Simple Solutions to
hard Problems in the Estimation and Prediction
# of Welfare Distributions

Dissertation
Presented for the Degree of Doctor of Economics
at the Faculty of Economic Sciences
of the Georg-August University of Göttingen

by
Jing Dai
from
Hunan, China

Göttingen, 2011

First Examiner:	Prof. Dr. Stefan Sperlich
Second Examiner:	Prof. Dr. Walter Zucchini
Third Examiner:	Prof. Dr. Inmaculada Martinez-Zarzoso
Disputation:	08 April 2011

Acknowledgement

There is a Chinese proverb that says, “The key to mastering any new task is finding the right mentor.”

When I began my doctoral research, the field of statistics and econometrics was really a “new task” for me. I therefore consider myself extremely lucky to have found Prof. Stefan Sperlich as my supervisor. In the period of his supervision he has demonstrated a deep sense of responsibility towards me. I have benefitted greatly from his experience and expertise in nonparametric statistical techniques and its applications in econometrics.

My special thanks goes to Prof. Walter Zucchini for his invaluable ideas, important comments and patience. He has set a great example of the scientific approach to solve research problem. I shall profit from his influence for life.

I thank Prof. Fred Böker who helped me a lot on my teaching work, and thus allowed me more time to work on my dissertation. Many thanks also go to my examiner Prof. Dr. Inmaculada Martinez-Zarzoso. Besides I thank all my colleagues at the Institute for Statistics and Econometrics for the valuable discussions and the friendly help. I thank Ta-chao and Ren for proof-reading and my officemate Duygu for the nice conversations.

My immense gratitude goes to my mom, Yingjian Kang and my sister, Na Dai. Their love is always the most important motivation for the completion of this dissertation. I am also intensely grateful to my husband Zheng Wan, who is always there to clear my mind, to inspire me with confidence and to provide me with constructive advices.

Without the loving support and encouragement, I can not adhere to come this far, and finally finish this work.

Contents

List of Figures

Boundary correction for kernel density estimation and regression 17

Predicting household expenditure and income distribution for welfare analysis 47

Estimating and predicting the distribution of the number of visits to the medical doctor **85**

Estimating the income distribution from a few quantiles **107**

List of Tables

Boundary correction for kernel density estimation and regression 17

Predicting household expenditure and income distribution for welfare analysis 47

Estimating and predicting the distribution of the number of visits to the medical doctor 85

Estimating the income distribution from a few quantiles **107**

Introduction and overview

The aim of my PhD projects is to investigate various problems in applied parametric and nonparametric estimation, and eventually in prediction. The main focus is on welfare analysis. The three major research objectives addressed are: (i) to overcome boundary effects in nonparametric density estimation and regression, (ii) to estimate and predict population distributions via data matching, and (iii) to construct a income distribution estimator from a few quantiles. The problems that we have dealt with are not new, even in the field of econometrics. However, in some specific application areas, new challenges are often presented to methodologies that have not been studied in (mathematical) statistics and are, unfortunately, probably not even known. I will highlight the three specific problems that are considered in my dissertation.

- Boundary correction. The reason why we are looking for a boundary correction method is that the application of kernel density estimation and regression often experience difficulties at the boundaries. For both kernel density estimation and regression, however, quite often, our interests are right up to the boundaries. For instance, if we are interested in poverty and inequality, it is necessary to have reliable estimates of the income distribution at the left tail i.e. near zero. Similarly, those interested in risk assessment looked at the performance of especially young or old, highly or poorly educated, compared large with small companies, etc.. These are all potential users of boundary correction methods, as they will definitely face problems with boundaries. The so-called boundary effect, i.e. the bias and variance increase due to one side data information, has been well studied in the literature. The methods used most often for boundary correction are the linear correction for density estimation (see Gasser and Müller, 1979; Gasser et al., 1985; Jones, 1993) as well as the local polynomial approaches, which were first applied in density estimation by Fan and Gijbels (1992), and later on improved by Cheng et al. (1997) with an optimal weighting. In many situations local polynomials are certainly an attractive remedy for boundary effects in regression, since it would automatically correct the boundary effects. Another option is to modify the bandwidth towards the boundaries, including Rice (1984), Gasser et al. (1985) and Müller (1991). See also Hall and Wehrly (1991). They believe that it is obvious that larger bandwidths should be used in the boundary area. The idea of the reflection method was first introduced by Schuster (1985) and Silverman (1986), and later on successfully extended by Cline and Hart (1991) by creating pseudo data. An alternative to Cline and Hart's extension are the more recent methods of Cowling and Hall (1996) as well as Zhang et al. (1999). I also mention the method for estimation performed on transformed variables cf. Wand et al. (1991), Ruppert and Cline (1994), Yang and Marron (1999), etc. Nevertheless, boundary correction methods are hardly used in density estimation or in regression, even though a considerable amount of theoretical studies and practical requirements exist. One important reason is that most procedures are only available in the literature, but not in any statistical or econometric software package. Another reason could be a disappointingly small performance improvement when using them. Finally, practitioners are often not willing to apply complex, and sometimes seemingly non-intuitive, methods. For this reasons we suggest a new boundary correction method that is simple and practical and can at least compete with Jones (1993) and local polynomials in both density and regression problems. As one will see, our method is much less complex and requires hardly more computational effort than the estimation without boundary correction does. A detailed methodological note with asymptotic insight, a comprehensive simulation study, and two applications are presented in Chapter 1.

- Data matching. In the second and third chapters I introduced an integration-based procedure to estimating and predicting population distributions. This is done by data matching with applications to the economics of wealth and health. From the methodological point of view the problem I dealt with is completely different from the problem dealt with in the first objective. However, we remain interested in welfare distribution estimation. Suppose we have a data set with which we want to conduct studies on welfare analysis. In the data set of our interest, however, the crucial information needed for household income and expenditure estimation is missing. In general, it is not particularly difficult, for the same country, region, similar year, etc., to find another data set, which not only preserve the household income variable but also has information about the other variables that are often used to construct the income prediction model. It is natural to estimate a regression model for household income and expenditure with the "auxiliary data set" and then use the estimated household income and expenditure model for the estimation of household income and expenditure that we are interested in. However, the use of this method only gives the mean income (or mean expenditure) conditioned on the available information and the specific model chosen. The resulting conditional distribution can by no means serve as an estimate of the distribution of the unconditional income or expenditure, and the subsequent poverty classification can only be "biased". In the present literature on poverty mapping and inequality studies, different approaches are applied to mitigate this problem like the adding of Gaussian errors to model based mean predictors. One could say, one does a kind of wild bootstrap under homoscedasticity to simulate the welfare distribution for the population of interest. This method, though quite popular, inherits several drawbacks, some of them are discussed in Chapter 2. We also mentioned two rather different approaches, which in some circumstances can provide more helpful solutions. The first approach is the quantile regression of conditional distributions and its marginals, see Koenker (2005), Firpo et al. (2009) and Rothe (2009). It sticks to the quantiles of a particular distribution instead of revealing the whole distribution of interest. The second approach is the imputation methods (see Dempster et al., 1977; Little and Rubin, 1987; Rubin, 1996; Schafer, 1997), where it is quite practical to impute some missing values in a survey or census. However, the 'imputation method' was not designed to provide an estimator or a predictor of the (marginal) distributions.

To prove that the proposed method also works well with a discrete data set, I estimate the (unconditional) discrete distribution of the number of doctor consultations for the population of interest. Further, I applied the proposed method to a moderate random sample from the population, then forecast that distribution for the population from which the sample was taken.

In conclusion, both Chapters 2 and 3, it is evident that the proposed method can be applied to estimate both (unconditional) discrete and continuous distributions. It is applicable irrespective of the mean regression or model, and can be easily extended to other contexts, such as small area statistics, nonparametric statistics, any latent variable model (e.g. Tobit regression), simultaneous equation systems, IV methods, etc..

- Income distribution estimation from a few quantiles. Again, the problem I faced is unlike those described in the previous sections. But I again solve problems in welfare distribution. A method for convex estimation of a regression function based on the spline smoothing technique is used to estimate the Lorenz curve from sparse data points. Compared to the currently available methods for Lorenz curve estimation, the new estimate does not require constrained optimization. The main contribution of this paper is to show how, based on a few quantiles, one can apply a functional form for the Lorenz curve to obtain a parametric density that is consistent with the given quantiles. Furthermore, one can easily derive inequality measures, such as Gini coefficient, based on the same information. In an application with quintile share data on the US income, it can be seen that the new estimate far outperforms the others. As an ongoing project the preliminary ideas and results are summarized in Chapter 4.

As I have mentioned previously, the three research objectives are totally different and are therefore independent of each other. Among the three objectives, objective 2 is however approached separately, using the discrete as well as the continuous data set. Although in most of my applications I have investigated the welfare distribution, my intention is to provide ideas that can be applied in a more general situation, cf. Chapter 3. Last, but not least, all these developments are only a source of help when they are provided in user-friendly software. The most popular software for statistical analysis at universities and research institutions currently is the freeware and open-source R-project. It is very similar to the commercial software S-Plus. I intend to provide an implementation of the proposed methods in R-modules to the general public, and thereby contribute to a rapid dissemination of my procedures.

1. Simple and Effective Boundary Correction for Kernel Densities and Regression with an Application to the World Income and Engel Curve Estimation

Abstract

In both nonparametric density estimation and regression, the so-called boundary effects, i.e. the bias and variance increase due to one-sided data information, can be quite serious. For estimation performed on transformed variables this problem can easily be elevated and may distort substantially the final estimates, and consequently the conclusions. After a brief review of some existing methods a new, straightforward and very simple boundary correction is proposed, applying local bandwidth variation at the boundaries. The statistical behavior is discussed and the performance for density and regression estimation is studied for small and moderate sample sizes. In a simulation study this method is shown to perform very well. Furthermore, it is an excellent method for estimating the world income distribution, and Engel curves in economics. This is joint work with Prof. Stefan Sperlich. My contributions in this paper are as follows: first I proposed a new method of boundary correction; and then I did complete implementation of this method in R.

1.1. Introduction

Boundary effects are a well-known problem in nonparametric estimation, no matter if we think of density estimation or regression. Moreover, if the estimation has been performed on transformed covariates, as recommended in the literature, see Wand et al. (1991), Ruppert and Cline (1994), Yang and Marron (1999), this problem may become elevated in two ways. Following these articles, a most appropriate transformation is the assignment $x_i \to \int_{-\infty}^{x_i} p(x)dx$ with p being a parametric prior (maybe with estimated parameters) of the density of X.

Firstly, after such a transformation we definitely face boundaries (here 0 and 1) with especially heavy tails. Secondly, what is just a boundary effect for the transformed data may then affect big and essential parts of the untransformed model. But also when we estimate an untransformed model directly, “boundaries” are not necessarily small nor are they mostly of minor interest. The larger the noise to sample size ratio or the smoother the function, the larger is the bandwidth and thus the affected boundary region. Furthermore, it is the boundaries that are of special interest; for example, in poverty analysis, it is necessary to have reliable estimates of the income distribution at the left side “close” to the natural boundary 0. Similarly, when using nonparametric regression in econometrics, spill-over effects, flexible returns to scale or multiple (dynamic) equilibria can typically, if at all, only be detected at, or close to, the boundaries. To conclude, if we are interested in risk, in poverty and inequality, the performance of especially young or old people, highly or poorly educated, compare large with small companies, etc., we always focus (also) on boundaries. In this article we will be confronted with boundary problems when studying the world income distribution, and when estimating the Engel curve for food expenditures in a poor country (Indonesia in our case).

As can be seen from these examples, we are concerned with boundary correction methods for both kernel density and kernel regression estimation. A quick internet search reveals that seemingly many boundary correction methods exist already, many are referred to the linear correction for density estimation, see Jones (1993), and can be considered as modifications of this method. A quite comprehensive discussion of boundary correction methods for density estimation is given in Cheng et al. (1997). In general, the existing methods can be divided in following groups:

The majority of researchers prefer the method of modifying the kernel, including Gasser et al. (1985), Jones (1993) and the local polynomial approaches (Cheng et al. 1997). Referring to the argument that local polynomial estimation would automatically correct for boundary effects in regression (see for example Fan and Gijbels, 1992) they apply this idea in density estimation. Effectively, however, a boundary correction takes place only if the polynomial is of the "correct" order; else it can even aggravate the boundary effect. In density estimation the use of local polynomial fitting has not prevailed, although Zhang and Karunamuni (1998, 2000) extended this method to the case of density estimation in combination with a bandwidth-variation function. Nevertheless, in many situations local polynomials are certainly an attractive remedy for boundary effects in regression, though the optimal weighting introduced by Cheng et al. (1997) has not been applied (much) until now.

The second set of boundary correction methods modifies the bandwidth near the boundaries. This group is much smaller and less known. Among them, Rice (1984), Gasser et al. (1985) and Müller (1991), see also Hall and Wehrly (1991), are maybe the most practical ones. They consider the regression context and suggest to fix the window size inside the support of the covariates. Somewhat similar to this idea, the loess and lowess smoother of Cleveland (1979, 1981) implemented in R and S, uses a fixed span thereby automatically addressing the boundary effects, see also Cleveland et al. (1992).

A quite old idea is the reflection method, introduced by Schuster (1985) and Silverman (1986), and later extended by Cline and Hart (1991). A further development of it is the more recent methods of creating pseudo data to correct for edges, see Cowling and Hall (1996). This method is more adaptive than the common data reflection approach in the sense that it corrects also for discontinuities in derivatives of the density. Zhang et al. (1999) suggested a method of generating pseudo data, combining the transformation and reflection methods. In some sense one could also add here the idea of Hall and Park (2002). They proposed an empirical translation of the argument of the kernels and a bootstrap method to translate the boundary estimate towards the body of the data set.

Finally we should mention again the transformation methods, see for example Wand et al. (1991), Ruppert and Marron (1994), and Yang (2000).

It is surprising that in spite of their importance in practice and the considerable (though not enormous) number of theoretical studies, boundary correction methods are hardly used either in density estimation or in regression. One obvious reason is the lack of implementation in statistical and econometric software; another could be a disappointingly small performance improvement when using them. Finally, practitioners are often not willing to apply complex, sometimes seemingly non-intuitive, methods.

For this reason we will concentrate mainly on comparing our method with that of Jones (1993) but also methods with fixed window size, the pseudo data approach (in particular Cowling and Hall, 1996) for densities, local linear for regression, and data transformation (in an application). However, to the best of our knowledge, even the quite well-known, and also reasonably successful method of Jones is neither much used nor implemented in standard software packages. Beside the lack of software, another reason for the scarce usage could be its complexity compared to the visible improvement in the final estimate. As will be shown, our method is much less complex and requires hardly more computational effort than does the estimation without boundary correction.

Summarizing, we are looking for a quick and easy boundary correction method that can at least compete with Jones (1993) and local polynomials in both, density and regression problems. Our method is driven by the idea of substantial bias reduction, c.f. Hall and Park (2002). Although the simplicity of our method allows for a (substantial) variance increase, in sum the boundary estimates improve in mean squared error. The method that handles the probability mass at or near the boundaries best is not at this point being looked into. We have introduced a new simple and practical method, given asymptotic insight, a comprehensive simulation study, a comparison with existing methods, and two applications.

1.2. Kernel estimators and boundary correction

Suppose we want to estimate a probability density f nonparametrically based on a random sample $\{X_1, X_2, \ldots, X_n\}$, $X_i \in R^d$. For the ease of presentation we restrict ourselves to univariate models ($d = 1$) in both density estimation and regression. The extensions to multivariate density and regression estimation are straight forward. The standard kernel density estimator of $f(x)$ is given by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x), \tag{1.1}$$

where $K_h(\bullet) = \frac{1}{h}K(\bullet/h)$ could be any common symmetric kernel with support $[-1, 1]$, satisfying $\mu_0(K) = 1$, $\mu_1(K) = 0$, $\mu_2(K) < \infty$, with $\mu_l(K) = \int_{-1}^{1} u^l K(u)du$ ($l = 0, 1, 2$; $u = \frac{X-x}{h}$) and h denoting the bandwidth. For such a kernel method to make sense, f is supposed to be smooth, typically expressed in the assumption of an existing second derivative f''.

However, if the support of f is bounded and has no exponentially falling tails, this estimator is well known to suffer from the so-called “boundary effects”. This means, for all points x being closer to the boundary than h, (1.1) underestimates (strongly) $f(x)$ since the kernel erroneously searches for information outside the support of f.

Now consider a random sample $\{(Y_i, X_i)\}_{i=1}^n$ for the regression model

$$Y_i = m(X_i) + \epsilon_i, \tag{1.2}$$

where ϵ_i are random errors with expectation zero and finite variance σ_i^2, and a smooth regression function $m(\bullet)$ that is assumed to have second derivatives. Then, the local polynomial estimator of degree α can be expressed as

$$\widehat{m}^{(v)}(x) = (v!)e_v^T(Z^T WZ)^{-1}Z^T WY\ , \tag{1.3}$$

where $m^{(v)}$ denotes the $v \leq \alpha$ derivative of m, Z is a $(n \times (\alpha + 1))$ matrix with elements $Z_{ik} = (X_i - x)^{k-1}$, $Y = (Y_1, \ldots, Y_n)$, $W = diag\{K_h(X_i - x)\}_{i=1}^n$, and e_v

is a vector of zeros with a 1 at position $(v+1)$. For $v=0$ and $\alpha=0$ we get the popular and simple Nadaraya-Watson estimator (Nadaraya, 1964). Also in this regression case, the problem of boundary effects is well-known and can become quite serious in practice.

