A Simple Method for Predicting Distributions by Means of Covariates with Examples from Poverty and Health Economics

Jing Dai[a], Stefan Sperlich[b] and Walter Zucchini[c]

JEL-Classification: C1, C4, I32, I15.
Keywords: predicting distributions, missing values, household expenditures, income distribution, health economics, impact evaluation.

SUMMARY

We present an integration based procedure for predicting the distribution *f* of an indicator of interest in situations where, in addition to the sample data, one has access to covariates that are available for the entire population. The proposed method, based on similar ideas that have been used in the literature on policy evaluation, provides an alternative to existing simulation and imputation methods. It is very simple to apply, flexible, requires no additional assumptions, and does not involve the inclusion of artificial random terms. It therefore yields reproducible estimates and allows for valid inference. It also provides a tool for future predictions, scenarios and ex-ante impact evaluation. We illustrate our procedure by predicting income distributions in a case with sample selection, and both current and future doctor visits. We find our approach outperforms other commonly used procedures substantially.

a Universität Kassel, Nora-Platiel-Str. 5, D-34109 Kassel, tel: +49(0)561 803 3044, dj-cn@hotmail.com

b Université de Genève, Geneva School of Economics and Management, 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.

c Department of Economic Sciences, Georg-August Universität, Platz der Göttinger Sieben 3, D-37073 Göttingen, tel: +49(0)551 39 7286, walter.zucchini@wi-wiss.uni-goettingen.de. We thank the editors and an anonymous referee for helpful discussion and comments. We also appreciated the discussions with the participants of the Annual meetings of the German Statistical Society 2010, and of the Swiss Statistical Society in 2011.

1. Introduction

The importance of the *distribution* of welfare indicators, such as household income or expenditure, is well-documented. It is used to compute a variety of poverty, inequality and development measures that provide policy makers with objective information on which to base decisions regarding the allocation of resources, and to monitor social security systems. There exist numerous initiatives, such as "Operationalising Pro-Poor Growth", of national and international institutions for which such information is essential. Ravallion (2001) emphasized the need for greater attention to be paid to micro-level distributions rather than just to averages. An advantage of estimating the entire distribution is that the various quantities of interest derived from it (e.g. Gini index) are coherent, which is not always the case if such quantities are estimated individually using separate models. Ideally the statistical methods used should allow for inferences, scenario simulations, and comparisons over time and space.

There is a large literature on the issue of poverty and inequality measurement focusing on the *distribution* of welfare indicators (see e.g. Paulin and Ferraro, 1994, on imputing income; Filmer and Pritchett, 2001, looking at missing data problems; Hentschel et al., 2000, on imputing the likelihood to be poor; or Atkinson and Bourguignon, 2000, as well as Chotikapanich, 2008, for general reviews). Many procedures are based on data matching: a mean regression, performed with a *source* sample (from population $\mathcal{S}$), of the variable of interest y^s on additional information x^s about the individuals or households to predict the income or expenditures for the group of interest (for which the information x^t is available, but not y^t).[1] The population of interest ($\mathcal{T}$) 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 (e.g. a census), a future or past panel wave (or cohort), or simply a hypothetical population in the case of scenario investigations.[2] The development of techniques to interpolate from one survey to a different, usually larger, *target* set of households or individuals has been summarized earlier by Davis (2003).

Standard regression-based predictors, $\hat{y}$ provide a biased estimate of the distribution of the indicator because the conditional density is "too narrow", rendering it useless for welfare assessment purposes, especially for assessing poverty and inequality. A popular remedy is to add normally distributed random noise

1 The superscripts *s* and *t* indicate *source* and *target* set, respectively.

2 E.g. to provide comparisons of the welfare indicator distribution over time and space (Sahn and Stifel, 2000).

to the individual predictions. In statistical terms, this constitutes a kind of wild bootstrap to simulate the distribution of the target population; see Gasparini et al. (2003). Such simulation-based methods are easy to apply but have various drawbacks resulting from the fact that the estimated distribution depends on the particular sequence of random errors that was added to the predictions. The estimate is not reproducible, which renders further valid inference complicated or unavailable. These drawbacks are mitigated if the method is just used for aggregates, e.g. in the context of small area statistics (Birkin and Clarke, 1989; Elbers et al., 2003). However, these small area methods are designed to estimate selected statistics, not the distribution itself. But even in that context, Tarozzi and Deaton (2009) discussed the approach quite critically; see also Zeller et al. (2005) and Azzarri et al. (2006).

Various methods on a related problem have been developed in the literature on policy evaluation where the focus is to estimate (or say 'predict') differences between the actual indicator (say y^s) and *counterfactual* outcomes (say y^t). Most of this literature looks at the difference in the mean, but its link to our problem comes from the fact that some have extended these considerations to (conditional) quantiles or distributions. The interest is focused on estimating causal differences, often in particular quantiles, so that a model is estimated for each. For our purpose these methods are unnecessarily computationally demanding and unsuitable in situations in which one wishes to have a single model instead of a large set of individual quantile models. Alternatively one could directly estimate the conditional distributions in $\mathcal{S}$ and then predict the unconditional distribution for $\mathcal{T}$; see Donald et al. (2012) and Chernozhukov et al. (2013) for reviews and inference.

More closely related to our problem is the literature on wage discrimination which focuses on distributions; see DiNardo et al. (1996) or Biewen and Jenkins (2005) for parametric approaches, and Rothe (2010) for a nonparametric case.[3] The main problem with the nonparametric approaches in practice is that, for our purpose, we wish to include as much information on x as possible, but the nonparametric estimators for conditional densities and cumulative distribution functions are particularly prone to the curse of dimensionality. Typically, in impact evaluation one includes as few covariates as possible.

We propose a novel but simple method that makes use of information from covariates x to estimate the distribution of a related indicator of interest, y. The difference to the above mentioned methods is that one looks at conditional distributions of y. The key assumption is that the conditional distribution of $y|x$ for

3 For related literature based on quantiles see e.g. Juhn et al. (1993), Melly (2005) or Noufaily and Jones (2013).

the target population $\mathcal{T}$ can be estimated in some reasonable way, e.g. that it is the same as the conditional distribution in a population $\mathcal{S}$, where $\mathcal{T}$ and $\mathcal{S}$ refer to comparable regions, or surveys in the same region but in different years. More problematic is the case where there is a single population in which the values of y are observed in a subpopulation $\mathcal{S}$ but missing in the target subpopulation $\mathcal{T}$. In the context of welfare surveys the proportion of missing values is often substantial and it can be seldom assumed that they are missing at random. One needs to take account of the factors that influence the probability that the value of y is missing. We note that this problem has not been addressed in the above-mentioned methods as their focus and context are different to those considered here.

The proposed estimation method can be based on parametric or nonparametric estimates of the conditional distribution but, for the reasons indicated above, the latter are unlikely to be useful in the applications for which the proposed estimator was developed. We do, however, also consider semiparametric models. This, as well as a variety of extensions to more complex models, can be done quite easily by choosing a flexible parametric conditional distribution, and by estimating its moments using appropriate (non-)parametric methods. The method is applicable to models such as mixed effects (multi-level) models, latent variable models and simultaneous equation systems. The resulting estimator for the target distribution of is a simple analytic formula and thus the estimates are reproducible and permit further valid inference.

A different approach is to estimate the missing values using so-called imputation methods; see e.g. Little and Rubin (2002) for which abundant software is available, Horton and Lipsitz (2001), Royston (2004) or Su et al. (2011). The estimates can then be used to estimate the target distribution based on the conditional independence assumptions. However, these methods are not designed for recovering distributions of one population from another or for making forecasts. The method illustrated in this paper is simpler and more transparent. Furthermore the results of a simulation study (using publicly available software), not reported here, suggest that the method introduced here outperforms the imputation-based approach.[4]

The outline of this paper is as follows: Section 2 introduces the statistical methodology. In Section 3 we consider different types of problems in estimating the income and predicting the expenditure distribution for data from Indonesia. Section 4 illustrates an application for a distribution of counts, namely the number of doctor consultations for a suburb of Sidney. Section 5 concludes.

4 In fairness it needs to be pointed out that the latter was not specifically designed to address the problem considered here.

2. A General Methodology for Estimating Welfare Distributions

Our objective is to estimate the distribution of some welfare indicator, y, for a set of households that we will refer to as the target population, $\mathcal{T}$. The target population need not physically exist; in scenario studies it might be a hypothetical population in which the covariates, x have been specified (e.g. forecast) by the researcher. The values of y are unknown but the values of a set of covariates x are known for all households in $\mathcal{T}$. The required marginal distribution of y can be written in terms of the conditional distribution of y given x:

$$\begin{aligned} f_t(y) &= \int f_t(y \mid x) dF_t(x) \\ \text{and } F_t(y) &= \int F_t(y \mid x) dF_t(x), \end{aligned} \tag{1}$$

where f_t, F_t represent the densities and cdfs of the target population, respectively. In order to predict $f_t(y)$ or $F_t(y)$ one replaces $f_t(y|x)$, $F_t(y|x)$ by estimates obtained from population $\mathcal{S}$. A popular approach in the treatment effect literature is to assume that $f_t(y|x) = F_s(y|x)$, called the *conditional independence assumption*. In many situations this is a reasonable assumption, but there also exist situations where it is not, for example if we want to predict the distribution of y for the non-responses (i.e. where $\mathcal{T}$ is simply the subsample for which the y-values are missing).[5] As will be illustrated in the applications to follow there are situations in which it makes sense to set $F_t(y|x)$ equal to $F_s(y|x)$ up to a calibrating additive and/or multiplicative factor.

