbidity and Mortality Project.
http://research.ncl.ac.uk/ammp/site_files/public_html/finalproxies.pdf
- Ahn, H. and Powell, J. L., 1993. Semiparametric estimation of censored selection models

- with a nonparametric selection mechanism. *Journal of Econometrics* 58(1-2), 3-29.
- Alisjahbana, A. S., Yusuf, A. A., Chotib, Yasin, M. and Soeprobo, T. B., 2003. Understanding the determinants and consequences of income inequality in Indonesia. Revised version of the paper presented at the “Bangkok Conference on Comparative Analysis of East Asian Income Inequalities” titled: “Income distribution and sustainable development: The Indonesian experience”. www.eadn.org/reports/iwebfiles/i04.pdf, accessible on August 7, 2009.
- Atkinson, A. B. and Bourguignon, F.,(Ed.) 2000. *Handbook of Income Distribution*. Amsterdam: North-Holland.
- Azzarri, C., Carletto, G., Davis, B. and Zezza, A., 2006. Monitoring poverty without consumption data. *Eastern European Economics* 44(1), 59-82.
- Birkin, M. and Clarke, M., 1989. The generation of individual and household incomes at the small area level using synthesis. *Regional Studies* 23(6), 535-548.
- BPS (various years), Statistik Indonesia, Badan Pusat Statistik, Jakarta.
<http://webapps.bps.go.id/cpi/tables.cfm>, accessible on February 26, 2009.
- Chaudhuri, S., Jalan, J. and Suryahadi, A., 2002. Assessing household vulnerability to poverty from cross-sectional data: A Methodology and Estimates from Indonesia. Discussion Paper Series, Department of Economics, Columbia University.
- Chotikapanich, D., (Ed.) 2008. *Modeling Income Distributions and Lorenz Curves*. Series: Economic Studies in Inequality, Social Exclusion and Well-Being 5, Springer.
- Davis, B., 2003. Choosing a method for poverty mapping. Food and Agriculture Organization of the United Nations, Rome.
<http://www.fao.org/docrep/005/y4597e/y4597e00.htm>, accessible on March 5, 2011.
- Demombynes, G., Elbers, C., Lanjouw, J. O. and Lanjouw P., 2007. How good a map? Putting small area estimation to the test. World Bank Policy Research Working Paper 4155.
- Elbers, C., Lanjouw J. O. and Lanjouw P., 2003. Micro-level estimation of poverty and inequality. *Econometrica* 71(1), 355-364.
- Filmer, D. and Pritchett, L. H., 2001. Estimating Wealth Effects without Expenditure Data - or Tears: An Application to Educational Enrollments in States of India. *Demography* 38(1), 115-132.
- Firpo, S., Fortin, N.M. and Lemieux, T., 2009. Unconditional quantile regression. *Econometrica* 77(3), 953-973.

- Gasparini, L., Cicowiez, M., Gutiérrez, F. and Marchionni, M., 2003. Simulating income distribution changes in Bolivia: a microeconomic approach. The World Bank Bolivia Poverty Assessment.
- Ghosh, M. and Rao, J. N. K., 1994. Small area estimation: an appraisal. *Statistical Science* 9(1), 55-93.
- Härdle, W., Huet, S., Mammen, E. and Sperlich, S., 2004. Bootstrap inference in semi-parametric generalized additive models. *Econometric Theory* 20, 265-300.
- Heckman, J. J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5(4), 120-137.
- Heckman, J. J., 1979. Sample selection bias as a specification error. *Econometrica* 47(1), 153-161.
- Hentschel, J., Lanjouw, J. O., Lanjouw, P. and Poggi, J., 2000. Combining census and survey data to trace the spatial dimensions of poverty: a case study of Ecuador. *World Bank Economic Review* 14(1), 147-165.
- Horton, N. J. and Lipsitz, S. R., 2001. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician* 55(3), 244-254.
- Hyman, G., Larrea, C. and Farrow, A., 2005. Methods, results and policy implications of poverty and food security mapping assessments. *Food Policy* 30(5-6), 453-460.
- Koenker, R., 2005. *Quantile Regression*. Cambridge University Press, USA.
- Little, R. J. A. and Rubin, D. B., 2002. *Statistical Analysis with Missing Data* (Second Edition). John Wiley, New York.
- Lombardía, M. J. and Sperlich, S., 2008. Semiparametric inference in generalized mixed effects models. *Journal of Royal Statistical Society: Series B* 70(5), 913-930.
- Marin, J. M., Mengersen, K. and Robert, C. P., 2005. Bayesian modelling and inference on mixtures of distributions. In: Dey, D. and Rao, C. R. (Eds.) *Handbook of Statistics* 25, 459-507.
- McLachlan, G. and Peel, D., 2000. *Finite Mixture Models*. Wiley Series in Probability and Statistics.
- Mincer, J., 1958. Investment in human capital and personal income distribution. *The Journal of Political Economy*, 66(4), 281-302.