To avoid confusion we shall assume (at least in the notation) that global bandwidths h_{global} were used unless otherwise stated, especially for the estimation at all interior points. Henceforth, the lower boundary - if it exists - is called a, and the upper boundary - if it exists - is denoted by c. In other words, the interior region is $[a+h_{global}, c-h_{global}]$ while $B_l = \{x : a \leq x < (a+h_{global})\}$ and $B_r = \{x : (c-h_{global}) < x \leq c\}$ are the left and right boundary regions.

Many methods have been proposed to correct for boundary effects, see Section 1.1. Probably the most popular one is the method of Gasser and Müller (1979), revitalized by and named after Jones (1993), namely the local linear estimation. Jones (1993) proposed to borrow more strength from inside of the support. More specifically, if f is supported on $[a,c]$, then the used kernel is given by

$$K^*(u) = \frac{w_3 - w_2 u}{w_1 w_3 - (w_2)^2} K(u) \mathbb{1}_{[c_2,c_1]}, \tag{1.4}$$

where the re-normalizing moments w_j are defined by

$$w_j = \int_{c_2}^{c_1} \left(\frac{t-x}{h_{global}}\right)^{j-1} K\left(\frac{t-x}{h_{global}}\right) dt,$$

with $c_1 = min(c, x+h_{global})$ and $c_2 = max(a, x-h_{global})$. Then the density estimate applying his linear boundary corrector is $\hat{f}$ in (1.1) but with the linearly corrected kernel $K^*(u)$. Similarly, for the regression estimator (1.3), we would use $K^*(u)$ in the definition of W.

An alternative is to choose local bandwidths in the boundary area. Typically, one would say it is obvious that larger bandwidths should be used there. Rice (1984) and Gasser et al. (1985) suggested choosing a bandwidth that keeps the window width fixed at the boundary; see also Hall and Wehrly (1991). To reach this we simply use, for all boundary points, a local bandwidth defined by

$$h_x = \begin{cases} 2h_{global} - (x-a) & \text{for } a < x < (a + h_{global}), \\ 2h_{global} - (c-x) & \text{for } (c - h_{global}) < x < c, \\ h_{global} & \text{otherwise.} \end{cases} \tag{1.5}$$

Hall and Wehrly (1991) extended this idea to first generate pseudo-data (with a kind of extrapolating bootstrap) and then estimate in the boundary region using the set of real and pseudo data. In the context of estimating a regression function $m(\bullet)$, Rice (1984) used a kind of Richardson extrapolation proposing a linear combination of uncorrected estimators $\hat{m}_{h_{global}}$ and corrected estimators $\hat{m}_{h_x}$. I.e. for all boundary points $x = a + h\rho$, $\rho < 1$ he set

$$\tilde{m}(x) = (1 + \beta_\rho)\hat{m}_{h_{global}}(x) - \beta_\rho \hat{m}_{h_x}(x), \tag{1.6}$$

with $\hat{m}$ as in (1.3) with $\alpha = 0$, h_x as in (1.5), and

$$\beta_\rho = \frac{w_1(\rho)w_0^{-1}(\rho)}{(2-\rho)w_1\left(\frac{\rho}{2-\rho}\right)w_0^{-1}\left(\frac{\rho}{2-\rho}\right) - w_1 w_0^{-1}} \qquad \text{for } w_k(v) = \int_{-1}^{v} u^k K(u)du.$$

In contrast to the idea of enlarging the bandwidth at the boundary, we suggest to reduce the bandwidth in the boundary regions. Our local bandwidth h_x for $a \le x \le c$ can be indicated by

$$h_x = \begin{cases} max(x-a, \varepsilon) & \text{if } a \le x < (h_{global} + a), \\ max(c-x, \varepsilon) & \text{if } (c - h_{global}) < x \le c, \\ h_{global} & \text{otherwise.} \end{cases} \tag{1.7}$$

where $\varepsilon > 0$ is just added for numerical reasons going to zero for $n \to \infty$. For theoretical discussion one could even skip ε and define h_x only for $a < x < c$ such that the density or regression estimator is not defined at the boundaries but arbitrarily close to them.

Inserting h_x, either (1.5) or (1.7) into (1.1), we have

$$\hat{f}_{h_x}(x) = \frac{1}{nh_x}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_x}\right) \tag{1.8}$$

for the kernel density estimator. As we can see, the local bandwidths h_x are adjusted within the boundary region while $\hat{f}_{h_x}(x)$ is identical to the usual kernel density estimator (1.1) if x is in the interior region. This also corresponds to Jones' method. It should be emphasized that the index x of h_x refers to a given point at which we wish to estimate the density or regression function. When we insert h_x into the regression estimator (1.3), we adjust only the weight W. In contrast, Jones' method is identical to (1.3) with global bandwidth inside the interior region, but using K^* in W.

We concentrate here on the situation where the boundary is (naturally) given; see also our applications in Section 1.4. For given boundaries and x the bandwidths h_x are neither in the interior nor at the boundary random. Therefore the statistical behavior of our resulting estimators is as simple as the method is. One might also imagine situations where the boundary is unknown and has to be estimated. Sometimes in the literature, the boundaries are set equal to the smallest and largest observation. Especially for density estimation, however, this is a quite questionable procedure to estimate the boundaries. In those cases the statistical behavior of our final estimate (density and regression) is very complex because it then has a random bandwidth. One would first have to establish assumptions and conditions on the boundary estimates etc. For simpler situations random bandwidths have been investigated e.g. in Abramson (1982), Hall (1983) or Hall and Marron (1988).

Recall that in our notation, a point x belongs to the boundary region when its distance to the boundary is smaller than h_{global}. In asymptotic theory a boundary point is a point x being closer to the boundary than the bandwidth used to estimate $f(x)$ or $m(x)$ respectively. In this sense, our method turns all support points into interior points and the asymptotics therefore remain unchanged. This was also the original idea of the reflection and of the pseudo data approach; they (artificially) changed the support, we change the bandwidth. Then, for the kernel density estimator (1.1) one obtains

$$Bias\{\hat{f}_{h_x}(x)\} = \frac{h_x^2}{2} f''(x)\mu_2(K) + o_p(h_x^2), \tag{1.9}$$

with $\mu_2(K) = \int_{-1}^{1} u^2 K(u)du$, and

$$Var\{\hat{f}_{h_x}(x)\} = \frac{1}{nh_x} f(x)||K||_2^2 + o_p(\frac{1}{nh_x}), \tag{1.10}$$

with $||K||_2^2 = \int K^2(u)du$. For the regression (1.3) one obtains

$$Bias\{\hat{m}_{h_x}(x)\} = \frac{h_x^2}{2}\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\}\mu_2(K) + o_p(h_x^2) \tag{1.11}$$

for the Nadaraya-Watson estimator with $\alpha = 0$, and

$$Bias\{\hat{m}_{h_x}(x)\} = \frac{h_x^2}{2} m''(x)\mu_2(K) + o_p(h_x^2) \tag{1.12}$$

for the local linear estimator with $\alpha = 1$, both with

$$Var\{\hat{m}_{h_x}(x)\} = \frac{1}{nh_x}\frac{\sigma^2(x)}{f(x)}||K||_2^2 + o_p(\frac{1}{nh_x}). \tag{1.13}$$

For consistency one needs $h_x \to 0$ and $nh_x \to \infty$ for $n \to \infty$. It is clear that our proposal of h_x, given in (1.7), gives full preference to bias reduction at the cost of increasing the variance. This becomes evident when we compare it with the methods of Jones (1993) and fixed window sizes. Nevertheless, in sum this can easily yield a reduction in mean squared error, as shown by our simulations in the next section. The pseudo data approach is constructed to control for both bias and variance at the edges.

Let us consider the asymptotics of a kernel density estimator when the method of Jones (1993) is applied. Without loss of generality we assume there is a lower bound a. Recall that we consider kernels bounded on $[-1,1]$. We skip the index *global* of bandwidth h and define implicitly a scalar p depending on

x and a via $x = p(a+h)$. Then, for $a_l(p) = \int_{-1}^{\min\{1,p\}} u^l K(u)du$ and $b(p) = \int_{-1}^{\min\{1,p\}} K^2(u)du$ the asymptotics can be approximated by

$$Bias\{\hat{f}_h(x)\} \simeq f(x)(a_0(p)-1) - ha_1(p)f'(x) + \frac{h^2}{2}f''(x)a_2(p), \qquad (1.14)$$

with

$$Var\{\hat{f}_h(x)\} \simeq \frac{1}{nh}f(x)b(p)\ . \qquad (1.15)$$

Note that for all interior points, the asymptotics coincide with the common expressions (1.9) and (1.10) respectively. In order to achieve a bias of order h^2 near the boundary, as well as in the interior, Jones (1993) defined a linear combination of K and a closely related function to obtain boundary kernel (1.4), such that $a_0(p) = \int_{-1}^{\min\{1,p\}} K^*(u)du = 1$ and $a_1(p) = \int_{-1}^{\min\{1,p\}} uK^*(u)du = 0$. Similar observations can be made for regression and the other boundary correcting methods.

The above, however, are asymptotic statements. In the next section we will study how these methods compare for finite samples of different sizes. We should emphasize once again that in the past it has been repeatedly stressed that local polynomial estimators do automatically correct for boundary effects. We mentioned already in Section 3.1 that this is only true if the order of polynomials is chosen accordingly. We should further remark that local polynomial estimators (in practice and theory) need larger bandwidths for increasing degrees. In boundary regions where data are sparse, it can even be recommendable to choose degree ≤ 1, i.e. to use the Nadaraya-Watson or local linear estimator. Applying Jones' or our method for local linear smoothers yielded poor numerical performance and is therefore skipped in the simulation section. The proposal of Cheng et al. (1997) to extend the local polynomial estimator by an additional weighting turns out to be rather complex in practice and still needs a reasonable amount of data.

We will also compare these simple methods with the reflection or pseudo data approach of Cowling and Hall (1996). Note, however, that this is by no means an easy-to-use or intuitive method. In fact the practitioners have to chose two further parameters which are essential for the success of the method. Cowling and Hall (1996) defined the density estimator at the boundaries as

$$\hat{f}(x) = \frac{1}{nh}\left\{ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) + \sum_{i=1}^{m} K\left(\frac{x - X_{-i}}{h}\right)\right\}, \tag{1.16}$$

where m is such that $O(nh) < O(m) < O(n)$, and X_{-i} are pseudo data. More specifically for positive constants $A_1, \ldots, A_s$, $s \geq r$, where r is related to the smoothness of the quantile function of X at the considered edge, and real numbers $a_1, \ldots, a_s$ they define

$$X_{-i} = \sum_{j=1}^{s} a_j X_{A_j i},\ 1 \leq i \leq \frac{n}{\max\{A_i\}}, \tag{1.17}$$

such that $\sum_{k=1}^{s} a_k A_k^j = (-1)^j$, $1 \leq j \leq r$. For example, in their article they recommend the so called best three-point rule $X_{-i} = -5X_{(i/3)} - 4X_{(2i/3)} + 10/3X_{(i)}$, $i = 1, 2, \ldots, n$ with $X_{(i)}$ indicating order statistics. Unfortunately, in Cowling and Hall (1996) say nothing about the choice of m, either in general or in their simulations. For more details and asymptotic behavior we refer to the paper of Cowling and Hall (1996).

Finally we would like to mention that there exist many other methods for nonparametric regression estimation, like different versions of splines, Fourier series, wavelets, etc. All these suffer a different kind of boundary effect. Fortunately, for our approach it is clear how it can be applied / extended to these other methods.

1.3. Finite sample comparison

We separate the simulation study into two parts: a more detailed one for density estimation, and a smaller study for regression. The reason is that in regression, the boundary performance depends on too many factors to provide a really comprehensive study; in fact, it depends on the distribution of the covariate(s), the functional form of the conditional mean of the response, on the degree of the (local) polynomial, and even on the heteroscedasticity. Therefore, the regression part of our simulation study has rather an illustrative character. In our simulations we set $\varepsilon = 0.001$ in (1.7).

1.3.1. Density estimation

To assess the effect of the correction methods near the boundaries, the following six models are investigated:

1. uniform distribution on $[0, 1]$;
2. gamma distribution $Gamma(2.25, 1.5)$ applied on $5x$;
3. log-normal distribution with $\mu = 0$ and $\sigma = 1$;
4. log-normal distribution with $\mu = 0$ and $\sigma = 1.5$;
5. log-normal distribution with $\mu = 0$ and $\sigma = 2$;
6. exponential distribution with $\lambda = \mu = 5$.

The density estimator was defined as in (1.1) with the Epanechnikov kernel $K(u) = 3/4(1 - u^2)\mathbb{1}\{|u| < 1\}$. For illustration issues we chose $h_{global} = 0.3$ thereby provoking substantial boundary effects. We estimated $f(\bullet)$ on a grid of 25 equidistant points $x_1 < x_2 < \ldots < x_{25}$, where $x_1 = 0$ and $x_{25} = 1$. Then the first 8 points lie in the left boundary region. The sample sizes were $n = 50$, $n = 100$ (not shown for brevity) and $n = 200$. All results were calculated from 1000 simulation runs.

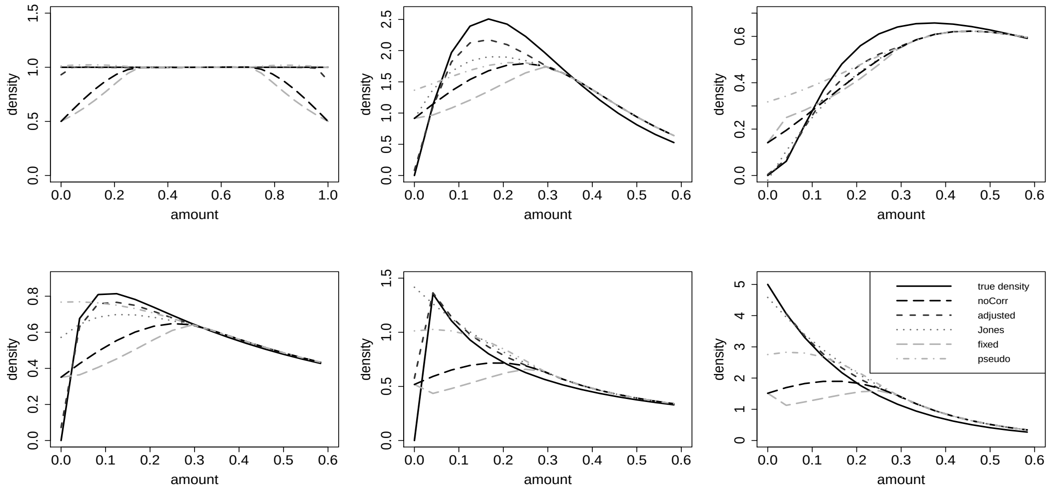


Figure 1.1.: The estimates for the six densities (upper left to the lower right) for n=50. Black line is the true density, black long dashes indicate the density estimate without boundary correction, grey long dashed is the method with fixed window size (1.5), black short dashed is our adjusted window method (1.7), grey dashed & dotted is the pseudo data method (1.16), and grey dotted line is Jones' estimate.

Figures 1.1 and 1.2 display the true density and the expectation of its kernel estimates, i.e. the averages over 1000 simulation runs. To highlight the behavior in the boundary region, we plotted the estimates in $[0, 0.6]$ for models 2 to 5, and in $[0, 1]$ for model 1. Maybe not surprisingly, see discussion in Section 1.2, our new method has the smallest bias and reflects best the true boundary behavior of the underlying densities. For both moderate sample size ($n = 50$) and relatively large samples ($n = 200$) our method outperforms the others, while Jones' method seems to be uniformly the second best. It should be remarked that Jones' estimator shows exactly the behavior indicated in (1.14); it strongly underestimates the curvature e.g. for model 2 and 4. The method with fixed window size is even worse than not correcting at all. As indicated, for the density estimation at the boundary we also tried the method of Cowling and Hall (1996) with the best three-point rule and the maximal possible resulting m. This maximal number seems to be $n - 1$, but it turned out that the performance improves (except for density 6) when we ignore all pseudo data X_{-i} lying in the support of X; cf.

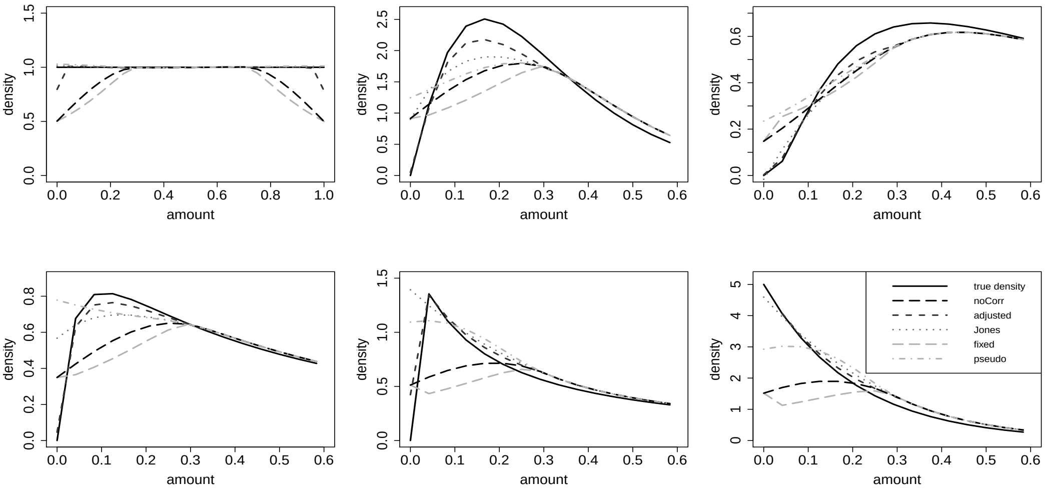


Figure 1.2.: The estimates for the six densities (upper left to the lower right) for n=200. Black line is the true density, black long dashes indicate the density estimate without boundary correction, grey long dashed is the method with fixed window size (1.5), black short dashed is our adjusted window method (1.7), grey dashed & dotted is the pseudo data method (1.16), and grey dotted line is Jones' estimate.

p. 555 of Cowling and Hall (1996). We also tried other choices, like $m = n^{9/10}$, but got worse results. Apart from the choice of pseudo generator and m the method of Cowling and Hall is computationally easy but its performance can only compete with Jones', or ours, when the original data are uniformly distributed.