In what follows estimators of the conditional distribution (and its moments) are obtained from the $\mathcal{S}$ sample but always used as predictors for their analog in $\mathcal{T}$ and so we will omit subindex for the conditional distribution predictors. The next step is to replace the integrals in (1) by the average over the set of covariates in the target population $\{x_i^t\}_{i=1}^m$. One obtains

$$\begin{aligned} \hat{f}_t(y) &= \frac{1}{m}\sum_{j=1}^{m} \hat{f}(y \mid x_j) \\ \text{and } \widehat{F}_t(y) &= \frac{1}{m}\sum_{j=1}^{m} \hat{F}(y \mid x_j). \end{aligned} \tag{2}$$

5 Suppose for example that y is income. The TE approach would predict the income distribution for $\mathcal{T}$ using the income (distribution) that would be expected if the non-responding households were representative of $\mathcal{S}$. In the context of welfare surveys with missing values such an assumption is usually unrealistic.

Simulation-based methods, in which random noise is added to the predicted values for each household, can be viewed as (random) approximations to (2). It can be proved that for repeated simulations, the average value of the simulated distributions converges to the estimator $\widehat{F}_t(y)$ as the number of repetitions increases.

We now turn to methods for predicting the conditional distribution in practice. In some cases one could use an existing estimate $\hat{f}(y\,|\,x)$, $\widehat{F}(y\,|\,x)$ or its moments, e.g. one computed for a similar region, or from a past survey in the same region. More common is the situation in which the conditional distribution is estimated from observations of both x and y from a sample of $\mathcal{S}$, say $\{(x_i^s, y_i^s)\}_{i=1}^n$. Any procedure that yields unbiased estimates of $f_t(y|x)$ (or $F_t(y|x)$ or its moments) for the range of x-values covered in $\mathcal{T}$ can be used.[6]

If one assumes that $f_t(y|x) = f_s(y|x)$ then one could try to estimate the conditional distribution function nonparametrically using the $\mathcal{S}$-sample. There are, however, some practical advantages in choosing a particular distribution (up to the conditional moments) for $f_t(y|x)$ and then estimating the unknown moments. One advantage is that there are many more semiparametric methods available for imposing structure on moments than for imposing structure on conditional densities or cdfs. This advantage becomes especially important if (a) economic theory requires a certain structure or (b) if x contains variables that are not binary.[7]

Parametric estimation of $f_t(y|x)$ has the additional advantage in situations where the assumption $f_t(y|x) = f_s(y|x)$ has to be relaxed in some simple way, for example if additional information on $f_s(y)$ or $f_s(y|x)$ is available, such as a known shift in the mean, the variance or the median. Then, the necessary calibration can directly be applied to these moments. Similarly, if $\mathcal{T}$ represents missing values and the objective is the estimation of their distribution, (or that of the population $\mathcal{S}+\mathcal{T}$ then the assumption that $f_t(y|x) = f_s(y|x)$ is often unrealistic. As is illustrated in the applications to follow the proposed estimator can accommodate methods that are available to predict the moment functions for the missings.[8] For both cases see our applications below.

6 The latter is known in the treatment effect literature as the 'common support assumption'.

7 Stock (1989) applied kernel regression to predict the mean effect of a change in the distribution of certain policy-related variables on a dependent variable. He pointed out and confirmed the problems of using purely nonparametric regression techniques in practice (curse of dimensionality, choice of bandwidth matrix, etc.).

8 We are not saying that analogues methods would not exist, or could not be developed for nonparametric densities, but they will clearly be much more cumbersome than the method we propose.

The obvious drawback of parametric modeling is that $f_t(y|x)$ may be misspecified even if the (first) moments are not. The question that cannot be answered in general is to what extent this negatively affects the final prediction of $f_t(y)$. In our applications, where the objective is the estimation of the marginal distribution of y and where sufficiently many covariates are involved, the choice of $f_t(y|x)$ turned out to be much less important than the proper specification of the moments.

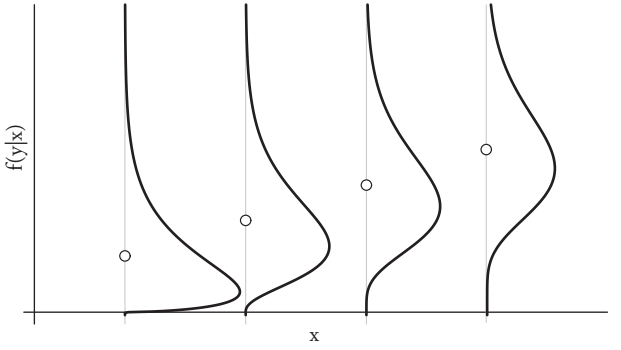
In what follows we focus on situations in which the empirical researcher choses a model for $f(y|x)$ and only estimates its unknown parameters, expressed in terms of moments. For example, a simple and popular model is the normal distribution whose mean and variance are functions of the covariates, say $\mathrm{N}(\mu(x),\sigma(x)^2)$ where the values are assumed to be independently distributed. The required conditional distribution is obtained by estimating $\mu(x)$ and $\sigma^2(x)$ from the observations in $\mathcal{S}$ enriched by additional information on $F_t(y)$ if available. The moment functions can be modeled parametrically, non- or semiparametrically; where necessary one can use Tobit models, apply selection bias corrections, appropriate weights for stratified samples, random effects and multilevel or panel models and so on. The data might suggest that the conditional distribution is heteroscedastic or that the skew also changes as a function of the covariates. The statistical literature offers a rich variety of models for conditional distributions, software is available for fitting them (e.g. the GAMLSS R-package) and model selection techniques to guide the choice of appropriate model.

Sometimes the pdf, $f_t(y|x)$, is bounded below by zero, there is evidence of heteroscedasticity and non-constant skewness. Figure 1 shows an example where all these features can be accommodated using a two-parameter family of models.[9] A simplification that can prove useful when dealing with heteroscedasticity is to assume that the coefficient of variation, $\mathrm{CoV} = \sigma(x) / \mu(x)$, is constant or, more generally, that $\sigma(x)$ is some simple function of $\mu(x)$. The plausibility of such assumptions can be assessed by examining the plot of $\hat{\sigma}(x)$ against $\hat{\mu}(x)$ computed from the observations in $\mathcal{S}$.

In summary, one simple general strategy is to use a flexible parametric family of distributions and fit it by equating its first two or three moment functions to those estimated from the observations in $\mathcal{S}$ under given restrictions (prior knowledge on f_t) where applicable. Consider a family of distribution with three parameters which can be expressed in terms of the first three moments. One

9 In practice it is sometimes difficult to determine which of the many possible models for $f_t(y|x)$ should be considered. For instance, Biewen and Jenkins (2005) chose to model the income distributions in the presence of covariates with the specifications of Singh-Maddala (1976) and Dagum (1977); see also Van Kerm (2013).

Figure 1: Gamma distributed conditional densities.



Notes: Gamma conditional pdfs shown as functions of a single covariate x for which the mean and variance functions are both straight lines. The circles show the conditional means.

estimates the moment functions $\mu(x) = E[y\,|\,x]$, $\mu_2(x) = \sigma^2(x) = \mathrm{Var}[y\,|\,x]$, and $\mu_3(x) = E[(y - E[y\,|\,x])^3\,|\,x]$ with the available data using mean regression. Where additional information regarding is available, for example the mean household expenditure of the target population, the estimate $\hat{f}_t(y)$ can be calibrated accordingly, as illustrated in the next section.

As justification of this strategy we first note that our objective here is not to test hypotheses, nor to interpret model parameters; it is to estimate $f_t(y)$ or $F_t(y)$. The estimator given in (2) is a mixture of m density functions. Mixtures are known to give excellent approximations and are consistent under (different sets of) typically mild conditions, see for example McLachlan and Peel (2000). Kernel density estimators with second order kernels are also local m-fold mixtures. In our case, the kernel $K_{\hat{h}_j}(\cdot - \hat{y}_j)$ is the conditional pdf with mean $\hat{m}(x_j)$ and variance $\hat{\sigma}^2(x_j)$ where $\hat{y}_j = \hat{\mu}(x_j)$ and $\hat{h}_j = \hat{\sigma}(x_j)$, i.e. the first two (estimated) moments of $f_t(y)$. In other words, controlling for the second moment corresponds applying local bandwidths in kernel density estimation. Controlling for the third moment (in a three-parameter distribution) corresponds to local kernels with asymmetric weighting, as in the so-called "knn smoothing".

3. Predicting Income and Expenditure Distributions in Indonesia

We use a longitudinal household-level data set from the Indonesia Family Life Survey (IFLS) that provides observations at the individual and household level on consumption, income, health, education, housing and employment. The IFLS sample is representative of about 83 % of the Indonesian population living in 13 of the 26 provinces in the country (Mishra, 2009). For 1997, 2000 and 2008 the IFLS covered between 6,500 and 10,000 households, in part from cross-sectional cohorts and partly from a panel. Household expenditure and income information is available but 20 % to almost 50 % of the values are missing. In our study household income per capita is taken as the sum of five sources, namely income from (1) wages and salary in both cash and in-kind transfers; (2) agricultural business; (3) non-agricultural business; (4) household non-labor sources, e.g. estimated rent, pensions, scholarships, etc; (5) household assets. In what follows we model expenditure and income in log(Rupiahs).