- Paulin, G.D. and Ferraro, D. L., 1994. Imputing Income in the Consumer Expenditure Survey. *Monthly Labor Review* 117(12), 23-31.
- Politis, D.N., Romano, J.P. and Wolf, M., 1999. *Subsampling*. Springer, New York.
- Pradhan, M., Suryahadi, A., Sumarto, S. and Pritchett, L., 2001. Eating like which ‘Joneses?’ an iterative solution to the choice of a poverty line ‘reference group’. *Review of Income and Wealth* 47(4), 473-487.
- Ravallion, M., 2001. Growth, inequality and poverty: Looking beyond averages. *World Development* 29(11), 1803-1815.
- Rodríguez-Póo, J. M., Sperlich, S. and Fernández, A. I., 2005. Semiparametric three-step estimation methods for simultaneous equation systems. *Journal of Applied Econometrics* 20, 699-721.
- Rothe, C., 2009. Nonparametric Estimation of distributional policy effects. *Journal of Econometrics* 155(1), 56-70.
- Royston, P., 2004. Multiple imputation of missing values. *The Stata Journal* 4(3), 227-241.
- Sahn, D. E. and Stifel, D. C., 2000. Poverty comparison over time and across countries in Africa. *World Development* 28(12), 2123-2155.
- Sahn, D. E. and Stifel, D. C., 2003. Exploring alternative measures of welfare in the absence of expenditure data. *Review of Income and Wealth* 49(4), 463-489.
- Su, Y.-S., Gelman, A., Hill, J. and Yajima M., 2010. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, forthcoming.
- Tarozzi, A. and Deaton A., 2009. Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics* 91(4), 773-792.
- Zeller, M., Alcaraz, G. and Johannsen, J., 2004. Developing and testing poverty assessment tools: Results from accuracy test in Bangladesh. College Park, Maryland: IRIS Center, University of Maryland.
- Zeller, M., Johannsen, J. and Alcaraz, G., 2005. Developing and testing poverty assessment tools: Results from accuracy test in Peru. College Park, Maryland: IRIS Center, University of Maryland.

Appendix

	Application 3 (linear model)
Constant	11.77 (.1251)
Average age	.0231 (.0042)
Average age squared	-.0003 (.0000)
Average year of schooling	.0488 (.0025)
Log of assets per capita	.1567 (.0061)
Share of asset to business	.0289 (.0438)
Farmer in family	-.2011 (.0247)
Share of working hhm	-.1082 (.0562)
Share of female hhm	-.0521 (.0612)
HH size	-.0567 (.0042)
Located in urban area	.1604 (.0212)
North Sumatera	-.3990 (.0463)
West Sumatera	-.1974 (.0578)
South Sumatera	-.3636 (.0561)
Lampung	-.2568 (.0556)
West Java	-.2754 (.0415)
Central Java	-.3456 (.0425)
Yogyakarta	-.5030 (.0534)
East Java	-.6584 (.0417)
Bali	-.4339 (.0500)
West Nusa Tenggara	-.3546 (.0482)
South Kalimantan	-.1474 (.0525)
South Sulawesi	-.6501 (.0493)
Number of observations	5406

Table 5: Regression results of mean expenditures. Figures in parentheses give the standard deviations

Province	Inflation Rate		
	1998	1999	2000
Aceh	78.71	6.09	9.57
North Sumatra	82.53	0.66	5.37
West Sumatra	87.87	4.23	10.99
Riau	64.35	2.04	9.67
South Sumatra	89.22	-1.01	8.49
Bengkulu	84.10	0.47	8.21
Lampung	84.66	3.34	10.18
Jakarta	74.78	1.77	10.29
West Java	72.89	2.94	6.55
Central Java	70.46	1.02	8.62
Yogyakarta	77.46	2.51	7.32
East Java	87.09	1.06	9.62
Bali	75.11	4.39	9.81
West Nusa Tenggara	90.14	0.59	5.19
Central Kalimantan	75.12	-2.56	10.22
South Kalimantan	75.50	1.47	7.57
East Kalimantan	71.70	3.35	11.29
South Sulawesi	79.35	1.64	9.73
Southeast Sulawesi	97.75	1.29	11.25

Table 6: Regional inflation rates in 1997, 1999 and 2000 (Rp per capita/month), see also Pradhan et al. (2001)