Clearly, as stated in Section 1.2, our method is tailored to reduce bias but may have very large variance. If so, it can not really be considered as an improvement since the outcome would be rather random. To check this we constructed - again from our 1000 simulation runs - pointwise confidence bands with a coverage probability of 80%. These bands are given in Figures 1.3 and 1.4. First, we have to admit that at the boundaries our method has often the widest intervals. A closer look, however, reveals that they are not much wider and sometimes even tighter than the bands corresponding to Jones' method; and they are the only confidence bands that always include the true function, except for design 2. For $n = 200$ the widths of all the confidence bands are almost the same for our and

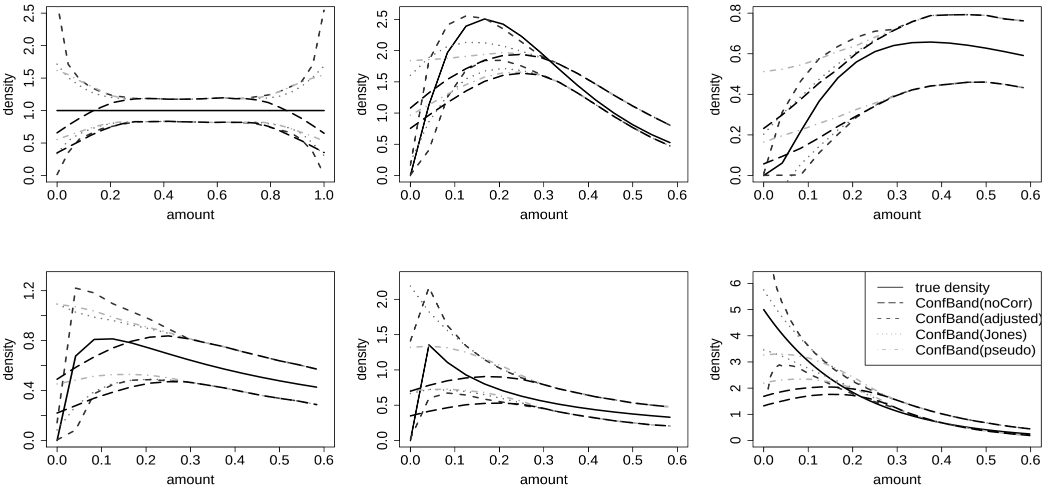


Figure 1.3.: The simulated confidence bands corresponding to Figure 1.1 with coverage probability of 80%.

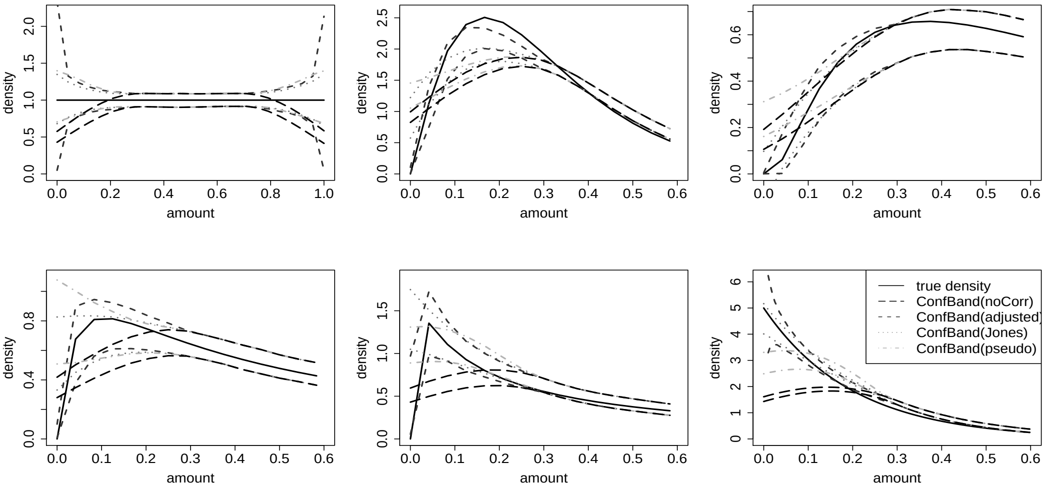


Figure 1.4.: The confidence bands corresponding to Figure 1.2 with coverage probability of 80%.

Jones' method.

To better quantify the gain in bias and mean squared error, we calculated the absolute bias and mean squared error averaged over the grid of 8 equidistant

n			$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
50	$\lvert Bias\rvert$	*no correction*	.2022	.5673	.0975	.1895	.2405	.7830
		Jones	.0013	.4696	.0678	.1280	.0765	.2370
		adjusted	.0105	.2093	.0474	.0365	.0518	.1436
		fixed	.2577	.7572	.1158	.2477	.3352	1.096
		pseudo	.0147	.6003	.1373	.1345	.1266	.4355
	MSE	*no correction*	.0596	.3384	.0208	.0527	.0776	.6261
		Jones	.0835	.3053	.0227	.0693	.0681	.1421
		adjusted	.6816	.1446	.0236	.0718	.1037	.2082
		fixed	.0776	.5811	.0215	.0723	.1239	1.206
		pseudo	.0708	.4053	.0354	.0522	.0516	.2512
200	$\lvert Bias\rvert$	*no correction*	.2014	.5665	.0951	.1888	.2428	.7811
		Jones	.0053	.4668	.0618	.1278	.0745	.2360
		adjusted	.0295	.2080	.0392	.0341	.0500	.1440
		fixed	.2575	.7563	.1142	.2460	.3365	1.096
		pseudo	.0142	.5962	.1109	.1528	.1168	.4498
	MSE	*no correction*	.0450	.3251	.0120	.0399	.0639	.6136
		Jones	.0195	.2383	.0085	.0292	.0203	.0780
		adjusted	.1484	.0670	.0071	.0158	.0259	.0678
		fixed	.0691	.5739	.0151	.0633	.1161	1.202
		pseudo	.0281	.3647	.0162	.0405	.0271	.2371

Table 1.1.: Absolute bias and MSE of density estimates in left boundary region for sample size $n = 50$ and $n = 200$, based on 1000 repetitions: *adjusted* refers to our method (1.7); *fixed* refers to a fixed window size (1.5); *pseudo* refers to (1.16).

points x_l over the left boundary region, i.e. we calculated

$$|Bias\{\hat{f}_h(x)\}| = \frac{1}{8}\sum_{l=1}^{8}\left|\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{f}_h^M(x_l) - f(x_l)\right)\right|, \tag{1.18}$$

$$\text{and MSE}\{\hat{f}_h(x)\} = \frac{1}{8}\sum_{l=1}^{8}\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{f}_h^M(x_l) - f(x_l)\right)^2 . \tag{1.19}$$

The results are displayed in Table 1.1. It can be seen from this table that, as expected, our method outperforms, by far, the competitors when looking at the bias. For the variance this is different, at least for small sample sizes (except for the $U[0,1]$ design). For $n = 100$ (not shown) the mean squared error is about the same for our method and Jones'; for $n = 200$ our new method outperforms all others considered, except for the $U[0,1]$ and $Log - N(0,2)$ design when comparing with Jones.

Before coming to the regression part we should briefly summarize. We have looked for a rather simple method, that is easy to implement and to interpret, for mitigating the boundary effects which in practice can cause rather serious problems and nuisance. As has been shown in Section 1.2, equations (1.7), our method complies with these requirements. Among all methods we have seen it is even the one with the simplest implementation. The ease of interpretation comes along with the insight that the statistical behavior is the same as for the interior points; it is a local bandwidth which - this we admit - can become rather small numerically although not in its rate. Fortunately, it has turned out in our simulation study that this method is not just the simplest one but also shows an excellent performance. In fact it outperforms even the popular method of Jones. The other alternatives considered seem not to work in our density examples.

1.3.2. Regression estimation

We recommend our new method not only for density estimation but also for kernel regression. As mentioned above, due to the fact that the boundary effects depend on too many factors, we have limited the following study to a brief illustrative simulation with only one design for the one dimensional covariate X, and a simple cubic polynomial for the regression function. That is, we consider random samples $\{(Y_i, X_i)\}_{i=1}^n$ from the nonlinear model

$$Y_i = m(X_i) + \epsilon_i \, , \text{ where } m(x) = -(10/3)x^3 + 5x^2 - 1.275x \tag{1.20}$$

is a smooth regression function, $X \sim U[0,1]$ i.i.d. and $\epsilon \sim N(0, 0.1)$ i.i.d. We estimated $m(\bullet)$ with the Nadaraya-Watson and the local linear estimator, i.e. (1.3) with $\alpha = 0$ or $\alpha = 1$ respectively. We used the Quartic kernel $K(u) =$

$15/16(1-u^2)^2\mathbb{1}\{|u|<1\}$ on a grid of 25 equidistant points $x_1 < x_2 < \ldots < x_{25}$, where $x_1 = 0$ and $x_{25} = 1$, as we did above. Then again, for a global bandwidth of $h_{global} = 0.3$ the first 8 points form an equidistant grid in the left boundary region. Note that the design choice favors now Jones' method; recall the results of Section 1.3.1. Like before, we did simulations for sample sizes $n = 50$ and $n = 200$.

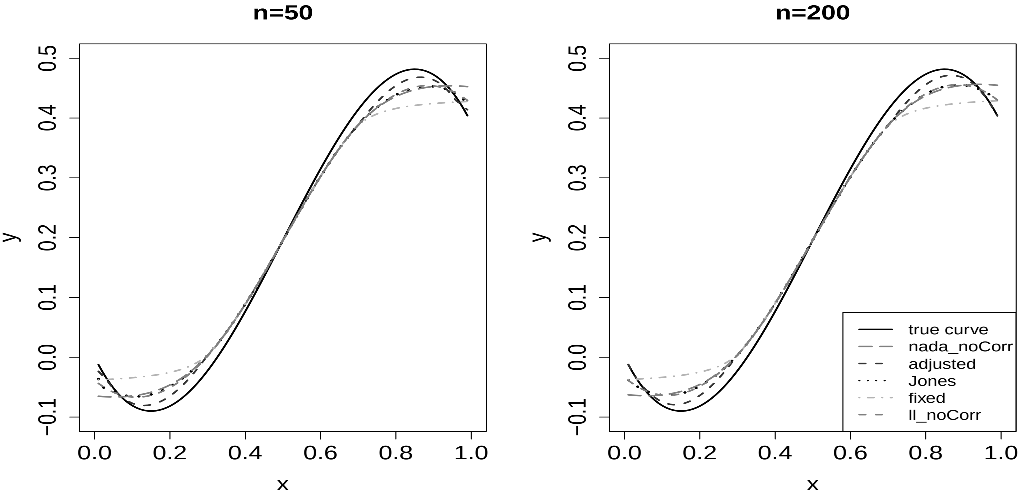


Figure 1.5.: Comparison of regression estimates: black line is true curve, grey long dashed is Nadaraya-Watson estimate without boundary correction, black short dashed is our (adjusted) method, black dotted line is Jones' estimate, grey dashed and dotted is the estimate with fixed window size, and grey short dashed is the local linear estimator.

As was the case for the density estimation context, a most serious problem is the bias at the boundary, and this is exactly what our method tries to mitigate. It can be seen from Figure 1.5, that the bias is corrected best by our method. Jones' method improves on the Nadaraya-Watson but not on the local linear estimator (not shown). It turned out that our method can also cause problems in combination with the local linear estimator (not shown), see our discussion about local polynomial estimation when data are sparse. Again, the method with fixed window size performs worst. We also tried Rice' (1984) more complex procedure, see (1.6), and found that it could not uniformly compete with simple

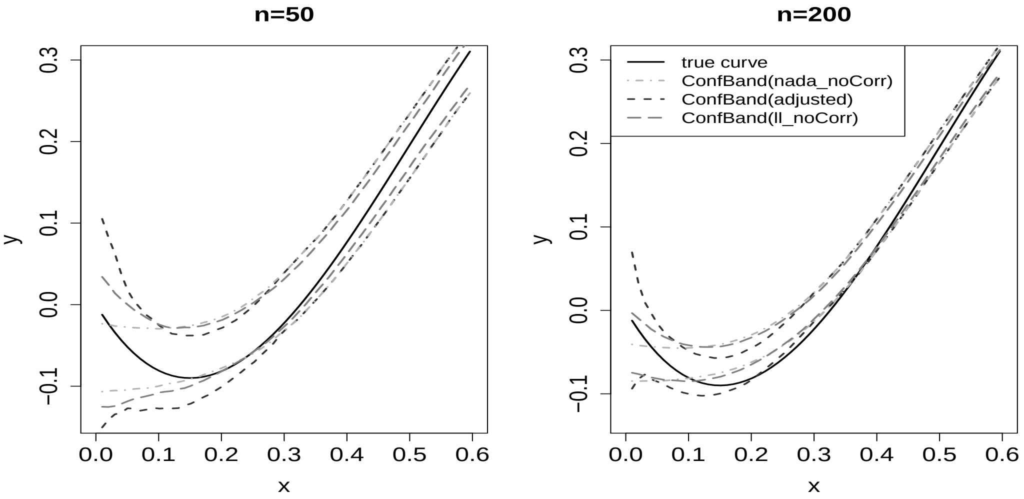


Figure 1.6.: The confidence bands for the left boundary, corresponding to Figure 1.5 for the non corrected Nadaraya-Watson, the non corrected local linear, and our method.

local linear nor with ours. Additionally, cannot be considered as a “simple and practical” methods. The local linear estimator turned out to be the strongest competitor compared to our method.

To have an idea about the variance of the estimators, we again constructed pointwise confidence bands with an 80% coverage probability, see Figure 1.6. As for the density estimation, the bands for our corrector are wider at the boundaries than for the other methods. Now the confidence bands are still much wider when increasing the sample size from $n = 50$ to $n = 200$. However, again it is only our method that really captures the curvature of the true data generating function such that the true function is almost always inside the 80% pointwise confidence bands, especially in the boundary region.

Our simulations conclude with Table 1.2 showing the average absolute biases and mean squared errors of the left boundary region. As we did for density estimation, we calculated

$$|Bias\{\hat{m}_h(x)\}| = \frac{1}{8}\sum_{l=1}^{8}\left|\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{m}_h^M(x_l) - m(x_l)\right)\right|, \tag{1.21}$$

$$\text{and } \mathrm{MSE}\{\hat{m}_h(x)\} = \frac{1}{8}\sum_{l=1}^{8}\frac{1}{1000}\sum_{M=1}^{1000}\left(\hat{m}_h^M(x_l) - m(x_l)\right)^2. \tag{1.22}$$

The results confirm what we have seen in Figures 1.5 and 1.6. Our method by far outperforms the others in terms of bias reduction at the boundary. Due to its large variance, however, its mean squared errors (on average) are clearly larger than for all in the small sample $n = 50$ and is still larger than others with sample size $n = 200$.

	$\|Bias\|$		MSE	
	$n = 50$	$n = 200$	$n = 50$	$n = 200$
adjusted	.0146	.0125	.0028	.0016
NW(no correction)	.0317	.0308	.0018	.0011
Jones	.0272	.0247	.0408	.0010
LL(no correction)	.0259	.0246	.0022	.0009
fixed	.0447	.0435	.0027	.0021

Table 1.2.: Absolute bias and MSE of regression estimates in left boundary region for sample size 50 and 200 based on 1000 repetitions.

1.4. World income distribution and Engel curve estimation

The potential of our method and the need of boundary correction is illustrated in the two following applications. First we estimate the world income distribution, and second we estimate the Engel regression curves for food expenditure in Indonesia.

The world income distribution is of ongoing concern for economists and scholars worldwide, see e.g. Acemoglu and Ventura (2002) and Sala-I-Martin (2006). The discussion of a two or even three mode shape (cf. Holzmann et al. 2007) of the world income distribution has been challenging the conventional findings of growth empirics. As a consequence, for example, the convergence literature established divergence among countries but found different convergence clubs. Further, from this world income distribution one can obtain measures for global inequality and poverty, as well as global growth incidence curves.

An often discussed question is how many convergence clubs do we find world wide, which should be certainly reflected in the shape of the income density function. The typical problem here is that of proper modeling, for example should one use a normal mixture or a log-normal mixture, and how should we bound the number of components (from above) or the variances (from below). This problem even appears in nonparametrics: when Holzmann et al. (2007) used the income, they encountered problems at the left boundary; when they considered log-income, the 'convergence club' of the rich countries (i.e. a bump on the right) was no longer visible. This can be seen quite well in our application in Figure 1.7. It shows kernel density estimates based on all available worlds real PPP GDP per capita for the year 2003 from the Penn World Table, Version 6.2. The available income data, and that used here, comprise 174 countries. In this analysis we estimate density $f(\bullet)$ with lower bound $a = 0$ on a grid of 200 equidistant points.