3.1 A Split-Sample Verification Exercise

This exercise was carried out to assess the accuracy of the proposed estimator in an artificial setting in which the true result is known. We took the 5,567 households in 2008 for which income was recorded and split them into two sets; 2,783 for $\mathcal{S}$ and 2,784 for $\mathcal{T}$. Since the income in $\mathcal{T}$ is in fact known (but not used in the estimation) we are able measure the accuracy of our estimate.

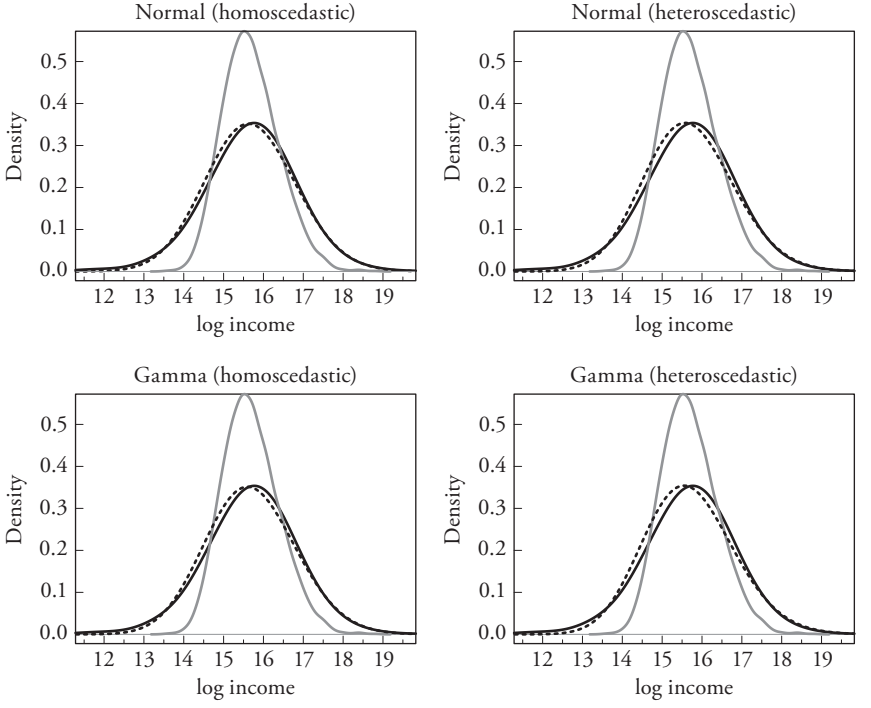
We applied different estimators and models to get an impression of their relative merits and of the sensitivity of our procedure with respect to changes to specific assumptions. We refrain from discussing the choice of the covariates (listed in Table 7). The first column of Table 7 gives the OLS estimates of the regression coefficients for the simple linear model used in most poverty assessments based on linear regression. The second column lists estimates of the coefficients (for the parametric components) for a fitted semi-parametric model, namely the additive partial linear model (APLM) using smooth flexible functions g_k, $k = 1,2,\ldots,K$

$$\mu(x) = c + u'\beta + \sum_{k=1}^{K} g_k(t_k), \tag{3}$$

with $\mathrm{x}' = (u', t')$ where u denotes the vector containing all covariates entering the model linearly and $t = (t_1, t_2, \ldots, t_K)$ the vector of those whose contributions to $\mu(x)$ is modeled non-parametrically.

We fitted the normal and the gamma to the conditional distribution of the log-income, given the covariates, and investigated a variety of specifications that allow for heteroscedasticity. We present a subset of the results to compare the use of the normal with the gamma distributions, the simple linear model with the APLM, and the homoscedastic with the heteroscedastic options. For the last comparison we show only the case that assumes a constant CoV (cf. Section 2).

Figure 2: Density estimates using an APLM regression model



Notes: Estimate of $f_t(y)$ based on the regression predictions $\{\hat{y}_j\}_{j=1}^m$ (grey line), $\hat{f}_t(y)$ computed using the proposed estimator (2) (dotted black line), under the 4 different models for $f(y|x)$ indicated above the panels, and $f(y)$ (solid black line).

Recall that in this artificial example the values of y in $\mathcal{T}$ are known. The solid line in the panels of Figure 2 shows a nonparametric smooth density fitted to these values (the true density $f_t(y)$). The dotted lines show a smooth of the conditional

means, $\{\hat{y}_j = \hat{\mu}(x_j)\}_{j=1}^m$, i.e. the predicted values from the regression[10]. The dashed lines show the estimates $\hat{f}_t(y)$ obtained using estimator (2).
The estimates for the log(income) distribution in $\mathcal{T}$ using an APLM are given in Figure 2. A striking feature in all displays is the enormous bias of the estimator based on the usual regression predictions. This is not surprising given an R^2 of just over 30 % for the regressions. A second feature is that the different estimates $\hat{f}_t(y)$ based on (2) differ very little. In this example (in which $f_t(y)$ is known) their relative accuracy was compared using the integrated squared error

$$\text{ISE} = \int_{-\infty}^{\infty} [\hat{f}_t(y) - f_t(y)]^2 \, dy.$$

The APLM is marginally better than parametric regression. The choice between homo- and heteroscedasticity, and also the choice of conditional distribution, seem to have slightly more impact than the choice of regression method.
A third noticeable feature is that, in all cases the mode of $\hat{f}_t(y)$ is slightly smaller than that of $f_t(y)$. To investigate this more closely we repeated the exercise several times, i.e. splitting the original 5,567 observations at random to obtain different (samples of) $\mathcal{S}$ and $\mathcal{T}$, and then redoing the estimation. That investigation revealed that the discrepancy is due to sampling variation rather than to a systematic bias. Apart from that blemish the estimator proved to work very well in this application. Particularly pleasing is how little the estimates depended on the type of models used for. In retrospect this finding is not surprising when one recalls that the estimator (2) is a m-fold mixture of distributions and not unlike a (nonparametric) kernel density estimator for which is well-known to be insensitive to the shape of the kernel. This is excellent news because in practice it is difficult to know which of the many possible models should be used for $f_t(y|x)$.

3.2 Estimating the Income Distribution when there Is Sample Selection Bias

The sets for $\mathcal{S}$ and $\mathcal{T}$ considered in Section 3.1 were artificially selected in such a way that the values of y in $\mathcal{T}$ were "missing completely at random". Such an assumption is seldom justified in the context of welfare surveys. Methods have been developed to take account of the likelihood that a particular respondent will disclose sensitive information, such as income. It is not our objective to discuss

10 These densities were computed using a kernel method with Gaussian kernel and twice the Silverman's rule-of-thumb bandwidth because the default bandwidth led to insufficient smoothing.

the circumstances under which such methods are appropriate; it is to illustrate that the proposed estimator is also applicable when corrections for selection bias are used.

We again consider the IFLS data from 2008 in which 5,567 households reported their income but 4,894 did not. We take the former as $\mathcal{S}$ and the latter as $\mathcal{T}$. In order to avoid the "missing at random" assumption we model the selection by

$$y = \mu(x) + \varepsilon, \text{ with selection } s = 1\{\nu(z) + \eta\}, \tag{4}$$

where y is log(income), x are covariates, s is a binary variable (1 if the income is known, and 0 if it is missing), z are covariates that provide information about the probability that y will be available; $\mu(x)$ is the mean of the distribution $f_t(y|x)$, ν is a function of the selectivity model covariates z, and ε and η are residuals. In our application z contains x (see Table 7) with two additional covariates, namely "respondent was household head" and "respondent is married", that turned out to be significant in the selection model of (4).

The expectation of the available (i.e. non-missing) y values is given by

$$E(y \mid x, s = 1) = \mu(x) + E(\varepsilon \mid x, s = 1) = \mu(x) + \alpha \cdot \lambda(\nu(z)). \tag{5}$$

If the joint distribution (ε, η) in (4) is specified parametrically then the function $\lambda(\cdot)$ is also parametric. One begins by estimating the function $\nu(\cdot)$ for the selectivity model using the observations from both $\mathcal{S}$ and $\mathcal{T}$. Then the data from $\mathcal{S}$ (for which the income values are available) is used to estimate the remaining parameters $\mu(x)$:[11]

$$y = \mu(x) + \alpha \cdot \lambda(\hat{\nu}(z)) + v, \text{ where } E[v] = E[v \mid x, \nu(z)] = 0. \tag{6}$$

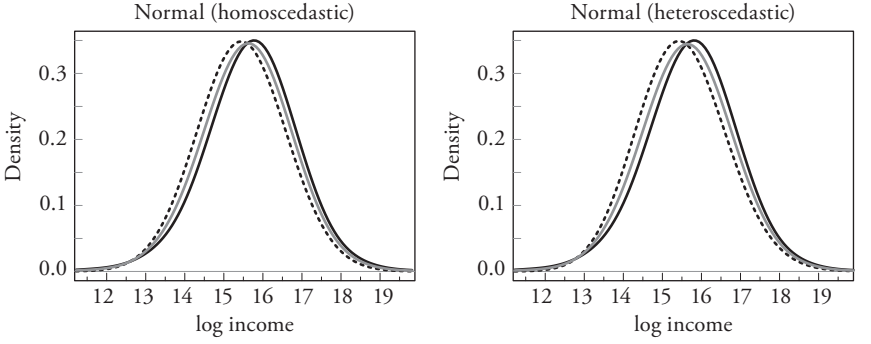
The mean regression prediction for the *j-th* missing value is $\hat{y}_j = \hat{\mu}(x_j)$, $j = 1,2,\ldots,m$.