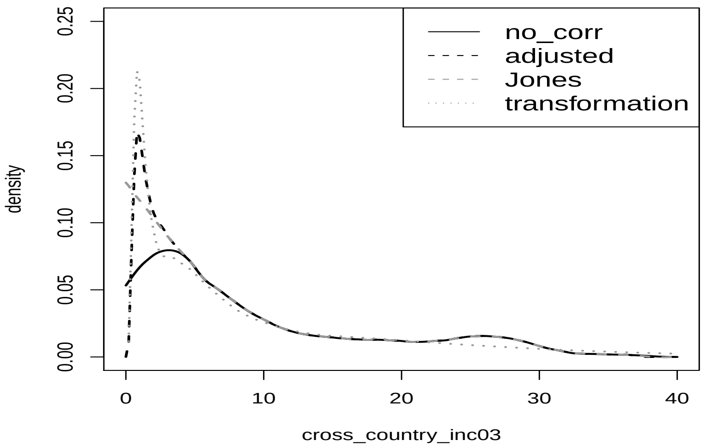


Figure 1.7.: Comparison of kernel density estimates for cross-country income distribution in 2003 with $h_{global} = 5.37$ (1.5 times Silverman's rule-of-thumb bandwidth): black solid line is kernel density estimate without boundary correction, black dashed is our method ('adjusted'), grey dashed is Jones' estimate, and grey dotted line is kernel estimate on log-transformed data with $h_{global} = 0.76$ (1.5 times Silverman's rule-of-thumb bandwidth). Scale: x-axis 10^3, y-axis: 10^{-3}.

The black line is the usual kernel estimate without boundary correction. The comparison with all other methods shows a serious boundary problem at 0. The global bandwidth has been chosen such that we could replicate the graphical results of Holzmann et al. (2007) where the bandwidth choice is not mentioned. The density estimation based on the log incomes and therefore facing no boundaries nicely resolves the very sharp peak at *income* $\approx$ 880 (very poor countries) and also makes visible a second convergence club of developing countries showing a plateau and a flat slope (to the left) at around *income* $\approx$ 3500. However, it does not exhibit the mode on the heavy right tail, i.e. the rich countries' mode. Jones method linearizes the slope until zero (from the right) which causes several problems in practice (density does not start from zero at zero nor does it exhibit the two first convergence clubs). Our simple boundary correction method is the only one that allows the estimator to reveal all interesting characteristics of this density. We refer to Vollmer (2009) for more discussion on the behavior of the cross-country income distribution.

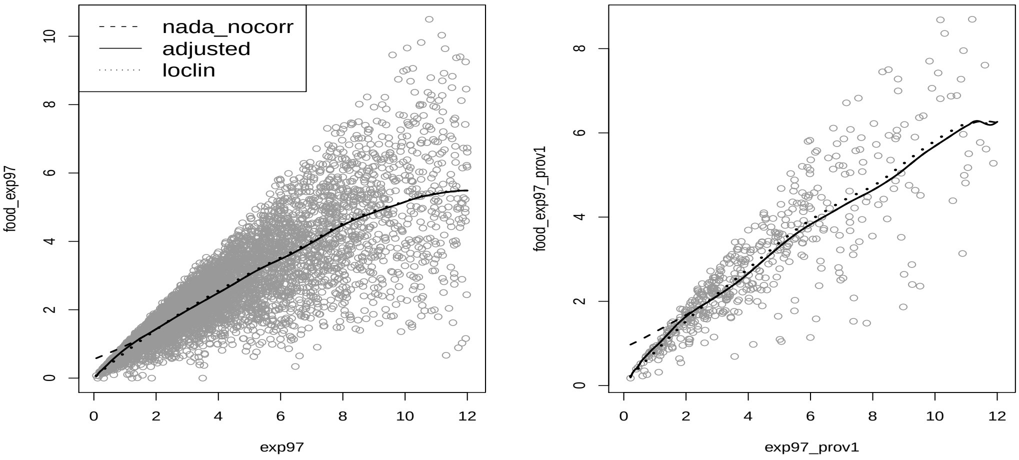


Figure 1.8.: Comparison of Engel curve estimates in 1997 with $h_{global} = 1.5$ (left, for Indonesia) and $h_{global} = 2.4$ (right, for North Sumatra) which is Silverman's rule-of-thumb times 3. Black dashed line is Nadaraya-Watson estimate without boundary correction, black solid is our method, and dotted line is local linear. Scale: x-axis 10^6, y-axis: 10^6.

The second application requires a nonparametric but boundary corrected regression. Since almost the beginning of econometrics, the specification and estimation of Engel curves has attracted the attention of many economists and applied econometricians. A detailed discussion and review of the parametric approaches to these problems are given in Deaton and Muellbauer (1980); an analysis of the cross-sectional consumer behavior in the context of fully nonparametric models can be found in Bierens and Pott-Buter (1990) or Engel and Kneip (1996). Still today, Engel curves are of special interest in welfare analysis. They are especially affected by a boundary problem at the left in poor countries like Indonesia. In Figure 1.8, we see $n = 6242$ observations of household annual food and total consumption expenditures per capita for the whole country (left), and among them $n = 502$ observations for the province North Sumatra (right). The source of these data is the second wave of Indonesia Family Life Survey (IFLS) in 1997. We fit an Engel curve to the left scatter plot of food versus total expenditure on a grid of 200 equidistant points with a natural left boundary at $a = 0$. Certainly, for poverty, welfare, and development analysis we pay special attention to the poorest, and these are exactly at the boundary. Again, the usefulness of our correction method is evident but one might argue that the local linear estimator does as well. There are several pros and cons and we do not want to enter the question which estimator has to be preferred. What we can say is that it seems that, with our boundary correction, the Nadaraya-Watson can compete with local linear estimation.

R. References

Abramson, I.S., 1982. On bandwidth variation in kernel estimates-a square root law. The Annals of Statistics 10(4), 1217-1223.

Acemoglu, D. and Ventura, J., 2002. The world income distribution. The Quarterly Journal of Economics 117(2), 659-694.

Bierens, H.J., and Pott-Buter, H.A., 1990. Specification of household engel curves by non-parametric regression. Econometric Reviews 9(2), 123-184.

Cheng, M.Y., Fan, J.Q and Marron, J.S., 1997. On automatic boundary corrections. The Annals of Statistics 25(4), 1691-1708.

Cline, D.B.H. and Hart, J.D., 1991. Kernel estimation of densities with discontinuities or discontinuous derivatives. Statistics 22, 69-84.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatter plots. J. Amer. Statist. Assoc. 74, 829-836.

Cleveland, W.S., 1981. LOWESS: A program for smoothing scatter plots by robust locally weighted regression. The American Statistician 35, 54.

Cleveland, W.S., Grosse, E. and Shyu, W.M., 1992. Local regression models, in: Chambers, J.M. and Hastie, T.J. (Eds.), Chapter 8 of Statistical Models in S. Wadsworth & Brooks/Cole.

Cowling, A., and Hall, P., 1996. On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society, Series B 58(3), 551-563.

Deaton, A., and Muellbauer, J., 1980. Economics and Consumer Behavior, Cambridge University Press, Cambridge.

Engel, J., and Kneip, A., 1996. Recent approaches to estimating Engel curves. Journal of Economics 63(2), 187-212.

Fan, J.Q. and Gijbels, I., 1992. Variable bandwidth and local linear regression smoothers. The Annals of Statistics 20(4), 2008-2036.

Gasser, T. and Müller, H.-G., 1979. Kernel estimation of regression functions, in: Gasser, T. and Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation (Lecture Notes in Mathematics 757). Springer Verlag, Berlin, pp. 23-68.

Gasser, T., Müller, H.-G. and Mammitzsch, V., 1985. Kernels for nonparametric curve estimation, Journal of the Royal Statistical Society, Series B 47(2), 238-252.

Hall, P., 1983. On near neighbour estimates of a multivariate density. Journal of Multivariate Analysis 13, 24-39.

Hall, P. and Marron, J.S., 1988. Variable window width kernel estimates of probability densities. Probability Theory and Related Fields 80, 37-49.

Hall, P. and Wehrly, T. E., 1991. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. Journal of the American Statistical Association 86(415), 665-672.

Hall, P. and Park, B.U., 2002. New methods for bias correction at endpoints and boundaries. The Annals of Statistics 30(5), 1460-1479.

Holzmann, H., Vollmer, S. and Weisbrod, J., 2007. Perspectives on the world income distribution - beyond twin peaks towards welfare conclusions, in: Proceedings of the German Development Economics Conference, Göttingen.

Jones, M.C., 1993. Simple boundary correction in kernel density estimation. Statistics and Computing 3, 135-146.

Müller, H.-G., 1991. Smooth optimum kernel estimators near endpoints. Biometrika 78(3), 521-530.

Nadaraya, E.A., 1964. On estimating regression. Theory of Probability and Its Applications 10, 186-190.

Rice, J., 1984. Boundary modification for kernel regression. Communications in Statistics - Theory and Methods 13(7), 893-900.

Ruppert, D. and Cline, D.B.H., 1994. Bias reduction in kernel density estimation by smoothed empirical transformations. The Annals of Statistics 22(1), 185-210.

Ruppert, D. and Marron, J.S., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society, Series B(Methodological) 56(4), 653-671.

Sala-I-Martin, X., 2006. The world distribution of income: falling poverty and ... convergence, period. The Quarterly Journal of Economics 121(2), 351-397.

Schuster, E.F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics - Theory and Methods 14(5), 1123-1136.

Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Vollmer, S., 2009. A Contribution to the Empirics of Economic and Human Development, Ph.D. Dissertation, Georg-August-University Göttingen. Peter Lang, Frankfurt am Main.

Wand, M.P., Marron, J. S. and Ruppert, D., 1991. Transformations in density estimation. Journal of the American Statistical Association 86(414), 343-353.

Yang, L. and Marron, S., 1999. Iterated transformation-kernel density estimation. Journal of the American Statistical Association 94(446), 580-589.

Yang, L., 2000. Root-n convergent transformation-kernel density estimation. Journal of Nonparametric Statistics 12(4), 447-474.

Zhang, S. and Karunamuni, R.J, 1998. On kernel density estimation near endpoints. Journal of Statistical Planning and Inference 70, 301-316.

Zhang, S., Karunamuni, R.J and Jones, M.C. 1999. An improved estimator of the density function at the boundary. Journal of the American Statistical Association 94(448), 1231-1241.

Zhang, S. and Karunamuni, R.J, 2000. On nonparametric density estimation at the boundary. Journal of Nonparametric Statistics 12(2), 197-221.

2. Predicting Household Expenditure and Income Distribution for Welfare Analysis

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 actually fails to predict the required information about tails and quantiles. In the present literature on poverty mapping and inequality studies, different devices are applied to mitigate this problem, like the adding of Gaussian errors to model-based mean predictors. Most methods yield simulated or numeric results, sometimes using strong non-testable assumptions. A new estimation and prediction method is introduced which can be combined with any reasonable mean prediction method. 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. Moreover, it allows us to impute welfare distributions for a census from survey data, as well as for synthetic populations under specific scenarios. Further inference is straight-forward, including the prediction of Lorenz curves, indexes like the Gini, or any distribution quantile, including confidence intervals. Chapter 2 and 3 are joint works with Prof. Stefan Sperlich and Prof. Walter Zucchini. My main contribution was a complete implementation of the proposed method in R.

2.1. 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 OPPG[1], 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 a statistical 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.

[1] Operationalising Pro-Poor Growth

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 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 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, 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 to introduce 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 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 statistics, 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 clearly theoretical rather than practical contributions. Computationally they can be quite cumbersome, especially the nonparametric approaches which are only recommendable for the one dimensional case; note that algorithms or computer codes are not available. 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 an 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 convemently 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 2.3 we consider two different types of problems of estimating the income distribution accounting for possible selectivity biases. In Section 2.4 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 2.5 concludes.

2.2. 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.

2.2.1. From marginalization to local n-fold mixtures

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

$$y_i = g(x_i) + \epsilon_i. \tag{2.1}$$

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 statistics, like in Elbers et al. (2003). 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

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

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.

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

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. In our opinion the latter is not recommended because, 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 Baysian 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 knn smoothing.

2.2.2. 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 very 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:

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

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 $T = (T_1,\ldots,T_q)$ 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.

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

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 $E[\epsilon^2|x] = c_0 + c_1E[Y|x] + c_2E^2[Y|x]$.

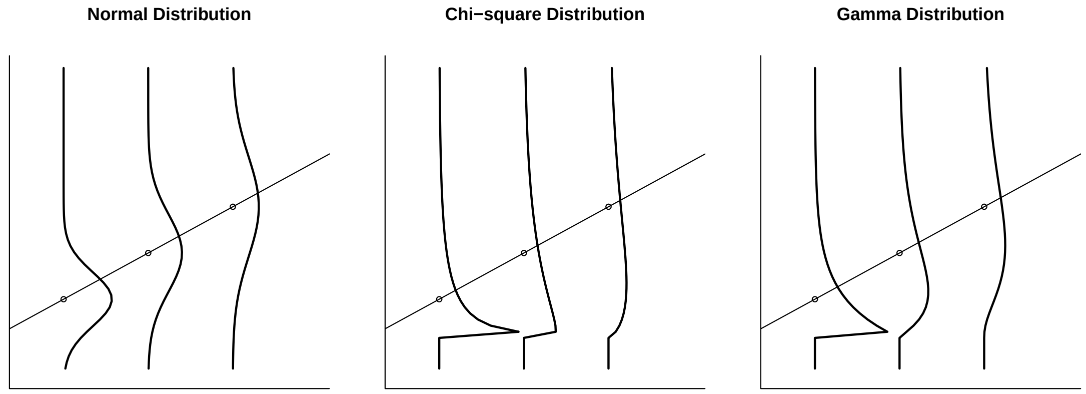


Figure 2.1.: Examples for typical prior conditional distributions with heteroscedasticity when the mean function is a straight line.

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 2.1 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

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

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 $k(x) = k$.

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.

2.3. 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.

2.3.1. An easy exercise

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; (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. Mincerian human capital theory suggests that education (here measured as average years of schooling) and experience of working household members (measured here as average

age) are chosen as explanatory variables. Other socioeconomic variables available are the share of working household member, 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 model 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 2.1. For the APLM we only give the coefficients for the parametric linear part without standard deviations.

As parametric prior distributions for the conditional density of $y|x$ in $\mathcal{S}_2$ we tried the 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 2.2.2. The resulting estimates for the income distribution in $\mathcal{S}_2$ are given in Figure 2.2 for the linear regression model, and in Figure 2.3 for the additive partial linear model. The density plots for the real income distribution and the distribution of conditional incomes, were created by kernel methods with Gaussian kernel and two times Silvermans rule-of-thumb bandwidths, as the default still gave wig-

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

gly 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

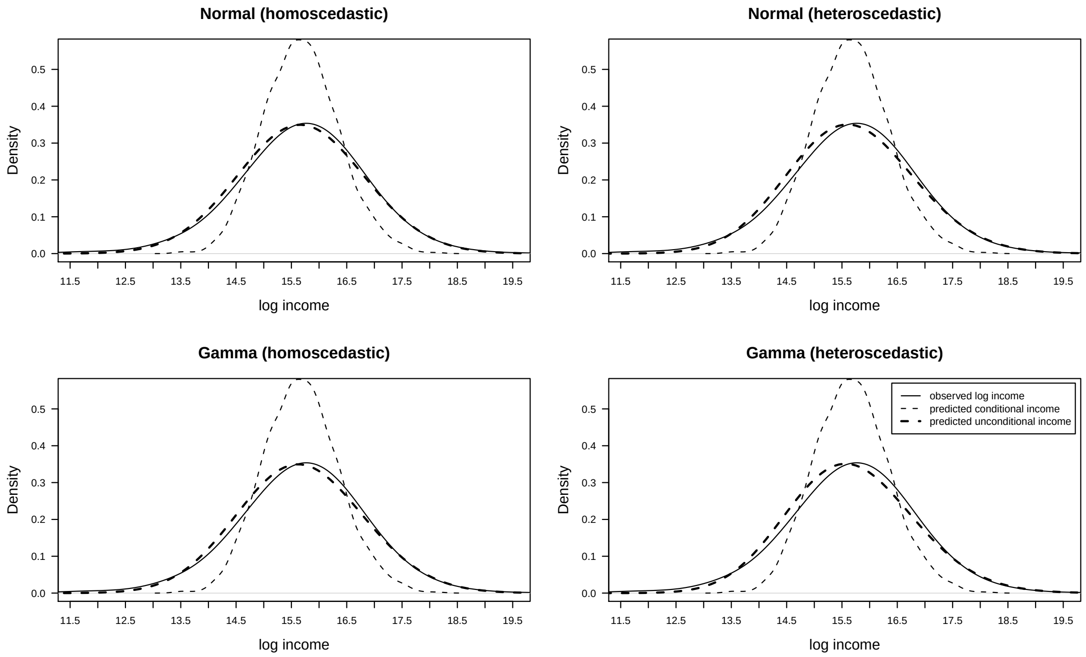


Figure 2.2.: Probability density curves based on a linear regression mean for the conditional (grey dashed) and the unconditional (dark dashed) predicted income, compared to a kernel density estimate based on the real incomes (black line). Different modeling approaches from the upper left to the lower right.

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 distribution) is

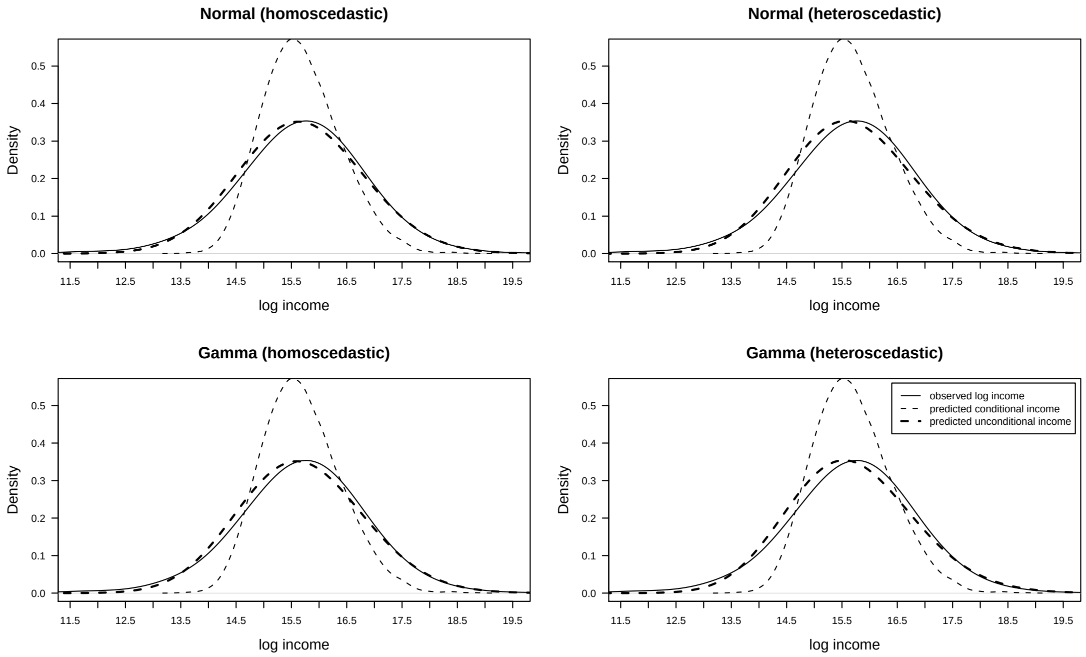


Figure 2.3.: Probability density curves based on an additive partial linear regression mean for the conditional (grey dashed) and the unconditional (dark dashed) predicted income, compared to a kernel density estimate based on the real incomes (black line). Different modeling approaches from the upper left to the lower right.

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

2.3.2. 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,

$$\begin{aligned} y^* &= x^T\beta + u, \text{ income} && (2.7)\\ s &= 1\!\!1\{z^T\theta + \epsilon\}, \text{ reports income or not} && (2.8)\end{aligned}$$

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 (2.8). Let y be the reported income (else $y = 0$), then we have

$$\begin{aligned} E(y|x, y > 0) &= x^T\beta + E(u|x, y > 0)\\ &= x^T\beta + \alpha \cdot \lambda(z^T\theta) && (2.9)\end{aligned}$$

where $\lambda(\cdot)$ is parametrically specified if the joint distribution of (u, ϵ) from equations (2.7) and (2.8) is. Therefore, the first step is the estimation of equation (2.8) to obtain θ, and the second step is the estimation of equation

$$y = x^T\beta + \alpha \cdot \lambda(z^T\hat{\theta}) + v \qquad (2.10)$$

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 (2.7).

We tried several parametric and semiparametric estimation methods; see references in Section 2.2.1. 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 2.4. Then we tried to use a semiparametric single index estimator for equation (2.8), 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 (2.10); see the next to last column of Table 2.1. Similar to what we found in the first exercise, Section 2.3.1, this semiparametric estimation procedure had hardly an impact on the final results for the unconditional income distribution of $\mathcal{S}_2$.

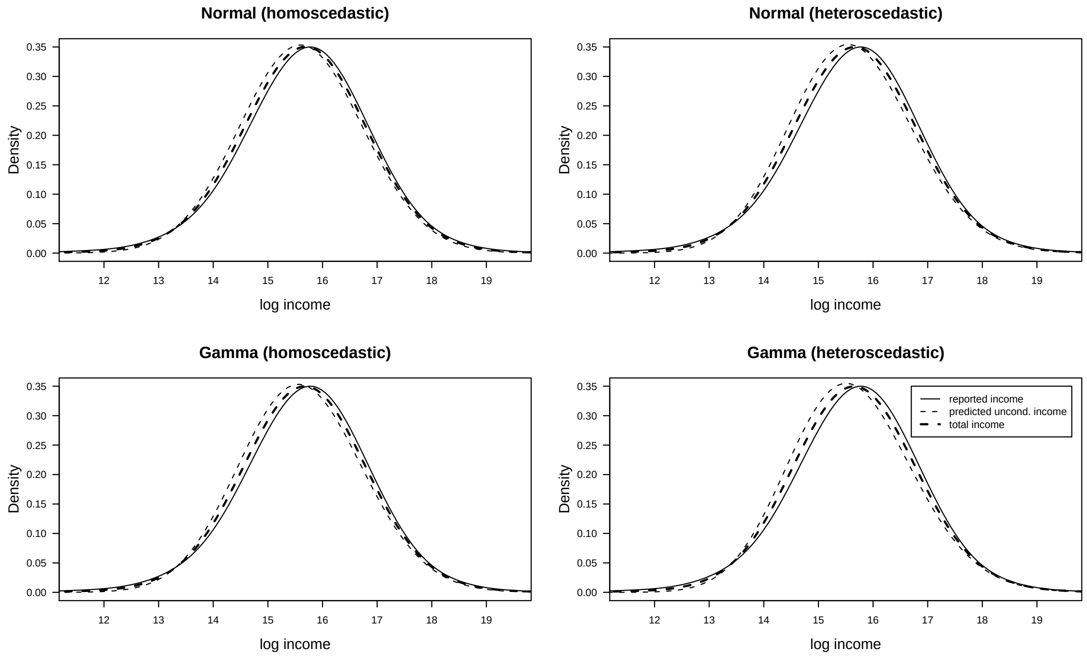


Figure 2.4.: Estimated and predicted density curves of unconditional income for households with not reported income (grey dashed), households with reported income (solid line), and for the whole sample (dark dashed) in 2008, based on different prior assumptions from the upper left to the lower right.

In Figure 2.4 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 income distribution which is also shown in Figure 2.4.

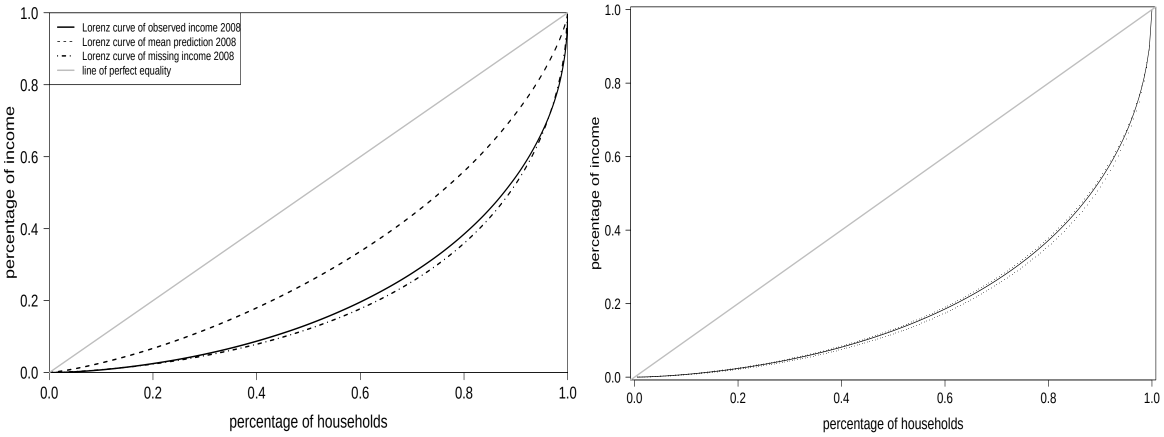


Figure 2.5.: Left figure: The Lorenz curves for the observed (solid) income, the conditional income (thick dashed) and the predicted income (dotted-dashed). Right figure: The Lorenz curve for the total survey, i.e. observed plus predicted with 99% point-wise confidence intervals.

In view of this potential source of bias, one should study the consequences e.g. for the Lorenz curve and Gini coefficient. In Figure 2.5, 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 2.5, 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.

2.4. 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 2.5 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 deflater 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 2.6 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 2.2 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 2.6.

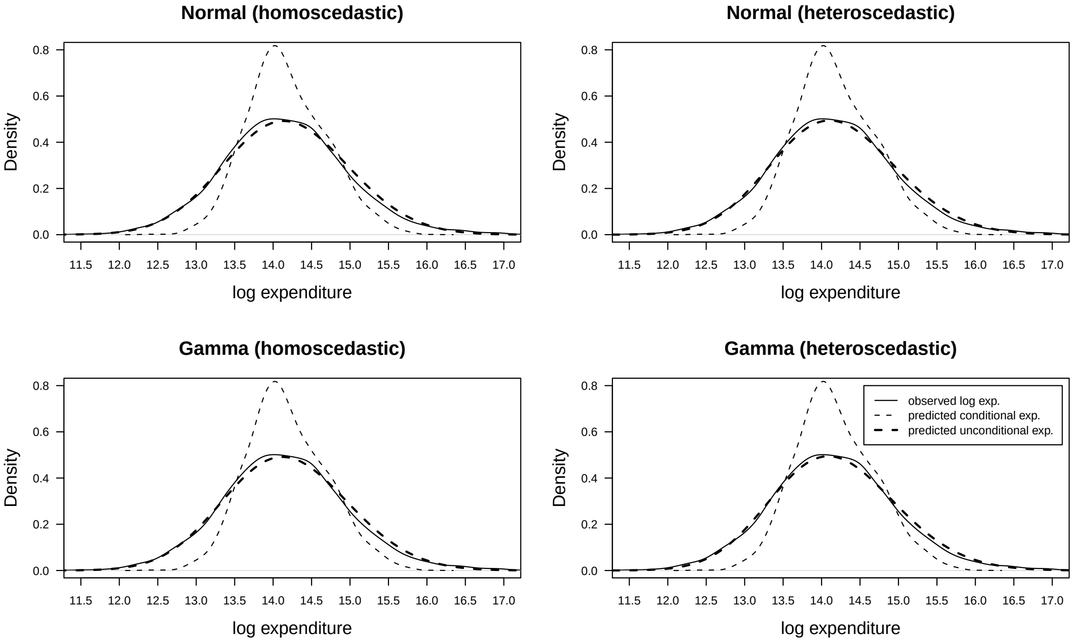


Figure 2.6.: Density curves for the conditional expenditures (grey dashed) the predicted unconditional expenditures (thick dashed) for 2000 based on a 1997 cohort, and a kernel density estimates of the observed expenditures (solid line) in 2000.

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

$$ISE = \int_{-\infty}^{\infty} [\hat{f}(y) - f(y)]^2 dy\,, \tag{2.11}$$

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

$$\hat{y}_{uncond} = F^{-1}(F_{\hat{y}}(\hat{y}|x)), \tag{2.12}$$

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 2.12 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 2.2 to 2.4. In Table 2.2 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 2.3 is shown what a purely mean prediction would tell us about poverty. Finally, in Table 2.4 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 2.3 and 2.4 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 2.2.: Number of Households below the relative poverty line according to the unconditional distribution prediction

	Observed	Predicted	90%Conf.Int.
NotPoor	4079	4495	[4476; 4509]
Poor	506	90	[76; 109]

Table 2.3.: Number of Households below the relative poverty line according to the mean prediction

Observed	Predicted	
	NotPoor	*Poor*
NotPoor	3711	368
Poor	352	154

Table 2.4.: Individual classification of households, predicted versus reported.

2.5. 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.

A. 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 2.5: Regression results of mean expenditures. Figures in parentheses give the standard deviations

	Inflation Rate		
Province	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 2.6: Regional inflation rates in 1997, 1999 and 2000 (Rp per capita/month), see also Pradhan et al. (2001)

R. 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, accessible on March 5, 2011.

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 microeconometric 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 semiparametric 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.

3. Estimating and Predicting the Distribution of the Number of Visits to the Medical Doctor

Abstract

This paper is concerned with the problem of estimating and predicting the distribution of the number of visits to the medical doctor. We are interested in predicting the distribution for a certain population, given a sample that is not necessarily taken from that population. The prediction is based on a pre-estimated conditional probability that is assumed to be the same for both the population of interest and that from which the sample was taken. We apply the method to data from Ryde (a suburb of Sidney) in 1994 and 1995. In a first step, we model the distribution of the number of visits to a general practitioner (GP) as a function of gender and age. The main contribution of this paper is to show how, based on a sample of observations, one can estimate the (unconditional) discrete distribution of the number of visits for the population of interest, and also forecast that distribution for the population from which the sample was taken. This is achieved using an integration-based procedure.

3.1. Introduction

The frequency distribution of doctor consultations is a primary indicator of health care utilization in a population and, as such, is of obvious importance for health care budgeting. Therefore, patterns in the frequency of consultations to the doctor, especially the dependency of the utilization of health resources on socioeconomic and geographic factors has been extensively documented, and its proper modeling is of central interest for empirical research in health economics and applied econometrics, respectively (Cameron et al., 1988; Pohlmeier and Ulrich, 1995; Windmeijer and Santos Silva, 1997; Deb and Trivedi, 1997; Jochmann and León-González, 2004; Winkelmann, 2004; to mention only few). Typically, the literature bemoans a lack of data on certain key variables on the one hand and the 'shrinkage effect' effect of conditional expectations on the other hand. More recently, Berzel et al. (2006) offered a plausible description of the number of doctor visits by modeling its dependence on a very limited number of demographic factors. In fact, it turned out that the mean number of doctor visits can already be estimated quite well when applying appropriate statistical modeling on simple available demographic factors such as age, gender and location.

Now, the target is to predict the numbers of visits for a population having at hand only these simple demographic factors for the population of interest, but full information - i.e. also the numbers of visits - observed for a particular sample. There exist several well-studied methods for estimating missing values (see Dempster et al., 1977; Little and Rubin, 1987; Rubin, 1996; Schafer, 1997), some of which could be used in our context. Then, instead of estimating the distribution of interest directly one could consider the numbers of visits as missing values and complete the data by simulated (imputed) values. Most of these imputation processes are iterative. A key feature of such approaches is to regard missing data as random variables, and then to replace them with multiple draws from the assumed underlying distribution. Therefore, these methods are often known as ‘multiple imputation’.

However, to our knowledge, no direct estimator of the population distribution of the doctor consultations has been reported yet. Note that the above-mentioned methods perform well for imputing missing values but have not been studied for imputing values for a whole population. Just thinking of the computational burden, if - as typically the case - the underlying sample is small but the population is large, these methods are not really attractive for our problem. Although the method we introduce here is straight forwardly applicable on many similar estimation or prediction problems, we concentrate on estimating the population distribution of doctor consultation frequencies, based on a moderate sample.

In a first step, Section 3.2, we search for a reasonable conditional distribution model based on only those covariates that are available in both the sample and the population of interest. As commonly acknowledged, the Poisson or negative binomial generalized linear model is the simplest way to model count data. The chosen link is typically the logarithm, i.e. the canonical link. According to the exploratory analysis, however, it may not be appropriate to use Poisson or negative binomial generalized linear models for our data problem since the generalized linear model does not allow for the overdispersion parameter to depend on covariates. As a way of overcoming these limitations associated with Generalized Linear Models (GLM, see Nelder and Wedderburn, 1972) we tried also the generalized additive model for location, scale and shape (GAMLSS), introduced by Rigby and Stasinopoulos (2005). Nevertheless, all distribution models in question should be adapted to the sample, as the final objective is the optimal prediction or estimation of the unconditional distribution(s) of the population(s) of interest. In a second step, Section 3.3, one can now derive these distributions of interest as being a mixture of $N(=population\ size)$ of the above calculated conditional distributions. All we need is a clear idea of the distribution of the covariates in the populations of interest and the assumption that the a priori fitted conditional models hold throughout. In Section 3.3 we present the numerical results. Section 3.4 concludes.

Table 3.1.: 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)	–	–

3.2. Modeling of the conditional distributions

The data set considered in this paper records 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. A more detailed description of these data can be found in Heller (1997). In the original data set there are 11 individuals (of the $23,618$) reporting more than 100 visits. As it turned out that this was due to an excessive misuse of the health insurance card by illegal immigrants for which is was impossible to obtain reliable corrections, we decided to truncate the data at a maximum of 100 visits. Note that then we have just 41 individuals in 1994, and 40 in 1995, with more than 52 visits, i.e. more than one each week. As no information is available that would allow for a sound detection of missmeasurement, we have not truncated these counts. The summary statistics for the remaining set of $N = 23,607$ inhabitants are given in Table 3.1.

There are mainly two prediction problems of interest. First, practitioners usually only have access to surveys, which for local areas can be of moderate size. From these they have to estimate the number of visits for a certain population, or to predict them for an artificial population to calculate scenarios. For example, in most industrialized countries a serious demographic change is expected in the next two decades which will effect the health systems and pension funds. In order to simulate these two situations we first draw a random sample of only

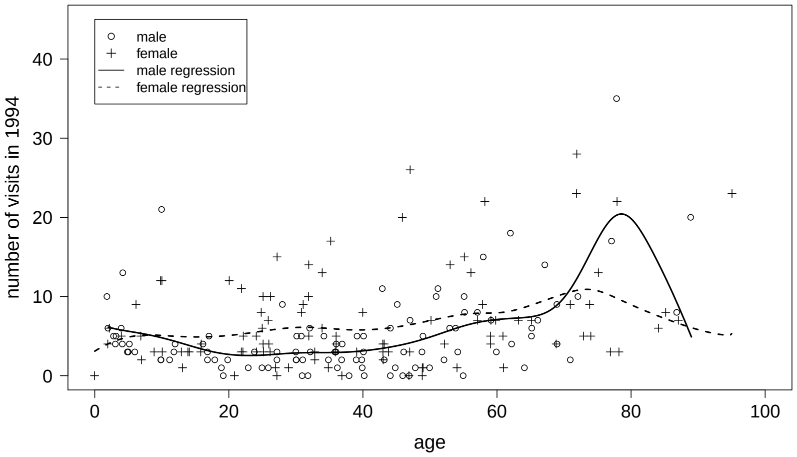


Figure 3.1.: The number of visits to a GP (left, in 1994; right, in 1995) plotted against age for a simple random sample of 200 residents in Ryde. Local linear regression estimate with cross-validation bandwidth $\hat{h}_{CV} = 2.78$ (black line, male) and $\hat{h}_{CV} = 2.78$ (grey line, female)

200 observations from 1994 with the summary statistics given in Table 3.1. The extension to stratified sampling or other sampling schemes is obvious. The aim is to estimate the distribution of the number of visits to the medical doctor in 1994, and afterwards to predict it for 1995.