We experimented with a number of parametric and semiparametric estimation methods starting with a fully parameterized version of the Tobit 2 model

11 There is some concern about a proper identification of all coefficients in the main equation of the structural model if only discrete variables fulfill the exclusion restriction. From a theoretical point of view, this criticism concerns mainly triangular systems, whereas our selection mechanism enters the main equation nonlinearly via the inverse Mill's ratio. Note that most variables in our data set are dummies, so the first mentioned criticism could still be maintained when we had more variables included in z. However, it needs to be kept in mind that our aim is not the identification of parameters but the prediction of the marginal distribution of y.

assuming that (ε,η) are jointly normally distributed, i.e. λ being the inverse Mills ratio. The estimates are shown in Figure 3, and column 3 and 4 of Table 7.[12] In the figure we compare the estimates using the normal distribution under homo- and heteroscedasticity, respectively. Again we show only results for the case in which heteroscedasticity was modeled under the assumption that the coefficient of variation is constant; more flexible options led to no improvement.

Figure 3: Normal distributed densities under homo- and heteroscedasticity



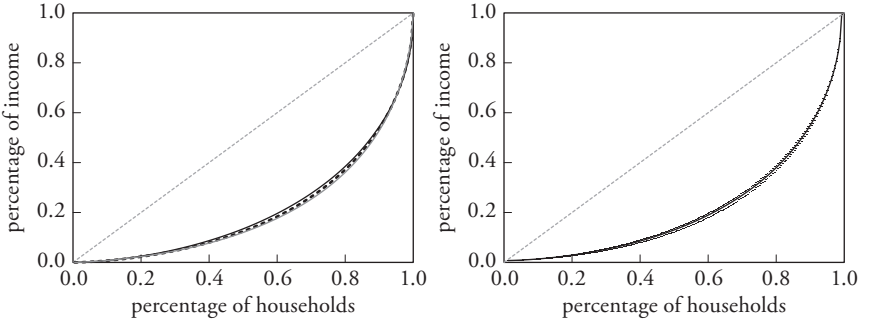
Notes: The $\hat{f}(y)$ based on (2) when using Tobit 2-step estimates for predicting $\mu(x)$. We see the $\hat{f}(y)$ for households that did not report income (dotted black line), a kernel smooth of the pdf for households that did report income (solid black line), and the estimate for the both samples merged (grey line). Left: for $f(x|x)$ with homoscedasticity; Right: with heteroscedasticity.

We can now examine the consequences for the Lorenz curve and Gini coefficient that result from the missings. The left panel of Figure 4 displays the three Lorenz curves: one for $\mathcal{S}$, for which income values are available, one computed using $\hat{f}_t(y)$ for the sample $\mathcal{T}$, and one based on the mean regression predictions. As expected the last one is a very poor estimate despite the fact that it provides the best estimates of the income of individual households. The Lorenz for $\mathcal{S}$ and its estimate for $\mathcal{T}$ differ only slightly, but the Lorenz curve of the joint $(\mathcal{S}+\mathcal{T})$ does clearly deviate from the one we obtained when neglecting the missings (only

12 We also experimented with a semiparametric estimator for equation (4), in particular a two-step estimator with $\mu(x)=x'\beta$, $\nu(z)=z'\gamma$ but a nonparametric $\lambda(\cdot)$. This enabled us to check whether the impact of the distributional assumption was substantial, but it turned out not to be the case.

taking $\mathcal{S}$). In contrast to what is often reported for surveys in wealthy, industrialized countries, our estimates suggest that, on average, the poorer rather than the wealthier households are less likely to report their income. Being able to compute confidence bands for the Lorenz curve would be useful for assessing whether that difference could be attributable to sampling variations. For a Lorenz curve computed from purely parametric models the standard bootstrap is available; one simply carries out the estimation for a number of random samples of size n from the original sample (with replacement). For more complex models the nonparametric bootstrap and alternative sampling methods are available cf. Politis et al. (1999). For bootstrap inference in semiparametric models we refer to Härdle et al. (2004), and for mixed effects or small area models to Lombardía and Sperlich (2008). The right-hand panel of Figure 4 shows a 99 % confidence bands for the Lorenz curve for the population $\mathcal{S}+\mathcal{T}$.[13]

Figure 4: Estimates of the Lorenz curve and the 99 % point-wise confidence intervals.



Notes: Left panel: The Lorenz curve for $\mathcal{S}$ (black line), its estimate for $\mathcal{T}$ (dotted black line), and the Lorenz curve for the entire population $\mathcal{S}+\mathcal{T}$ (grey line). Right panel: The Lorenz curve for the entire survey with 99 % point-wise confidence intervals.

13 The construction of confidence bands for Lorenz curves computed from simulation-based estimates $f_r(y)$ is not as straight-forward. Bootstrap methods have been used in the literature to compute confidence intervals in such cases, but the intervals reflect only the uncertaity due to sampling variation in the *artificially generated residuals* rather than the uncertainty due to variation of simulation plus estimation.

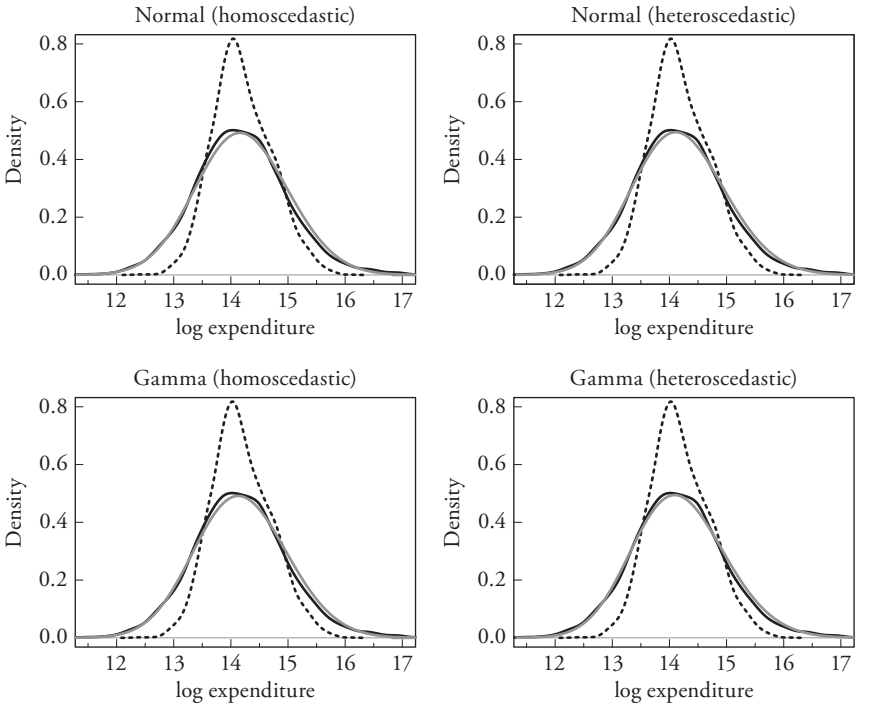
As pointed out in the introduction, measures of inequality (such as the Gini coefficient or selected quantiles of distribution of the welfare indicator), are sensitive to the changes in tails of the distribution. We examined the performance of estimates of the Gini coefficient computed from(2). The Gini coefficient for $\mathcal{S}$ is 0.579; our estimate for $\mathcal{T}$ is 0.587; it is as low as 0.379 when computed using the mean regression predictions. Our estimate for all the households in $\mathcal{S}$ and $\mathcal{T}$ combined is 0.586 with a 90 % bootstrap confidence interval of [0.580,0.595]. Note that the confidence interval is not symmetric (since the distribution of the estimator of the Gini coefficient is not). Note also that the Gini coefficient calculated only from the reported incomes suggests a lower level of inequality that is not even inside the above 90 % confidence interval.

3.3 Predicting the Expenditure Distribution

We now illustrate the application of the proposed estimator for prediction rather than for estimation. This could be done for a purely hypothetical population but here we predict the distribution of the log(expenditure) of the 2000 cohort (of the IFLS data) using information from the 1997 cohort. For the purpose of illustration we will predict the 2000 distribution only for the subpopulation of 4,585 households for which expenditure information is available because this allows us to compare the predictor with the observed distribution $f_t(y)$ in 2000. The assumption being made here is that the estimated conditional distribution, $\hat{f}(y \mid x)$ for the subpopulation of 5,406 households that reported expenditure 1997 is applicable for the (inflation-adjusted) subpopulation of 4,585 households in 2000. Specifically, to make expenditure in 1997 and 2000 comparable we adjusted for inflation using the regional inflation rates (available on request). These were obtained from the Badan Pusat Statistic consumer price index (CPI) reported for 45 cities in Indonesia. For provinces with more than one city we used a simple average of the price indices for those cities; cf. Chaudhuri et al. (2002). We will show the results for only fur different models for $f_t(y|x)$, all of which use linear regression (Table 7) for the mean and constant CoV. These contrast the impact of using different distributional assumptions (normal vs. gamma, homoscedasticity vs. heteroscedasticity).

As indicated in Section 2 one can make adjustments to the estimate of $f(y)$ if additional information is available. For example, the real GDP per capita provided by the WDI in 2003 indicates that there was a decline of 11.97 % between 1997 and 2000. If we assume that the average household expenditure decreased by the same percentage then we can calibrate $\hat{f}(y)$ accordingly, as was done for the estimates displayed in Figure 5.

Figure 5: A kernel smooth of the *observed* log(expenditure) in 2000



Notes: A kernel smooth of the *observed* log(expenditure) in 2000, i.e. $f_t(y)$ (solid black line) and two predictors of it: a (GDP per capita calibrated) estimate using the proposed estimator (grey line) and a kernel smoothed estimate of the mean regression predicted values (dotted line).

For computing $\hat{f}(y)$ we used a Gaussian kernel smooth of the observed log-expenditures in 2000 using Silverman's rule-of-thumb bandwidths. The relative accuracy of the four predictors based on $\mathcal{S}$ (i.e. information taken from 1997) was assessed using the ISE. For the normal and gamma homoscedastic models they were 0.0012 and 0.0010, respectively; and 0.0008 and 0.0007 for the heteroscedastic versions. The values in this example suggest that the estimates are less sensitive to distributional assumptions (normal vs. gamma) than to the choice of homoscedastic or heteroscedastic model.

The predicted Lorenz curves is almost indistinguishable from the true Lorenz curve computed from the actually observed expenditures in 2000. The resulting predicted value of the Gini coefficient for 2000 is 0.447 (the 90 % bootstrap

confidence interval is [0.429,0.455]); the true value was 0.451. These results, and those that follow, are for the heteroscedastic gamma model, but they differ very little from those obtained using the heteroscedastic normal distribution.

One of the important uses of the distribution of welfare indicators is to monitor poverty and vulnerability to poverty. There exist different definitions of poverty but we will consider only the simple definition: "the expenditures of the household fall below a given level". There are also different definitions of this level (the poverty line); it can be defined in absolute or relative terms. An important statistic in poverty assessment is the number (or percentage) of households that are classified as poor. That statistic is sensitive to errors in the mean of the predicted distribution of the welfare indicator if one uses an absolute poverty line; a small shift in the location of the predicted distribution can lead to a substantial change in the resulting statistic. Estimates based on a relative poverty line are less sensitive in that respect.

We examine the accuracy of that statistic using a relative poverty line, namely 40 % of the median expenditure. The resulting poverty line (in log Rps pa.) is at about 13.215 in 1997 and 13.205 in 2000. Table 1 shows the observed and predicted counts (and percentage) of poor and non-poor households in $\mathcal{T}$ in 2000 as well as the 90 % confidence intervals. The predicted counts were computed using the calibrated prediction of $f_t(x)$. The confidence intervals cover the true counts.

Table 1: Counts (and percentage) of poor and non-poor households in $\mathcal{T}$ in 2000

	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%)

Notes: Number (and percentage) of households in $\mathcal{T}$ that were poor/non-poor in 2000, and predictions using data from 1997.

The proposed estimator is not specifically designed for predicting the y-values for the individual households in $\mathcal{T}$. In any case the best predictors of those are the mean regression predictions, $\{\hat{y}_j = \hat{\mu}(x_j)\}_{j=1}^m$, despite the fact that, collectively, they provide a poor prediction of the distribution of y in the target population. The following transformation of the $\hat{y}_j$-values projects the 'position' of each household in $\mathcal{T}$ in the predicted distribution $\hat{f}(y)$:

$$\hat{y}'_j = \widehat{F}^{-1}(\widehat{F}^{*}(\hat{y}_j)) \text{ for } j = 1,2,\ldots,m, \tag{7}$$

where $\hat{F}^{*}(\cdot)$ is the empirical distribution function of the mean regression predictions; $\hat{F}(\cdot)$ is the distribution function corresponding to $\hat{f}(y)$ and can be computed by numerical integration of $\hat{f}(y)$. The inverse $\hat{F}^{-1}(\cdot)$ can be computed using, e.g., interpolation.

We emphasize that $\hat{y}'_j$ is not the best predictor for the missing value y_j ($\hat{y}_j = \hat{\mu}(x_j)$) but it is superior to $\hat{y}_j$ for the purpose of assessing whether y_j falls above or below the poverty line, i.e. whether household j is poor or non-poor. This point is verified by comparing the accuracy of the predicted counts (for 2000) in Table 1.

4. Predicting the Distribution of the Number of Visits to the Doctor

This section illustrates the application of the proposed method for discrete-valued variables.

4.1 Data and Model Selection

We have records of the 23,607 inhabitants of the Sydney suburb Ryde in 1994 and 1995. The available information comprises age, gender, and the number of doctor visits for both years.[14]

We consider two typical prediction problems. Practitioners usually only have access to data on a sample from which to estimate the distribution of number of visits for the population. In other cases that distribution needs to be predicted for a hypothetical future population. The latter is particular important for formulating public health policy in countries where substantial demographic change is expected to occur.

To illustrate the application of the method in these two situations we drew a random sample, say $\mathcal{S}$, of 200 observations from 1994. The age and gender for all individuals in the population were taken as known, but the number of visits was assumed to be known only for the 200 individuals in the sample. Summary

14 For a more detailed description of these data see Heller (1997). In the original data set 11 individuals reported more than 100 visits. These excessively high counts were attributed to misuse of the health insurance cards and, as it was not possible to obtain reliable corrections, we decided to truncate the data at a maximum of 100 visits.

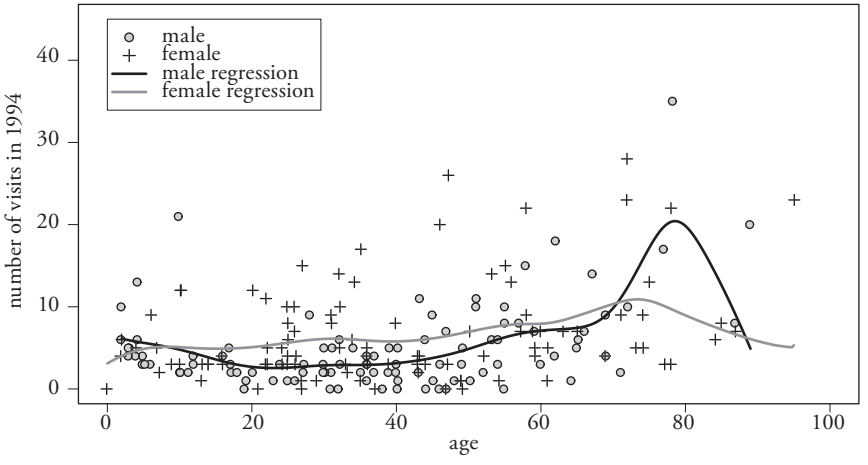
statistics are given in Table 2. The aim is to estimate the distribution of the number of visits in 1994, and afterward to predict it for 1995.

Table 2: Summary statistics, standard deviations in parentheses.

	population		sample	
	men	*women*	*men*	*women*
number of individuals	11302	12305	101	99
average age in 1994	36 (22)	39 (24)	37 (21)	40 (23)
average age in 1995	37 (22)	40 (24)	–	–
average number of visits in 1994	5.2 (6.3)	6.9 (7.2)	5.0 (5.3)	6.8 (6.1)
average number of visits in 1995	5.6 (6.5)	7.2 (7.2)	–	–

Plots of the number of visits against age, separately for male and female, are shown in Figure 6. Simple local linear regression estimates (the lines) indicate a non-linear relationship between the mean number visits with age and gender. The lines for males and females differ and there is evidence of overdispersion.

Figure 6: The number of visits to a GP against age



Notes: The number of visits to a GP (left, in 1994; right, in 1995) plotted against age for a simple random sample of 200 residents in Ryde. Local linear regression estimate with cross-validation bandwidth $\hat{h}_{CV} = 2.78$ (black line, male) and $\hat{h}_{CV} = 2.78$ (grey line, female).

Given that the observations are discrete-valued we start with the negative binomial (NB)

$$f(y \mid \mu,\sigma) = \begin{cases} \dfrac{\Gamma(y+1/\sigma)}{\Gamma(y+1)\Gamma(1/\sigma)} \dfrac{(\mu\sigma)^y}{(\mu\sigma+1)^{(y+(1/\sigma))}} & \text{if } x = 0,1,\ldots \\ 0 & \text{otherwise} \end{cases}$$

with mean μ, variance $\mu+\mu^2\sigma$, where σ is a scalar to be estimated. In order to check whether the overdispersion is attributable to zero inflation, we also consider zero inflated Poisson (ZIP) and zero inflated NB. To compare these three models we calculate the log-likelihood (llh), the deviance difference ΔD (relative to the simple Poisson) and the AIC of the fitted models. The results are listed in Table 3 separately for males and females. The ZIP fits is slightly better than the Poisson model (not shown) but substantially worse than the NB. The zero-inflated NB shows no improvement compared to the NB.