On the one hand it is well known that *gender* strongly interacts with age when looking at visits to the doctor; on the other hand, *age* is the only additional variable. Therefore, we first have to decide whether for a reasonable model fit the sample should be split by gender. In order to study this, we plot the number of doctor visits against age in Figure 3.1, separately for male and female. The solid and dotted lines are simple local linear regression estimates. They indicate a non-linear relationship between the mean of the number of doctor visits with age and gender. Furthermore, the differences between males and females seem to be quite complex and hard to capture in one common model.

Secondly, we analyze the variance-mean ratio to check for under or over dispersion. Figure 3.2 shows the variance-mean ratio by age and gender for the random sample of 200 inhabitants in 1994. The ratio is clearly greater than one

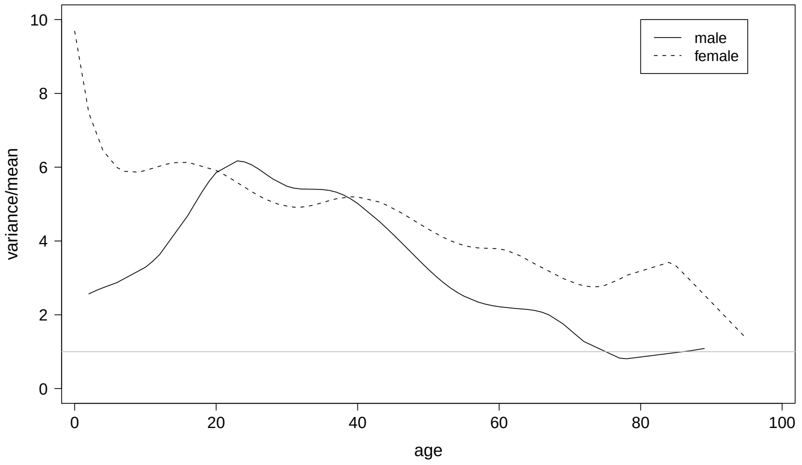


Figure 3.2.: Variance by mean, separate for males and females, based on 200 random samples in 1994.

for all levels of *age*, indicating inappropriateness of the Poisson model because of over dispersion. This exploratory analysis also reveals that *age* and *gender* have a strong influence on both the mean as well as the variance of visits; compare Figure 3.1 and 3.2.

Recall that the negative binomial regression model allows for overdispersion by introducing an unobserved heterogeneity term for each observation i. Observations are assumed to differ randomly in a manner that is not fully accounted for by the observed covariates. It assumes a negative binomial distribution for the response variable y in which its mean μ is modeled as a function of explanatory variables and a variance of the form $\mu + \mu^2\sigma$, where σ is an unknown overdispersion parameter which in turn shows no extra dependency on the covariate values. However, from Figure 3.2 we notice that the variance-mean ratio varies substantially over the covariate values. Consequently neither the standard Poisson nor negative binomial generalized linear models seem to be appropriate in this case.

As indicated, we will need to fit appropriate models of conditional distributions to our data. Given our count data and the above findings we start with the negative binomial model (see for example, Cameron and Trivedi, 1998, Section 4.2.2), defined by

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

with mean μ and variance $\mu + \mu^2\sigma$, see above. If the overdispersion is mainly due to zero inflations, an alternative extension of the simple Poisson is the zero inflated Poisson, i.e.

$$f(y|\mu,\alpha) = (1-\alpha)\cdot Po(y,0) + \alpha \cdot Po(y,\mu), \quad Po(y,\mu) = e^{-\mu}\mu^y/y!, (3.1)$$

where again μ is modeled as a function of the covariates whereas α is an unknown scalar. An alternative to this extension of the Poisson we can also consider a zero inflated negative binomial having μ as a function of covariates and two unknown parameters σ and α. Different approaches to tackle the zero-inflation or other finite mixtures are proposed e.g. by Gurmu (1997), Deb and Trivedi (1997). See that issue also for further suggestions though in different contexts. As we mentioned before, for modeling linear functions, the linear models, lm(), and generalized lineal models, glm() of Hastie and Pregibon (1992) in the R language can be used. However we are restricted to model only the mean using lm() and glm().

In order 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 as quality of fit statistics. The results are listed in Tables 3.2 and 3.3 respectively, separated by gender. Note first that the different criteria do not contradict each other. The zero-inflated Poisson model provides a slightly better fit than the Poisson model (not shown). However, the model which is superior (according to the AIC) is the negative binomial. The zero-inflated negative binomial shows no improvement compared to the negative binomial because the zero inflation is unnecessary after the inclusion of σ. Consequently, the observed deviance difference is zero relative to the negative binomial. The AIC even

Table 3.2.: Quality of fit statistics using GLM (for males)

Model	*Link*	*Terms*	*llh*	ΔD	*AIC*
zero-inflated Poisson	$log(\mu)$	$age + age^2$	−294	–	596
negative binomial	$log(\mu)$	$age + age^2$	−253	82	515
zero-inflated negative binomial	$log(\mu)$	$age + age^2$	−253	82	517

Table 3.3.: Quality of fit statistics using GLM (for females)

Model	*Link*	*Terms*	*llh*	ΔD	*AIC*
zero-inflated Poisson	$log(\mu)$	$age + age^2$	−360	–	728
negative binomial	$log(\mu)$	$age + age^2$	−286	148	579
zero-inflated negative binomial	$log(\mu)$	$age + age^2$	−286	148	581

indicates that the improvement in fit is insufficient to justify the use of the more flexible but also more complex model. Recall that our main objective is not the optimal fitting but prediction, which is much more sensitive to overfitting due to complexity. Indeed, complexity is often one of the worst enemies of good prediction.

However, the generalized linear considered so far is restricted to allow only the location parameter to depend on covariates, and this only in a known parametric way. Rigby and Stasinopoulos (1996, 2005) developed a general class of univariate regression models, called the Generalized Additive Model for Location, Scale and Shape (GAMLSS) with two important extensions. First, they allow all distribution parameters to depend on a predetermined set of covariates. Second, the modeling of these parameter functions may include random effects or even be nonparametric, but being always of an additive structure. The model assumes independent observations of the response variable given the parameters, the covariates and the values of the random effects. It provides a very general distribution family for univariate continuous or discrete response variables. In our case, under the negative binomial distributional assumption, both the mean and the dispersion parameter can be modeled as a function of *age*. To summarize, we consider the negative binominal density $f(y|\mu,\sigma)$ and will estimate

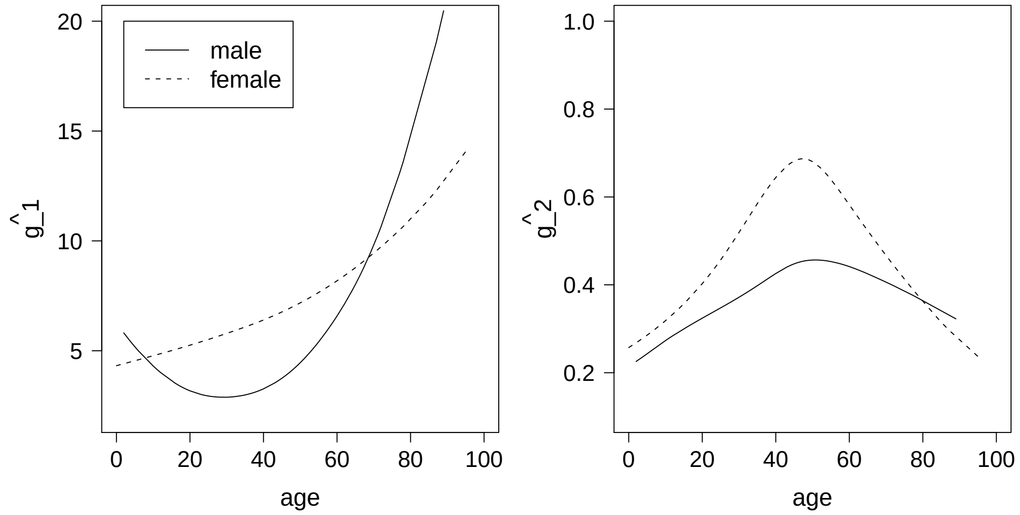


Figure 3.3.: Impact of age and gender on the GAMLSS nonparametric regression estimates for mean g_1 (left) and dispersion g_2 (right), based on a random sample of 200 residents in Ryde in 1994.

$$\log(\mu) = g_1(age), \qquad \log(\sigma) = g_2(age), \tag{3.2}$$

where we first will set g_1, g_2 to be parametric quadratic function, and afterwards nonparametric cubic splines (cs). For the latter we have plotted the functions g_1, g_2 in Figure 3.3.

For comparing these two GAMLSS models, we use the well known fitted global deviances $GD = -2l(\hat{\theta}) = -2\sum_{i=1}^{n} l(\hat{\theta}^*)$, the Akaike information criterion AIC of Akaike (1974) and the Schwarz Bayesian criterion SBC of Schwarz (1978). AIC and SBC are asymptotically justified as predicting the degree of fit in a new data set, i.e. approximations to the average predictive error. The global deviance, SBC and AIC are summarized as statistics relating to the fit of the parametric and nonparametric GAMLSS models in Table 3.4 and 3.5, again separately for males and females. Fortunately, the different criteria do give the same selections so that it is enough to look at the AIC here.

A further possibility to model dispersion in parametric or nonparametric nega-

Table 3.4.: Quality of fit statistics using GAMLSS (for males)

Model	*Link*	*terms*	*GD*	*AIC*	*SBC*
negative binomial	$log(\mu)$	$age + age^2$	506	518	533
(parametric model)	$log(\sigma)$	$age + age^2$			
negative binomial	$log(\mu)$	$cs(age)$	502	515	534
(nonparametric model)	$log(\sigma)$	$cs(age)$			

Table 3.5.: Quality of fit statistics using GAMLSS (for females)

Model	*Link*	*terms*	*GD*	*AIC*	*SBC*
negative binomial	$log(\mu)$	$age + age^2$	568	580	596
(parametric model)	$log(\sigma)$	$age + age^2$			
negative binomial	$log(\mu)$	$cs(age)$	568	580	595
(nonparametric model)	$log(\sigma)$	$cs(age)$			

tive binomial regression is the Vector Generalized Additive Model introduced by Yee and Wild (1996). One can also find some discussions about applying the provided R-package VGAM for count data in Berzel et al.(2006). However, already now we can see, compare Tables 3.2 to 3.5 that the AIC always selects the negative binomial generalized linear model throughout. This confirms our statement that, depending on the amount of information (data and signal-noise ratio), complexity is one of the worst enemies of prediction. Consequently, it is questionable to what extend other flexible, semi- or non-parametric model approaches can improve in our prediction problem. Nevertheless, in the final step we will also consider the GAMLSS results for the following reason. Our objective is not the conditional but the unconditional density of visits, and we do not know which model yields the best results there. Figure 3.3 shows that the data fit indicates a nonlinear, nonconstant dispersion parameter. While limiting to a quadratic modeling seems adequate, ignoring this finding might cause prediction loss in the final step.

3.3. Predicting population distributions

In this section, we come to this final step when applying our new method to the two described problems. Both are for prediction: ones is using a random sample of 101 males, and 99 females respectively, in 1994 to estimate the distribution of number of visits for the male (and female) population in Ryde in the same year (case study 1); another one is to predict the number of visits prediction for male (or/and female) population in Ryde in 1995 using the same random sample of males (or/and females) in 1994 (case study 2).

3.3.1. Case study 1

We start with estimating the distribution of number of doctor visits for the population, given a sample in the same year. The method we suggest is not simulation based; it provides transparent and reproducible estimators. We have a sample $\{(x_i^s, y_i^s)\}_{i=1}^n$, and the covariates $\{x_j\}_{j=1}^N$ of the population of interest. Recall that the required unconditional distribution of the population of interest, say $f_N(y)$, is simply the marginal distribution of the joint density $f_N(y,x)$ such that

$$f_N(y) = \int f_N(y,x)dx = \int f_N(y|x)f_N(x)dx\,. \tag{3.3}$$

For finite populations we can simplify to

$$f_N(x) = \begin{cases} \frac{1}{N} & \text{if } x = x_j \\ 0 & \text{if } x \neq x_j \end{cases} \tag{3.4}$$

and then obtain

$$f_N(y) = \frac{1}{N}\sum_{j=1}^{N} f_N(y|x_j). \tag{3.5}$$

Thus, what we need is a reasonable substitute in equation (3.5) for the conditional densities $f_N(y|x)$. An obvious choice here is one of the conditional densities fitted to the sample data, say $f_n(y|x)$. If $f_n(y|x)$ is a consistent estimate of $f_N(y|x)$, the consistency for $f_N(y)$ follows immediately. Also the asymptotic properties can be derived directly for most cases via Taylor expansion. In the nonparametric case this can be quite tedious, compare e.g. Van Keilegom and Veraverbeke (2002) or Sperlich (2009). In both the nonparametric and the parametric world, the estimator of the unconditional density will inherit consistency and convergence rate from the conditional density estimate.

What happens if $f_n(y|x)$ is not a consistent estimate of $f_N(y|x)$? In that case our procedure will still give a good approximate for $f_N(y)$ as long as the relation between y and covariates x specified and estimated from the sample can be carried over to the population reasonably well. In that case one could think of

$$\hat{f}_N = \frac{1}{N}\sum_{j=1}^{N} f_n(y|x_j)$$

as an N-fold mixture of pre-determined densities relating y to some covariates x.

In our case we considered the negative binomial (NB) as a reasonable description of the relation between y and x. With the estimates $\hat{g}_1, \hat{g}_2$ obtained from our sample $\{(x_i^s, y_i^s)\}_{i=1}^n$ in Section 3.2 we estimate then the unconditional probability function of y by

$$\hat{f}_N(y) = \frac{1}{N}\sum_{j=1}^{N} NB[y|\hat{g}_1(x_j), \hat{g}_2(x_j)]. \tag{3.6}$$

For the different specifications, NB with quadratic g_1 and constant g_2 (the GLM), g_1 and g_2 quadratic functions of *age*, and finally cubic splines for g_1 and g_2 (GAMLSS), the results are given in Figure 3.4. As in our example we have records of the real number of visits, we can evaluate our predictions exactly. It can be seen that the distribution of the estimated conditional mean (circles) is much too narrow to be of use when we are interested in the distribution of real visits (bars). In contrast, the predicted unconditional distribution (solid circles)

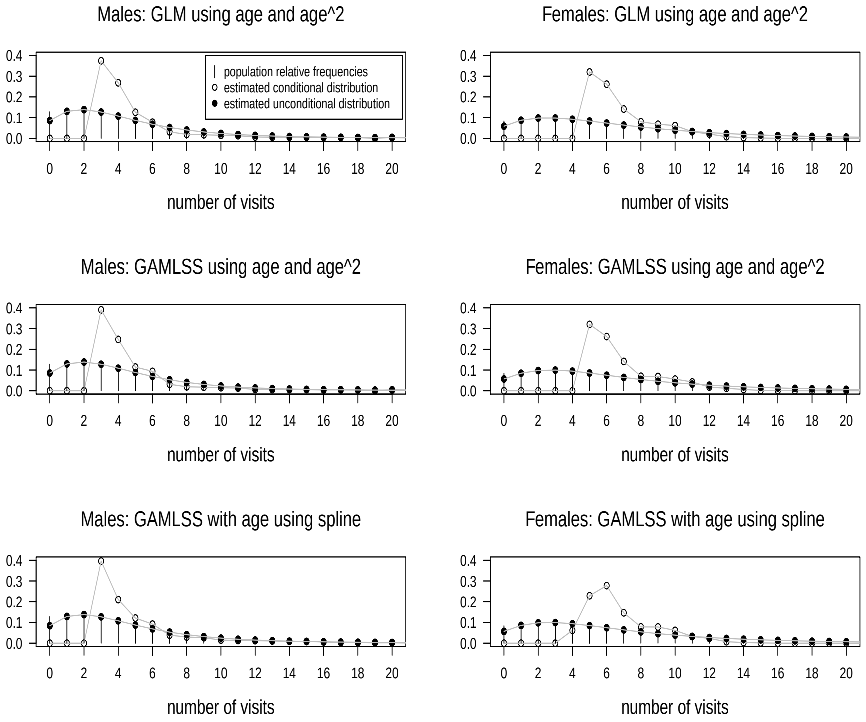


Figure 3.4.: Predicted population distribution based on negative binomial GLM estimates (upper), negative binomial GAMLSS parametric (middle) and cubic spline (lower) specification for 1994.

fits very well. While for men, it seems that it would be worth to pay more attention on a possible zero inflation, the problem is less emphasized for females.

For comparing the different GLM and GAMLSS specifications we need a more careful analysis. In order to do so, we calculated the prediction error

$$\frac{1}{M}\sum_{m=1}^{M} LOSS\,[\hat{f}_N(y_m) - f_N(y_m)], \tag{3.7}$$

where M is the number of values y does take, i.e. $1, 2, \cdots, 42$ for males and $1, 2, \cdots, 34$ for females. $LOSS\,[\cdot]$ stands simply for $abs[\cdot]$ (L1-norm) and $[\cdot]^2$ (L2-norm) respectively. The outcome is listed in Table 3.6. According to this, the negative binomial GAMLSS using spline 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).