Table 3: Quality of fit statistics using GLM

sex	*Model*	*Link*	*Terms*	*llh*	ΔD	*AIC*
males	ZIP	$\log(\mu)$	$age+age^2$	–294	–	596
	NB	$\log(\mu)$	$age+age^2$	–253	82	515
	zero-inflated NB	$\log(\mu)$	$age+age^2$	–253	82	517
females	ZIP	$\log(\mu)$	$age+age^2$	–360	–	728
	NB	$\log(\mu)$	$age+age^2$	–286	148	579
	zero-inflated NB	$\log(\mu)$	$age+age^2$	–286	148	581

The generalized linear models considered above allow only the location parameter to depend (parametrically) on the covariates. These restrictions can be relaxed in two important ways using the Generalized Additive Model for Location, Scale and Shape (GAMLSS) developed by RIGBY and STASINOPOULOS (2005). First, these models allow each of the distribution parameters to depend on covariates. Second, the parameter functions may include random effects, or may even be nonparametric (though always of an additive structure).[15] In our case we can

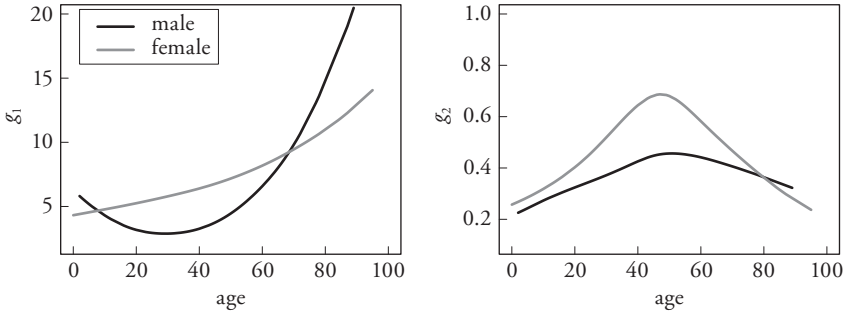
15 An alternative option is to model the dispersion in parametric or nonparametric NB regression using the Vector Generalized Additive Model introduced by YEE and WILD (1996); for an application see BERZEL et al. (2006).

model both the mean and the dispersion parameter as functions of *age* by parameterizing $f(y|\mu,\sigma)$ using

$$\begin{aligned} \log(\mu) &= g_1(age), \\ \log(\sigma) &= g_2(age), \end{aligned} \tag{8}$$

where g_1, g_2 are quadratic functions, or alternatively, nonparametric cubic splines (cs). Estimates obtained for the latter case are displayed in Figure 7.

Figure 7: Impact of age and gender on the GAMLSS nonparametric regression estimates, separately for mean and dispersion



Notes: Impact of age and gender on the GAMLSS nonparametric regression estimates for mean g_1 (left) and dispersion g_2 (right), based on a random sample of 200 residents in Ryde in 1994.

Table 4 gives the fitted global deviances GD, the AIC and the Schwarz Bayesian criterion (SBC) that were used to compare the two GAMLSS models.

One could also apply non- and semiparametric specification tests (see González-Manteiga and Crujeiras (2013) for a recent review) but conclusions based on these test would be applicable to $\mathcal{S}$ whereas our objective is to make predictions about $\mathcal{T}$. Furthermore non- and semiparametric specification tests suffer in practice from calibration problems, especially for $dim(X) \geq 2$; see Sperlich (2014). In this application we therefore prefer to base selection on the AIC and SBC which provide (asymptotic) estimates for the quality of the fit in a new data set.

Table 4: Quality of fit statistics using GAMLSS

sex	*Model*	*Link*	*terms*	*GD*	*AIC*	*SBC*
males	NB (parametric model)	$\log(\mu)$ $\log(\sigma)$	$age + age^2$ $age + age^2$	506	518	533
	NB (nonparametric model)	$\log(\mu)$ $\log(\sigma)$	$cs(age)$ $cs(age)$	502	515	534
females	NB (parametric model)	$\log(\mu)$ $\log(\sigma)$	$age + age^2$ $age + age^2$	568	580	596
	NB (nonparametric model)	$\log(\mu)$ $\log(\sigma)$	$cs(age)$ $cs(age)$	568	580	595

The criteria in Tables 3 and 4 indicate that the AIC selects the NB generalized linear model (GLM) throughout. Nevertheless, in what follows we will continue to consider the GAMLSS models for the following reason: The criteria considered so far assess the fit of the *conditional distribution* $f(y|\mu,\sigma)$ but our objective is to estimate the *unconditional distribution* of number of visits, $f(y)$, and it is not clear that the model selected as best for the former will also be best for estimating the latter. This issue is investigated in the next section. Figure 7 suggests that dispersion parameter is neither constant nor linear.

4.2 Predicting the Population Distribution

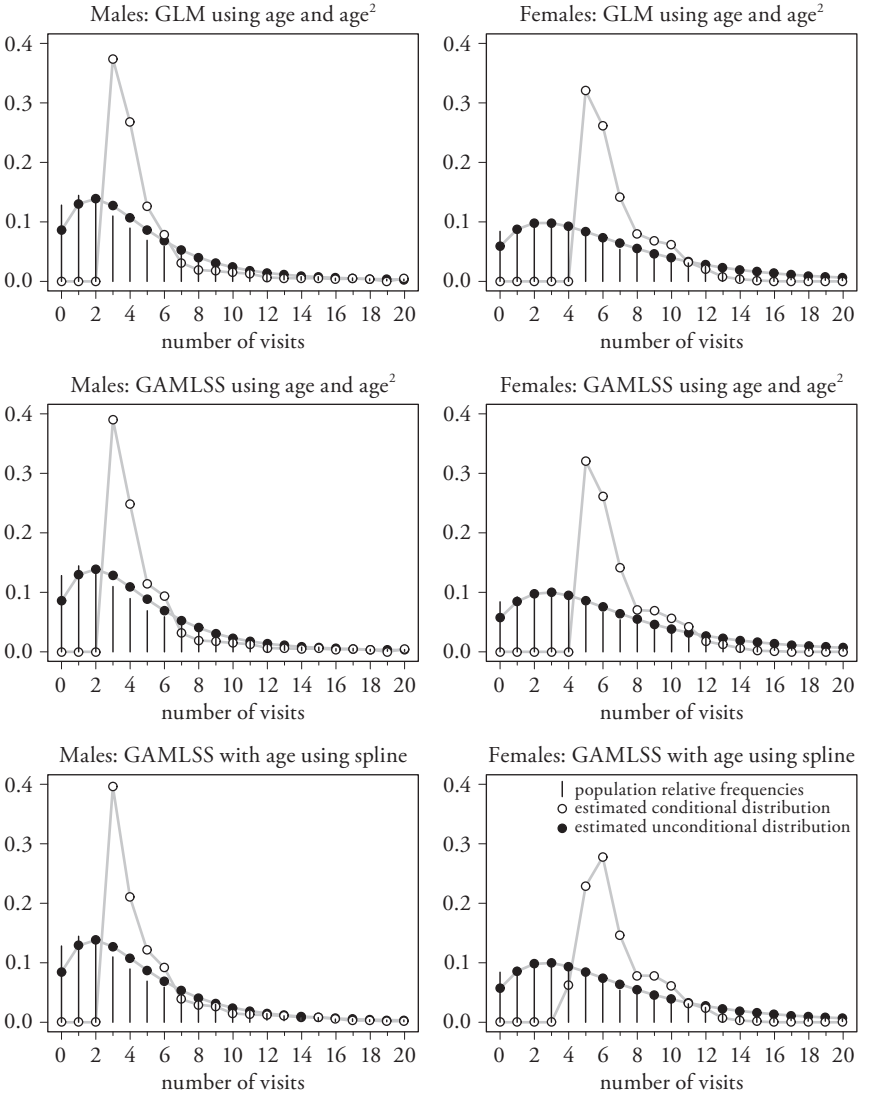
We apply our method to two prediction problems: one is using a random sample of 101 males and 99 females, in 1994 to estimate the distribution for number of visits for the (male and female) population in Ryde in the same year, and the other is to predict the distribution of the number of visits in 1995 using the (same) 1994 sample.

4.2.1 Estimating the Distribution of Number of Doctor Visits

Figure 8 displays the true distribution of number of visits for the population in 1994 together with the fits using three NB model specifications for g_1, g_2 and fitted using the sample from $\mathcal{S}$ from the same year. The estimate of the target distribution is thus of the form

$$\hat{f}(y) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{NB}[y \mid \hat{g}_1(x_i^t), \hat{g}_2(x_i^t)]. \tag{9}$$

Figure 8: Predicted population distribution for 1994



Notes: Predicted population distribution based on NB GLM estimates (upper), NB GAMLSS parametric (middle) and cubic spline (lower) specification for 1994. The *conditional distribution estimate*, in analog to the last section, has been calculated as follows: you predict for each individual i its conditional outcome by $\hat{y}_i = round\{\hat{E}[Y \mid x_i^t]\}$ and take the resulting frequencies, where the conditional expectation results from $\hat{g}_1$.

Since we have records of the number of visits for the entire population, $\mathcal{T}$, we can precisely quantify the accuracy of the estimate. The distribution obtained from estimated conditional means (circles) is clearly hugely biased (too narrow) to be of use as an estimate of the target distribution. In contrast, the predicted unconditional distribution (solid circles) fits very well.[16]
To quantify the accuracy of the estimates for different model specifications we used

$$\frac{1}{J}\sum_{j=1}^{J} LOSS\left[\hat{f}\left(y_j\right) - f_t\left(y_j\right)\right], \tag{10}$$

where J is the number of different y values (42 for males, 34 for females), and $LOSS[\cdot]$ was taken as the absolute value (L1-norm) or the squared value (L2-norm). The results are listed in Table 5. The NB GAMLSS using splines performs best for both males and females. It might however be surprising that for males it does much better than GLM although the AIC was the same (515 for both CS-GAMLSS and GLM). We note here that the determination of the degrees of freedom that are needed to compute the AIC can be quite problematic in nonparametric contexts, see e.g. Sperlich et al. (1999). Finally note that the zero-inflated NB GLM never outperformed the NB GAMLSS using splines.

Table 5: L1 and L2-Norm prediction errors of case 1

	L1-Norm		L2-Norm	
Model	*males*	*females*	*males*	*females*
NB GLM	.02480	.03558	.00385	.00451
zero-inflated NB	.02481	.03558	.00390	.00451
NB GAMLSS	.02483	.03531	.00389	.00448
NB GAMLSS using spline	.02334	.03193	.00371	.00346

4.2.2 Predicting the distribution of the number of future visits to the doctor

More challenging – and also more interesting for decision-making regarding public health – is predicting the distribution of the number of future visits to a GP. Here one must assume that the relationship between y and the covariates

16 The fit for the case of males might be improved if we were to allow for zero inflation.

will remain unchanged and thus the prediction performance will depend on the degree to which this assumption is violated.

To illustrate the method we use the same (1994) sample of size 200 that we used in Section 4.1, and the same procedure, to now predict the distribution of number of visits in 1995. The resulting predictions are displayed in Figure 9. The prediction performance looks even better than the estimation performance in Section 4.2.1. The improvement can be attributed to the lack of zero inflations in the 1995 (population) visits.

As in Section 4.2.1 we analyzed the prediction errors for our different specifications; see Table 6. Again, the semiparametric NB GAMLSS clearly gives the best predictions for both the male and female population. The ranking of the specifications is the same as before but, as is apparent from Figure 9, the overall prediction accuracy is even better.

Table 6: L1 and L2-Norm prediction errors of case 2

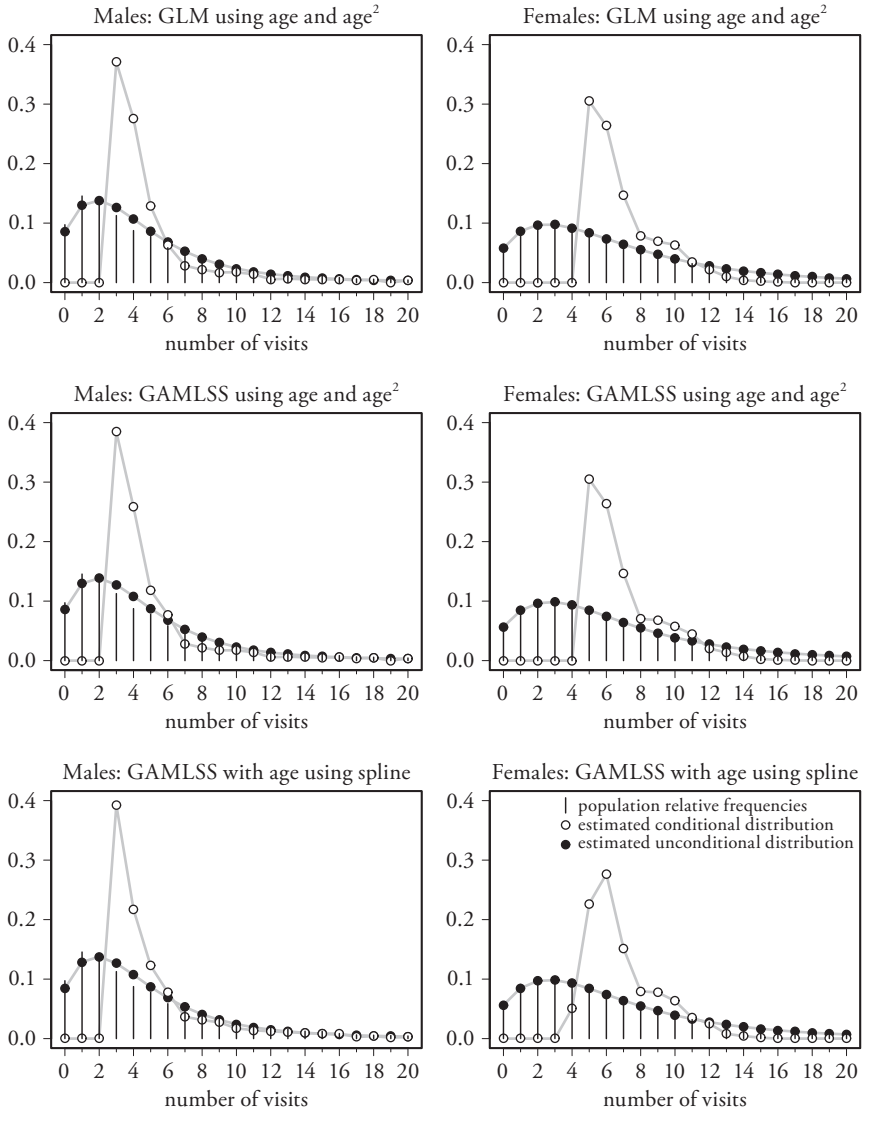
	L1-Norm		L2-Norm	
Model	*males*	*females*	*males*	*females*
NB GLM	.02402	.03407	.00367	.00411
zero-inflated NB	.02403	.03407	.00371	.00411
NB GAMLSS	.02405	.03369	.00368	.00409
NB GAMLSS using spline	.02238	.03110	.00348	.00325

5. Conclusions

Mean regression predictions can lead to severely biased estimates of the marginal distribution and hence of the statistics derived from them; the estimated density function is 'too narrow'. Simulation-based methods that add random noise to such predictions were developed to overcome that problem. We believe that the method proposed here is more appropriate in that it is analytic and does not involve the use of artificial random terms. That makes the estimates reproducible and it simplifies further inference.

Our estimator can be viewed in a number of different ways. From a Bayesian perspective the distribution that we are estimating is a random function that can be described by its moment functions together with appropriate prior conditional distributions. From a frequentist point of view we are modeling the

Figure 9: Predicted population distribution for 1995



Notes: Predicted population distribution based on NB GLM estimates (upper), NB GAMLSS parametric (middle) and cubic spline (lower) specification for 1995.

target distribution by an m-fold mixture of models whose parameter functions are fitted using regression techniques. As a special case we can even regard the estimator in terms of a nonparametric kernel density in which the assumed conditional distribution is the kernel, and the scedasticity function determines the data-adaptive local bandwidth. From that perspective assuming homoscedasticity corresponds to using a common global bandwidth; the use of asymmetric conditional distributions corresponds to the case of applying special kernels typically used for boundary correction or asymmetric information (like the knn estimators). In each of these three interpretations the choice of the conditional density plays a relatively minor role, the scedasticity function is more important, and the choice of model for the mean regression mainly impacts on the variability of the final estimate.

The method is applicable to a wide range of applications and models including parametric, nonparametric and semiparametric, selectivity correction or mixed effects models for cross section, panel, times series, or a combination of these. It covers discrete distributions by replacing densities with probability functions. Inference can be based on resampling methods. Given an underlying model it provides the researcher with clear interpretations from which to draw conclusions.

We illustrated the application and practical usefulness of the estimator in different contexts: data matching from one sample to another, the completion of surveys with (possibly endogenous) missing values, and the prediction of a target distribution from past values. One could add survey-to-census, cross-survey or cross-country data matching, or scenarios for the prior evaluation of treatment effects.

Specifically, we apply it to two settings: income distribution and doctor visits. We find the proposed method is simple to apply, especially as the functions needed for its implementation are available in software packages such as gretl, R, SAS or Stata, thereby placing it within easy reach of practitioners and empirical researchers. The R-programs we have written and used, as well as the data used in our applications are available on the web-page of the Swiss Journal of Economics and Statistics.

Appendix

Table 7: Estimated regression coefficients for log(income) under different models.