Table 3.6.: L1 and L2-Norm prediction errors of case 1

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

The problem with the AIC is that one needs to calculate the degrees of freedom which can be quite problematic in nonparametric statistics, see e.g. Sperlich et al. (1999) or Müller (2001). Note finally that based on our observation in Figure 3.4 we also calculated the prediction errors for the zero-inflated NB GLM; surprisingly, it never outperforms the NB GAMLSS using spline.

3.3.2. Case study 2

Even more challenging - and also more interesting for health economics and political decision making - is the prediction of visits to the medical doctor for the future.

Clearly, the theoretical findings from equations (3.3) to (3.6) stay all the same. The only difference is that, at least for far horizons, it is to be expected that the relation between y and the used covariates will change. The prediction performance of our method to the future depends on the persistency of the relation we estimate from the sample. To keep the problem simple we will use the same sample, i.e. the results obtained in Section 3.2 to now predict the distribution of visits to the doctor for 1995. Applying the same procedure as we used in the first case study, we get the predictions illustrated in Figure 3.5. At first glance the prediction performance looks even better than the estimation performance in case 1. This is due to the lack of zero inflations in the recorded real visits. This now also explains why the different criteria in Section 3.2 opted for models without zero-inflation. All these criteria are constructed thinking of an infinite

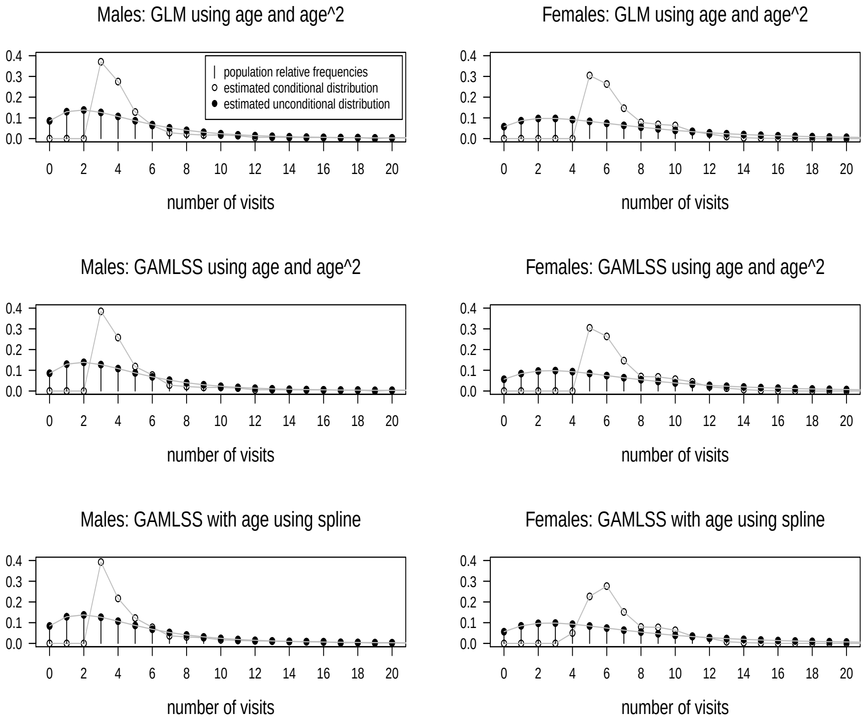


Figure 3.5.: Predicted population distribution based on negative binomial GLM estimates (upper), negative binomial GAMLSS parametric (middle) and cubic spline (lower) specification for 1995.

hyperpopulation, i.e. of a distribution from which the populations in 1994 and 1995 are just random samples. Then, a zero-inflation would fit better the 1994 data but constitutes an overfit for the hyperpopulation.

As for the first case study we again analyzed the prediction errors, see equation (3.7), of our different specifications, summarized in Table (3.7). We get a similar ranking of the specifications as in case study 1 but, as already noted from Figure 3.5, with better total performance. Again, the nonparametric NB GAMLSS clearly gives the best predictions for both the male and female populations.

Table 3.7.: L1 and L2-Norm prediction errors of case 2

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

3.4. Conclusions

The number of visits to the medical doctor is of essential interest in health economics, be it for political decision making or the planning of health care institutions and insurance. While there exist many econometric model approaches to study the demand for health care, it is hard to find a simple but effective method to estimate and predict the unconditional distribution of the number of visits. Often demographic information which is easily available even on the census level turns out to be more helpful for estimating – not to mention for prediction and scenario simulations – the distribution of the number of visits than complex econometric models. Doubtless the latter have their particular justification in more sophisticated (welfare) analysis.

A careful model specification and selection is the necessary prior step to obtain the conditional relationship between the number of visits and the demographic factors. Here, we fitted different conditional densities to the sample data. Then a well known integration principle yields a predictor for the required unconditional distribution of visits. In case the conditional sample distribution is a consistent estimator for the population analogue, this predictor inherits its asymptotic properties, in particular consistency and convergence rate. In case the population of interest follows a different distribution than the sample at hand, we still have the interpretation of our predictor as an intuitively appropriate N-fold mixture distribution. This may explain the excellent performance of our method despite its simplicity, in both case studies: for the estimation of the unconditional population distribution, and for the future prediction.

The model selection might be done via standard criteria as we used. However, one should not forget that these try to select the best estimator for the conditional distribution whereas the final objective is the unconditional one. This or the problematic calculation of the degrees of freedom for nonparametric estimators would explain that, for example, the AIC did not select the optimal model for the prediction of female visits to the medical doctor. In our case studies we were in the fortunate situation of knowing the outcome and could therefore compare the prediction with the real number of visits. In practice we recommend to evaluate the final predictor of the unconditional distribution on the observed sample. Note that due to the explicit analytic form of our estimator / predictor our results are transparent and reproducible. We do not use any random-, simulation- or resampling methods. As a flexible method it can be incorporated in any mixed effects or more sophisticated econometric models. Furthermore, it can be easily extended to any other context and allows for further inference.

R. References

Akaike, H., 1974. A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55, 117-128.

Berzel, A., Heller, G.Z. and Zucchini, W., 2006. Estimating the number of visits to the doctor. Australian & New Zealand Journal of Statistics 48(2), 213-224.

Cameron, A.C. and Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge: Cambridge University Press.

Cameron, A.C., Trivedi, P.K., Milne, F. and Piggott, J., 1988. A microeconometric model of the demand for health care and health insurance in Australia. Review of Economic Studies 55, 85-106.

Deb, P. and Trivedi P.K., 1997. Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics 12, 313-336.

Gutmu, S., 1997. Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. Journal of Applied Econometrics 12, 225-242.

Hastie, T. J. and Pregibon, D., 1992. Genaralized Linear Model. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Heller, G.Z., 1997. Who visits the GP? Demographic patterns in a Sydney suburb. Technical report, Department of Statistics, Macquarie University.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum-Likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1-38.

Jochmann, M. and León-González, R., 2004. Estimating the demand for health care with panel data: a semiparametric Bayesian approach. Health Economics 13, 1003-1014.

Little, R.J.A. and Rubin, D.B., 1987. Statistical Analysis with Missing Data. New York, John Wiley.

Müller, M., 2001. Estimation and testing in generalized partial linear models – A comparative study. Statistics and Computing 11, 299-309.

Nelder, J.A. and Wedderburn, R.W.M., 1972. Generalized linear models. Journal of the Royal Statistical Society A 135, 370-384.

Pohlmeier, W. and Ulrich, V., 1995. An econometric model of the two-part decision making process in the demand for health care. Journal of Human Resources 30, 339-361.

Rigby, R.A. and Stasinopoulos, D.M., 1996. A Semi-parametric Additive Model for Variance Heterogeneity. Statistical Computing 6, 57-65.

Rigby, R.A. and Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. Applied Statistics 54, 507-554.

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473-489.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6(2), 461-464.

Sperlich, S., 2009. A note on non-parametric estimation with predicted variables. The Econometrics Journal 12, 382-395.

Sperlich, S., Linton, O., and Härdle, W., 1999. Integration and backfitting methods in additive models. – Finite sample properties and comparison. Test 8, 419-458.

Van Keilegom, I. and Veraverbeke, N., 2002. Density and hazard estimation in censored regression models. Bernoulli 8, 607-625.

Windmeijer, F.A.G. and Santos Silva J.M.C., 1997. Endogeneity in count data models: a application to demand for health care. Journal of Applied Econometrics 12, 281-294.

Winkelmann, R., 2004. Health care reform and the number of doctor visits – an econometric analysis. Journal of Applied Econometrics 19, 455-472.

Yee, T. W. and Wild, C. J., 1996. Vector generalized additive models. Journal of Royal Statistical Society, Series B, Methodological 58, 481-493.

4. Estimating the Income Distribution from a few Quantiles

Abstract

For different welfare studies, often an estimate of the income or consumption distribution is needed even if only few quantiles (e.g. quartiles and quintiles) are available. A method for estimation of a convex function based on spline smoothing is applied for estimating the Lorenz curve from sparse data points. Compared to the currently available methods, the new estimate does not require constrained optimization. The use of the functional form for the Lorenz curve enables us to provide a parametric density that is consistent with the given quantiles. Further, we can easily derive inequality measures such as Gini coefficient. In the simulation study and an application with quintile share data on US income, it can be seen that the new estimate performs well. This oeuvre summarizes first ideas and results of an ongoing joint collaboration with Ignacio Moral-Arce from the Institute of fiscal studies in Madrid and Prof. Stefan Sperlich. The results described in this paper represents my independent work.

4.1. Introduction

When Sala-I-Martin (2006), in his seminal paper, calculated the income distribution of 138 countries he used nonparametric kernel density estimation. Typically, for such an estimation sufficiently large data sets have to be available or to be constructed artificially by different data matching, forecasting, or extrapolation methods (from neighboring countries). Often, however, quartiles, quintiles or even more quantiles are indeed available and probably more reliable than particular (small) samples or own artificial constructs. Sometimes, for many small areas we have the basic information of quantiles, we would like to be able to estimate the distribution for each as well as the total distribution, which respects the given quantiles, however here we come across a scaling problem. Our target is to construct easy-to-calculate analytical estimators based on only a few quantiles which will allow us to construct the distribution functions as well as derivatives, like the Lorenz curve, or inequality, or poverty measures. The main idea is to use a constraint spline estimator for the Lorenz curve which provides an analytic parametric specification of all the other functions and parameters of interest. The minimum of information needed are quartiles while there is no limit. However, with deciles we already get excellent approximations of the underlying distribution, and the use of more than centiles typically does not contribute new information.

A function f is a Lorenz curve if it satisfies the following properties (Ortega et al., 1991):

- $f(x) \geq 0$ for every $x \in [0, 1]$,
- $f(0) = 0$ and $f(1) = 1$,
- $f'(x) \geq 0$ and $f''(x) \geq 0$ for every $x \in (0, 1)$.

The literature on the construction of Lorenz curves is abundant. In general, the existing methods can be divided into two groups; cf. Ryu and Slottje (1999). Most methods are based on parametric estimation: either estimate the Lorenz curve under a distributional assumption for the income distribution, cf. Dagum (1980), McDonald(1984), Arnold (1983,1986), and Villaseñor and Arnold (1989), or fit different functional forms for estimating the Lorenz curve, cf. Kakwani and Podder (1973, 1976), Rasche et al. (1980), Kakwani (1980), Gupta (1984), Ortega et al. (1991), Basmann et al. (1990), Ryu and Slottje (1996). The former set of contributions starts by summarizing the given observations of microdata and then estimating the parameters under a pre-specified underlying hypothetical distribution. One then uses the estimated parameters to approximate the Lorenz curve. The latter set tries to fit the given data with a functional form that satisfies all the properties of a Lorenz curve. We noted that starting with the welfare implications of Atkinson's (1970) notion of Lorenz dominance, there exists a large literature about constructing asymptotically distribution-free statistical Lorenz curves, see Beach and Davidson (1983), Beach and Richmond (1985), Bishop et al. (1989), Bishop et al. (1991), and Gastwirth and Gail (1985). The problem with these works is that they only focus on a few Lorenz curve ordinates and provide information about discrete piecewise segments without considering the shape of the entire curve.

Although the method of Lorenz curve estimation has been much discussed, none of the above contributions was constructed to deal with the sparse data problem. In case only a few points of the distribution are available, instead of microdata sets, how can one retrieve the unknown Lorenz curve? Three different approaches to this task were reported and summarized by Braulke (1988). The approaches considered are the curve fitting method by specified functionals, as we mentioned above. It was tested using five inner points on the Lorenz curve. Unfortunately, it only gives a poor performance. The method of interpolation by a well-behaving (monotone, convex, differentiable) quadratic spline (Passow, 1977 and Lam, 1990) performs very well, as the fitted curve will definitely go through the given data points, and the estimate always lies inside the range suggested by the theoretical bounds. The midpoint of the range, spanned by the constructed upper and lower bounds, would be a possible candidate compared to the quadratic spline estimate, because the bounds were constructed so that the true Lorenz curve will be between these bounds. Unlike them, the main purpose of the present paper is to use a spline estimate of a regression function under certain shape restrictions such as monotonicity and convexity, and based on sparse data points on the Lorenz curve for estimating the entire Lorenz curve, the density, etc. The convex regression function estimate was described in detail in Birke and Dette (2007). We will illustrate that it provides good fits to the Lorenz curve of many income distributions, and allows us to easily compute Gini's Index and other inequality measures. Moreover, one can derive an explicit income density function form from our Lorenz curve estimator.

4.2. A new estimator for convex functions

Much effort has been spent on the problem of estimating a regression function under certain shape constraints, such as the underlying regression curve being positive, monotone, convex or concave, etc. using the least squares approach; cf. Hildreth (1954), Wu (1982) and Fraser & Massam (1989). One problem with the commonly used least squares technique, and also other projection-based techniques, is that they produce a rather unsmooth convex estimate, even if the underlying regression function is 'known' to be smooth.

In the present paper we apply an alternative estimate of convex functions based on the conventional smoothing methods; cf. Birke and Dette (2007). The new estimate is constructed in three steps. It starts with a strictly isotonic estimate $\hat{m}'$ of the derivative of the regression function m. Any unconstrained estimate (kernel type, local polynomial, series or spline estimator) of the regression function could be used for this purpose. In a second step a density estimate of the observations $\hat{m}'(U_i)$ is calculated, and then integrated to obtain an estimate of the inverse of the regression function. The corresponding estimate of m is finally obtained by inversion of this estimate, which is (at least) two times continuously differentiable. The consistency and asymptotic normality of the new estimate with the common rates of convergence in nonparametric regression has been proved in their paper.

Consider the nonparametric regression model

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, \cdots, n, \tag{4.1}$$

where $X_1, \cdots, X_n$ are independent and identically distributed (i.i.d.) random variables with density f and $\epsilon_1, \cdots, \epsilon_i$ are i.i.d. with mean 0, variance 1 and finite fourth moment. Further, assume that the variance function σ^2 is continuous and the density and regression function are three times continuously differentiable. Note that m is strictly convex if and only if its derivative m' is strictly increasing.

Suppose, for a set of specified population quantile indexed in increasing order

$(p_i < p_{i+1})$ and bounded in $[0,1]$, one has observed the corresponding share of aggregate income y_i. Then, the approach can be described as follows:

- Construct a strictly isotonic estimate of the derivative of the regression function $m(p)$. The penalized least squares smoothing spline is used as an unconstrained estimate of the regression function. Motivations and reasons for choosing smoothing splines as a solution for curve fitting problems were fully described, cf. the overall survey work of Wegman and Wright (1983). Silverman (1985) extended the practicability of the spline smoothing methodology.

- Construct a density estimate

  $$\frac{1}{Nh_d}\sum_{i=1}^{N} K_d\Big(\frac{\hat{m}'(i/N)-u}{h_d}\Big) \tag{4.2}$$

  from the estimated values $\hat{m}'(i/N)(i = 1,\cdots,N)$ of function m'. K_d denotes a symmetric kernel with compact support on $[-1,1]$ with finite second moment, and h_d the corresponding bandwidth converging to zero with increasing sample size n. We assume that K_d is twice continuously differentiable on its support.

- Consequently, referring to the ideas of Dette et al. (2006) for making a function estimate isotonic (which is in our case the estimate of m');

  $$\hat{\Upsilon}_{h_d} = \frac{1}{Nh_d}\sum_{i=1}^{N}\int_{-\infty}^{t} K_d\Big(\frac{\hat{m}'(i/N)-u}{h_d}\Big)du \tag{4.3}$$

  is a consistent estimate of the function $(m')^{-1}$ at the point t. Note that this function estimate is strictly increasing if h_d is sufficiently small. Consequently, its inverse is a strictly isotonic and smooth estimate of the derivative of the function $m(\cdot)$.

- The final step is to integrate the inversion of this estimate. Since the estimate $\hat{\Upsilon}_{h_d}^{-1}$ is strictly increasing (and continuous), the estimated functional

$$\hat{m}_l(p,u_0) = m(u_0) + \int_{u_0}^{p} \hat{\Upsilon}_{h_d}^{-1}(z)dz \tag{4.4}$$

is strictly convex. $\hat{\Upsilon}_{h_d}^{-1}$ is an arbitrary point at $u_0 \in (0,1)$. In our case the choice of the initial point u_0 is not crucial for the performance of the proposed procedure.

Three different spline estimates are considered in the present study while applying the convex regression function estimate of Birke and Dette (2007):

- S1: A smoothing spline estimate that minimizes the criterion:
$$\sum_{i=1}^{n}(y_i - m(p_i))^2 + \lambda \int (m'')^2. \tag{4.5}$$
The solution of this minimization problem is a piecewise cubic polynomial with the joint points at the unique set of P values. The constructed spline curve has continuous first and second derivatives, and the second and third derivatives are zero at the endpoints of the spline. It is implemented in R in the 'mi' package, see the function $sreg()$.