	Application 1		Application 2		Application 3
Model	Linear	APLM	Selection	Main	Linear
Constant	10.957 (.2529)		2.513 (.2027)	10.97 (.1832)	11.77 (.1251)
Average age	.0415 (.0102)		–.0271 (.0081)	.0421 (.0071)	.0231 (.0042)
Average age squared	–.0006 (.0001)		.0002 (.0000)	–.0006 (.0000)	–.0003 (.0000)
Average year of schooling	.0400 (.0053)		–.0084 (.0041)	.0366 (.0037)	.0488 (.0025)
Log of assets p.c.	.2261 (.0122)		–.0178 (.0091)	.2301 (.0084)	.1567 (.0061)
Share of asset to business	.4672 (.0808)		.1016 (.0644)	.4870 (.0583)	.0289 (.0438)
Farmer in family	–.2351 (.0489)	–.2102	.1199 (.0378)	–.2409 (.0363)	–.2011 (.0247)
Share of working hhm	1.5472 (.0977)		–.9768 (.0755)	1.427 (.0951)	–.1082 (.0562)
Share of female hhm	–.6385 (.1037)		.0207 (.0803)	–.5608 (.0727)	–.0521 (.0612)
HH size	–.0685 (.0099)		–.2477 (.0062)	–.0858 (.0161)	–.0567 (.0042)
Urban area	.2871 (.0430)	.2717	.0130 (.0327)	.2579 (.0303)	.1604 (.0212)
North Sumatera	–.2365 (.0894)	–.2162	.1508 (.0690)	–.1020 (.0619)	–.3990 (.0463)
West Sumatera	–.0194 (.1102)	.0056	–.2052 (.0794)	–.0216 (.0784)	–.1974 (.0578)
South Sumatera	–.1287 (.0942)	–.0738	.1574 (.0760)	–.0803 (.0682)	–.3636 (.0561)
Lampung	–.4017 (.0956)	–.3702	.2332 (.0774)	–.3995 (.0685)	–.2568 (.0556)
West Java	–.2611 (.0678)	–.2206	–.0044 (.0520)	–.2215 (.0470)	–.2754 (.0415)

	Application 1		Application 2		Application 3
Model	Linear	APLM	Selection	Main	Linear
Central Java	–.6629 (.0739)	–.6095	.0730 (.0560)	–.6223 (.0519)	–.3456 (.0425)
Yogyakarta	–.6850 (.1022)	–.6261	–.2391 (.0750)	–.6432 (.0746)	–.5030 (.0534)
East Java	–.5239 (.0723)	–.4890	–.1451 (.0534)	–.4921 (.0508)	–.6584 (.0417)
Bali	–.4119 (.0931)	–.4343	.0884 (.0735)	–.3727 (.0664)	–.4339 (.0500)
West Nusa Tenggara	–.5801 (.0846)	–.5296	.1338 (.0666)	–.5251 (.0590)	–.3546 (.0482)
South Kalimantan	–.0314 (.0965)	.0295	–.1433 (.0739)	–.0163 (.0697)	–.1474 (.0525)
South Sulawesi	–.6058 (.1068)	–.5632	–.1285 (.0772)	–.5953 (.0757)	–.6501 (.0493)
Head of the family			–.0844 (.0345)		
Marriage			.5856 (.0336)		
rho * sigma				.1291 (.0982)	
Number of observations	2783	2783	5567	5567	5406

Notes: Estimated regression coefficients for log(income) under different models. Estimated standard errors are given in parentheses. For the APLM only the coefficients for the parametric part are given (without standard errors).

References

Atkinson, Anthony B., and François Bourguignon (2000), *Handbook of Income Distribution*, Amsterdam: North-Holland.

Azzarri, Carlo, Gero Carletto, Benjamin Davis, and Alberto Zezza (2006), "Monitoring Poverty without Consumption Data", *Eastern European Economics* 44(1), pp. 59–82.

Berzel, Andreas, Gillian Z. Heller, and Walter Zucchini (2006), "Estimating the Number of Visits to the Doctor", *Australian & New Zealand Journal of Statistics* 48, pp. 213–224.

Biewen, Martin, and Stephen P. Jenkins (2005), "A Framework for the Decomposition of Poverty Defferences with an Application to Poverty Defferences Between countries", *Empirical Economics* 30, pp. 331–358.

Birkin, Mark, and Martin Clarke (1989), "The Generation of Individual and Household Incomes at the Small Area Level Using Synthesis", *Regional Studies* 23(6), pp. 535–548.

Chaudhuri, Shubham, Jyotsna Jalan, and Asep Suryahadi (2002), "Assessing Household Vulnerability to Poverty from Cross-Sectional Data: A Methodology and Estimates from Indonesia", Discussion Paper Series, Department of Economics, Columbia University.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly (2013), "Inference on Counterfactual Distributions", *Econometrica* 81(6), pp. 2205–2268.

Chotikapanich, Duangkamon (2008), *Modeling Income Distributions and Lorenz Curves*, Series: Economic Studies in Inequality, Social Exclusion and Well-Being 5, Springer.

Davis, Benjamin (2003), *Choosing a Method for Poverty Mapping*, Food and Agriculture Organization of the United Nations, Rome, www.fao.org/docrep/005/y4597e/y4597e00.htm.

DiNardo, Jone, Nicole M. Fortin, and Thomas Lemieux (1996), "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach", *Econometrica* 65, pp. 1001–1046.

Donald, Stephen G., Yu-Chin Hsu, and Garry F. Barrett (2012), "Incorporating Covariates in the Measurement of Welfare and Inequality: Methods and Applications", *Econometrics Journal* 15, pp. C1–C30.

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw (2003), "Micro-Level Estimation of Poverty and Inequality", *Econometrica* 71(1), pp. 355–364.

Filmer, Deon, and Lant H. Pritchett (2001), "Estimating Wealth Effects without Expenditure Data – Or Tears: An Application to Educational Enrollments in States of India", *Demography* 38(1), pp. 115–132.

Gasparini, Leonardo, Martín Cicowiez, Federico Gutierrez, and Mariana Marchionni (2003), "Simulating Income Distribution Changes in Bolivia: A Microeconometric Approach", The World Bank Bolivia Poverty Assessment.

González-Manteiga, Wenceslao, and Rosa M. Crujeiras (2013), "An Updated Review of Goodness-of-Fit Tests for Regression Models", *Test* 22(3), pp. 361–411.

Härdle, Wolfgang, Sylvie Huet, Enno Mammen, and Stefan Sperlich (2004), "Bootstrap Inference in Semiparametric Generalized Additive Models", *Econometric Theory* 20, pp. 265–300.

Heller, Gillian Z. (1997), "Who Visits the GP? Demographic Patterns in a Sydney Suburb", Technical report, Department of Statistics, Macquarie University.

Hentschel, Jesco, Jean Olson Lanjouw, Peter Lanjouw, and Javier Poggi (2000), "Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador", *World Bank Economic Review* 14(1), pp. 147–165.

Horton, Nicholas J., and Stuart R. Lipsitz (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables", *The American Statistician* 55(3), pp. 244–254.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce (1993), "Wage Inequality and the Rise in Returns to Skill", *The Journal of Political Economy* 101(3), pp. 410–442.

Little, Roderick J. A., and Donald B. Rubin (2002), *Statistical Analysis with Missing Data (Second Edition)*, John Wiley, New York.

Lombardía, María J., and Stefan Sperlich (2008), "Semiparametric Inference in Generalized Mixed Effects Models", *Journal of Royal Statistical Society: Series B* 70(5), pp. 913–930.

McLachlan, Geoffrey, and David Peel (2000), *Finite Mixture Models*, Wiley Series in Probability and Statistics.

Melly, Blaise (2005), "Decomposition of Differences in Distribution Using Quantile Regression", *Labour Economics* 12(4), pp. 577–590

Mishra, Satish C. (2009), "Economic Inequality in Indonesia: Trends, Causes, and Policy Response", Strategic Asia, commissioned by UNDP Regional Office, Colombo.

Noufaily, Angela, and M. C. Jones (2013), "Parametric Quantile Regression Based on the Generalized Gamma Distribution", *Journal of the Royal Statistical Society, Series C, Applied Statistics* 62(5), pp. 723–740.

Paulin, Geoffrey D., and David L. Ferraro (1994), "Imputing Income in the Consumer Expenditure Survey", *Monthly Labor Review* 117(12), pp. 23–31.

Politis, Dimitris N., Joseph P. Romano, and Michael Wolf (1999), *Subsampling*, Springer, New York.

Ravallion, Martin (2001), "Growth, Inequality and Poverty: Looking Beyond Averages", *World Development* 29(11), pp. 1803–1815.

Rigby, R. A., and Stasinopoulos, D. M. (2005), "Generalized Additive Models for Location, Scale and Shape", *Applied Statistics* 54, pp. 507–554.

Rothe, Christoph (2010), "Nonparametric Estimation of Distributional Policy Effects", *Journal of Econometrics* 155, pp. 5670.

Royston, Patrick (2004), "Multiple Imputation of Missing Values", *The Stata Journal* 4(3), pp. 227–241.

Sahn, David E., and David C. Stifel (2000), "Poverty Comparison Over Time and Across Countries in Africa", *World Development* 28(12), pp. 2123–2155.

Sperlich, Stefan, Oliver B. Linton, and Wolfgang Härdle (1999), "Integration and Backfitting Methods in Additive Models – Finite Sample Properties and Comparison", *Test* 8, pp. 419–458.

Sperlich, Stefan (2014), "On the Choice of Regularization Parameters in Specification Testing: a critical discussion", *Empirical Economics* 47, pp. 427–450.

Stock, sc James H. (1989), "Nonparametric Policy Analysis", *Journal of the American Statistical Association* 84(406), pp. 567–575.

Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima (2011),"Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box", *Journal of Statistical Software* 45(2), pp. 1–31.

Tarozzi, Alessandro, and Angus Deaton (2009), "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas", *Review of Economics and Statistics* 91(4), pp. 773–792.

Van Kerm, Philippe (2013). "Generalized Measures of Wage Differences", *Empirical Economics* 45(1), pp. 465–482.

Yee, T. W., and C. J. Wild (1996), "Vector Generalized Additive Models", *Journal of Royal Statistical Society, Series B, Methodological* 58, pp. 481–493.

Zeller, Manfred, Julia Johannsen, and Gabriela Alcaraz V. (2005), "Developing and Testing Poverty Assessment Tools: Results from Accuracy Test in Peru", College Park, IRIS Center, University of Maryland.