- S2: A so-called penalized spline estimate of the function m is the function that minimizes
$$\sum_{i=1}^{n}(y_i - m(p_i))^2 + \lambda \int (D^r m)^2, \tag{4.6}$$
which is seemingly similar to S1 but has different numerical performance. Let D^r be a linear differential operator, where r denotes the order of the derivative to be penalized. For more details about its specification see Heckman and Ramsay (2000). This modification of (4.5) is implemented in R in the 'pspline' package; see the command function $smooth.Pspline()$. Due to the sparse data we have to smooth the data using a second order polynomial spline. Then the penalty is $\int(D^2 m)^2$ and $m(p)$ is a piecewise polynomial of order $2r - 1 = 3$.

- S3: Is basically like S2 but now $m(p)$ is a piecewise polynomial of order

$2r - 1 = 4$ by setting $r = 2.5$.

There are some suggestions in the literature about choosing and using smoothing splines under certain shape constraints. For instance, Ramsay (1998) proposed to impose the smoothness penalty in (4.6), on a coefficient function, say w, which is related to the D via a differential equation. By choosing an appropriate differential equation one can ensure that D has some desired properties. More recently, Turlach (2005) proposed another approach to shape constrained smoothing using smoothing splines. However, they all propose methods that can hardly be applied on a problem with very sparse data like we face it in this work.

4.3. A simulation study

We generated a sample of size 1000 from a lognormal distribution with $\mu = 5$ and $\sigma = 1$. Consider a set of population quantiles $p_i = (0, 0.2, 0.4, 0.6, 0.8, 1)$ and the corresponding share of aggregate value, which is computed from the generated sample. First, we estimated densities from the smoothing spline estimator on a grid of 51 equidistant points $t_1 < t_2 < \cdots < t_{51}$, where $t_1 = 0$ and $t_{51} = 1$.

In this and the following sections, the kernel K_d from above (step 2) is always the Epanechnikov kernel. The bandwidth h_d is chosen as $h_d = (\sigma^2/n)^{8/35}$ with Rice's estimate (1984) $\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{[i+1]} - Y_{[i]})^2$, where $Y_{[1]}, \cdots, Y_{[n]}$ denote the observations ordered with respect to their corresponding X-values.

Due to the sparse data we restrict ourselves to either the simple quadratic, which doesn't work and is therefore not shown, the cubic and the quartic spline estimate. Figure 4.1 shows that Birke and Dette's method with S3 – i.e. the dash line – really works satisfactorily whilst capturing all the relevant information. To highlight the behavior in the region where relative large differences exist, we plotted the estimates for the vertical range of [0.2; 1]. In this region, the difference is increasing for higher quantiles.

Figure 4.1.: Estimated Lorenz curves

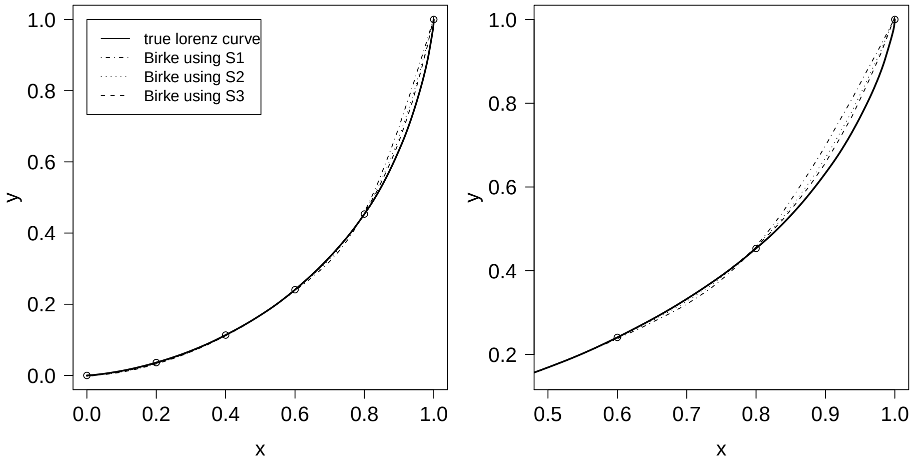


Figure 4.2.: Estimated density curves

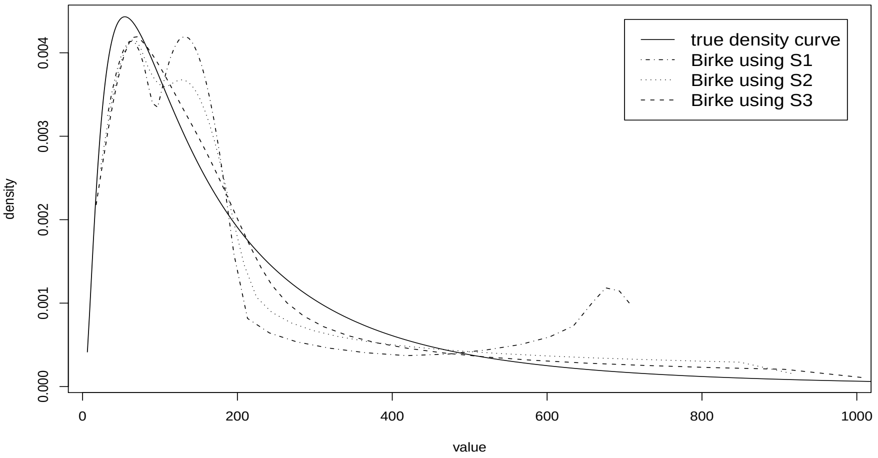


Table 4.1.: Share of aggregate value received by $10th, \cdots, 90th$ quantile

	$S1$	$S2$	$S3$
$L1-Norm$	.0975	.0468	.0306
$L2-Norm$	.0046	.0012	.0005

Following Gastwirth's definition (1971), the Lorenz curve can be derived as a function of the cumulative distribution function $F(x)$, where x is the income sample. Let $\hat{m}_l(p)$ denote the estimated empirical Lorenz curve. We have

$$\hat{m}_l(p) = \frac{\int_0^p F_X^{-1}(y)dy}{\mu_X}, 0 \leq p \leq 1, \tag{4.7}$$

where $\mu_X = \int xf(x)dx = \int xdF(x)$. To evaluate the goodness of the estimated share of aggregate values received by every 10th population quantile, we computed the L1- and L2-Norm estimation errors separately for S1, S2 and S3. The outcome, given in Table 4.1, shows that Birk and Dette's method using S3 does better than the other two.

Note that m is constructed as a continuous, nondecreasing convex function that is two times continuously differentiable and positive everywhere in $[p_1, p_2]$ (in our case $[0, 1]$). Thus the cumulative distribution function F_X has a finite positive density which is given by

$$f_X(x) = \frac{1}{\mu \hat{m}_l''(F_X(x))}. \tag{4.8}$$

Given this explicit expression, one can estimate the density at any point. For instance, according to the above formulas, the unknown income values

$$x(F) = \mu_X \cdot \hat{m}_l'(p) \tag{4.9}$$

can easily be calculated. By putting the estimated values in (4.8) one can obtain the densities at those points. The resulting density estimates are plotted in Figure 4.2. Once again it is confirmed that the method of Birke and Dette using S3 outperforms by far the others. It is the only one that really captures the mode of the true distribution.

4.4. Application on US data

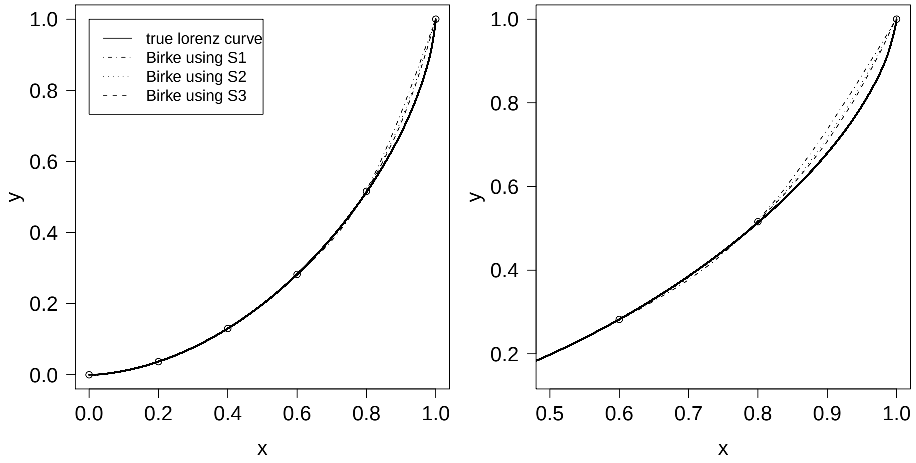
We consider the US income quintiles listed in Table 4.2. The income quintile measure is derived from the Statistics US Census data by ranking the average household income from the poorest to the wealthiest, aggregating them, and then grouping them into 5 income quintiles (1 being poorest and 5 being wealthiest), each quintile represents 20% of the population. In order to compare the estimated income distribution with the original one, we calculate the share of aggregate income received by each fifth of US households in 2000 from a micro data sample with $483,094$ observations. As can be seen, the poorest 20 percent of the population had roughly 3.7 percent of total income, the next poorest 20 percent of the population had roughly 9.3 percent of total income, etc.

Table 4.2.: Percent share of aggregate income (dollars) received by each fifth of US households in 2000

Quintile	1	2	3	4	5	mean income	number
Share of aggregate income	.037	.093	.152	.234	.484	$57,195$	$483,094$

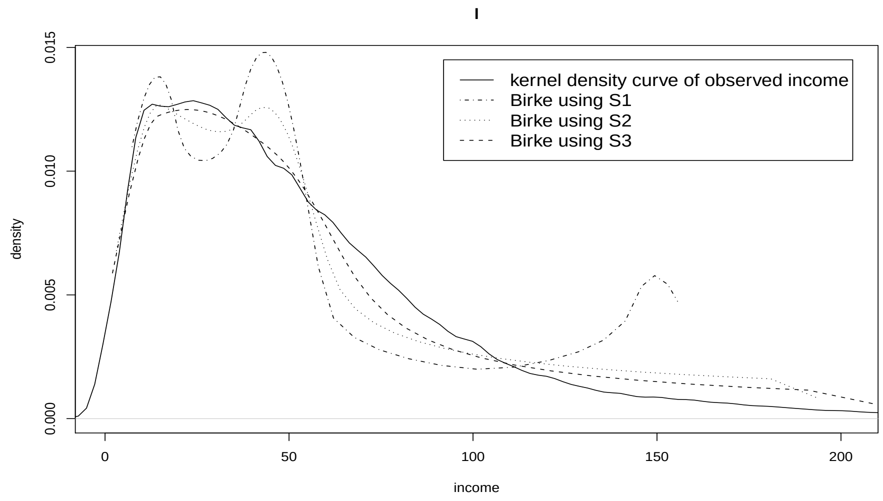
Source: U.S. Census Bureau: ACS Public Use Microdata Sample (PUMS) 2001

Figure 4.3.: Estimated Lorenz curves



Figures 4.3 and 4.4 give the function estimates of the Lorenz curve and income distribution based on the 483,094 observations, compared to our different esti-

Figure 4.4.: Estimated income distributions. Scale: x-axis 10^3, y-axis: 10^{-3}.



mates which are only based on the quintile information. As in the simulation study, the estimators using Birke and Dette's method perform best and describe the shape of the true underlying densities much better. Note that, again, all estimators pass almost exactly through the given quintiles.

Table 4.3.: Gini estimates

census estimate	$S1$	$S2$	$S3$
0.448	0.432	0.437	0.438

The Gini coefficient is commonly used as a measure of inequality of income or wealth. Based on the Lorenz curve the Gini coefficient is defined mathematically as the ratio of the area that lies between the line of equality and the Lorenz curve over the total area under the line of perfect equality, i.e. the 45% line. This ratio can be determined by $1 - 2\int_0^1 L(X)dX$. Table 4.3 compares the estimates based on the 483,094 observations to the different estimates based on the given quintiles. The resulting Gini estimates confirm what we had already seen in Figure 4.3: The estimate based on method of Birke and Dette using S3 is clearly closer to the census estimate.

R. References

Arnold, B. C., 1983. Pareto distributions. International Cooperative Publishing House, Fairland, MD.

Arnold, B. C., 1986. A class of hyperbolic Lorenz curves. Sankhyā: The Indian Journal of Statistics, Series B 48(3), 427-436.

Atkinson, A. B., 1970. On the measurement of inequality. Journal of Economic Theory 2, 244-263.

ACS Public Use Microdata Sample (PUMS), 2001. U.S. Census Bureau. http://factfinder.census.gov/home/en/acspums2001.html, accessible on March 5, 2011.

Basmann, R. L., Hayes, K. L., Slottje, D. J. and Johnson, J. D., 1990. A general functional form for approximating the Lorenz curve. Journal of Econometrics 43, 77-90.

Beach, C. M. and Richmond, J., 1985. Joint Confidence intervals for income shares and Lorenz curve. International Economic Review 26(2), 439-450.

Beach, C. M. and Davidson, R., 1983. Distribution-free statistical inference with Lorenz curve and income shares. The Review of Economic Studies 50(4), 723-735.

Bishop, J. A., Chakraborti, S. and Thistle, P. D., 1989. Asymptotically distribution-free statistical inference for generalized lorenz curves. The Review of Economics and Statistics 71(4), 725-727.

Bishop, J. A., Formby, J. P. and Smith, W. J., 1991. Lorenz dominance and welfare: changes in the U.S. distribution of Income, 1967-1986. The Review of Economics and Statistics 73(1), 134-139.

Birke, M. and Dette, H., 2007. Estimating a convex function in nonparametric regression. Scand. J. Statist. 34, 384-404.

Braulke, M., 1988. How to retrieve the lorenz curve from sparse data. In W.

Eichhorn (Ed.), Measurement in Economics, Heidelberg, Physica-Verlag.

Dagum, C., 1980. The generation and distribution of income, the Lorenz curve and the Gini ratio. Economie Appliquée 33, 327-367.

Dette, H., Neumeyer, N. and Pilz, K. F., 2006. A simple nonparametric estimator of a monotone regression function. Bernoulli 12, 469-490.

Fraser, D. A. S. and Massam, H., 1989. A mixed primal-dual bases algotithm for regression under inequality constraints. Application to concave regression. Scandinavian Journal of Statistics 16(1), 65-74.

Gastwirth, J. L., 1971. A general definition of the Lorenz curve. Econometrica 39(6), 1037-1039.

Gastwirth, J. L. and Gail, M., 1985. Simple asymptotically distribution-free methods for comparing Lorenz curves and Gini indices obtained from complete data. in: R. L. Basmann and G. F. Rhodes, Jr., (Eds.), Advances in Econometrics 4. JAI Press, Greenwich, CT.

Gupta, M. R., 1984. Functional form for estimating the Lorenz curve. Econometrica 52(5), 1313-1314.

Heckman, N. E. and Ramsay, J. O., 2000. Penalized Regression with Model-Based Penalties. The Canadian Journal of Statistics 28(2), 241-258.

Hildreth, C., 1954. Point estimates of ordinates of concave functions. Journal of the American Statistical Association 49, 598-619.

Kakwani, N.C. and Podder, N., 1973. On the estimation of Lorenzcurves from grouped observations. International Economic Review 14, 278-292.

Kakwani, N.C. and Podder, N., 1976. Efficient estimation of the Lorenzcurve and associated inequality measures from grouped observations. Econometrica 44, 137-148.

Kakwani, N.C., 1980. Functional forms for estimating the Lorenz curve: a reply. Econometrica 48, 1063-1064.

Lam, M. H., 1990. Monotone and Convex Quadratic Spline Interpolation. Virginia Journal of Science 41(1).

McDonald, J. B., 1984. Some generalized functions for the size distribution of income. Econometrica 44, 963-970.

Ortega, P., Martín, G., Fernández, A., Ladoux, M. and García, A., 1991. A new functional form for estimating Lorenz curves. Review of Income and Wealth 37(4).

Passow, E., 1977. Monotone quadratic spline interpolation. Journal of Approximation Theory 19, 143-147.

Rasche, R. H., Gaffeney, J., Koo, A. Y. C. and Obst, N., 1980. Functional forms for estimating the Lorenz curve. Econometrica 48(4), 1061-1062.

Ramsay, J. O., 1998. Estimating smooth monotone functions. Journal of the Royal Statistical Society, Series B 60(2), 365-375.

Rice, J., 1984. Bandwidth choice for nonparametric regression. The Annals of Statistics 12(4), 1215-1230.

Ryu, H. K. and Slottje, D. J., 1996. Flexible functional form approaches for approximating the Lorenz curve. Journal of Econometrics 72, 251-274.

Ryu, H. K. and Slottje, D. J., 1999. Parametric approximations of the Lorenz curve. In: Silber, J.(Ed.), Handbookk of Income Inequality Measurement, Kluwer Academic Publishers, Boston, 291-312.

Sala-I-Martin, X., 2006. The world distribution of income: falling poverty and ... convergence, period. The Quarterly Journal of Economics 121(2), 351-397.

Silverman, B. W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. Journal of the Royal Statistical Society, Series B 47(1), 1-52.

Turlach, B. A., 2005. Shape constrained smoothing using smoothing splines.

Computational Statistics 20, 81-103.

Villaseñor, J. A. and Arnold, B. C., 1989. Elliptical Lorenz curve. Journal of Econometrics 40, 327-338. North-Holland.

Wegman, E. J. and Wright, I. W., 1983. Splines in statistics. Journal of the American Statistical Association 78(382), 351-365.

Wu, C. F., 1982. Some algorithms for concave and isotonic regression. In: J.S. Rustagi and S.H. Zanakis, (Eds.) Studies in the Management Sciences: Optimization in Statistics 19, 105-116. North-Holland, New York.