Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 240

Vulnerability to poverty revisited: flexible modeling and better predictive performance

Maike Hohberg, Katja Landau, Thomas Kneib, Stephan Klasen, Walter Zucchini

October 2017

Platz der Göttinger Sieben 5 · 37073 Goettingen · Germany
Phone: +49-(0)551-3921660 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

Vulnerability to poverty revisited: flexible modeling and better predictive performance

Maike Hohberg*[1], Katja Landau[1], Thomas Kneib[1], Stephan Klasen[1], and Walter Zucchini[1]

[1] *University of Goettingen*

October 12, 2017

Abstract

This paper analyzes several modifications to improve a simple measure of vulnerability as expected poverty. Firstly, in order to model income, we apply distributional regression relating potentially each parameter of the conditional income distribution to the covariates. Secondly, we determine the vulnerability cutoff endogenously instead of defining a household as vulnerable if its probability of being poor in the next period is larger than 0.5. For this purpose, we employ the receiver operating characteristic curve that is able to consider prerequisites according to a particular targeting mechanism. Using long-term panel data from Germany, we build both mean and distributional regression models with the established 0.5 probability cutoff and our vulnerability cutoff. We find that our new cutoff considerably increases predictive performance. Placing the income regression model into the distributional regression framework does not improve predictions further but has the advantage of a coherent model where parameters are estimated simultaneously replacing the original three step estimation approach.

Keywords: vulnerability to poverty, distributional regression, generalized additive model for location, scale and shape; receiver operating characteristic curve

JEL classification: C13, C18, C52, I32

*Corresponding author, Chair of Statistics and Econometrics, University of Goettingen, Humboldtallee 3, 37073 Goettingen, Germany, mhohber@uni-goettingen.de

We thank two anonymous referees and Stephen Jenkins for helpful comments on earlier versions of this paper. We are grateful for funding from the Ministry of Science and Culture (Lower Saxony).

1 Introduction

Knowing which households are vulnerable to poverty and which are not can guide policy makers on how to efficiently allocate resources in order to prevent households from falling into poverty in the future. Although closely related, poverty and vulnerability to poverty are two different concepts. Poverty refers to a state at a (static) point in time usually measured *ex post* using household income or expenditure surveys whereas vulnerability to poverty refers to a potential state in the future, i.e. an occurrence that may or may not happen in future (Moser, 1998). Therefore, unlike poverty, vulnerability has the nature of a probability forecast, or an *ex ante* assessment of poverty risk.

Even though a few empirical applications of vulnerability to poverty measures evaluated predictive performance of their estimates (Bergolo, Cruces, & Ham, 2012; Celidoni, 2013; Feeny & McDonald, 2016; Jha & Dang, 2010; Ligon & Schechter, 2004; Zhang & Wan, 2009), very little attention has been paid to ways to improve their predictive performance. One reason is that assessing and improving the accuracy of probability forecasts requires knowledge of the outcome, that is whether or not the household did become poor. Hence, such an analysis relies on panel data which is not always available. Therefore, most authors have been concerned with developing vulnerability measures in the context of cross-sectional data (e.g. Chaudhuri, Jalan, & Suryahadi, 2002; Christiaensen & Subbarao, 2005; Günther & Harttgen, 2009; Jha & Dang, 2010; Suryahadi & Sumarto, 2003). However, the increased availability of good quality panel data in both industrialized and developing countries allows an analysis on how to improve predictive performance of vulnerability estimates in a panel data context.

Ideally, vulnerability to poverty correctly identifies households that will be poor at some point in the future while minimizing the number of households that are classified as vulnerable but will not be poor. In addition to correct identification, for practical relevance it is desirable to keep a vulnerability measure comprehensible and relatively easy to implement with data that is widely available or can easily be collected. One popular approach considers vulnerability as expected poverty (VEP), i.e. the probability of a household falling into poverty in a future period (Chaudhuri, 2003; Chaudhuri et al., 2002; Pritchett, Suryahadi, & Sumarto, 2000). However, this classification method does not always perform well in terms of prediction (Bergolo et al., 2012; Celidoni, 2013). In order to improve this existing measure, we present and assess several modifications.

The first modification is related to the form of the regression model empirical researchers use to model income, consumption or any other measure of welfare. We embed the income model in the flexible framework of distributional regression that allows for a variety of potential distributions and relates all parameters of this distribution, such as mean, scale and shape, to a structured additive predictor.[1] The

[1] Distributional regression is equivalent to generalized additive models for location scale and shape (GAMLSS Rigby & Stasinopoulos, 2005). We prefer the term "(structured additive) distributional regression" as some distributions neither have location nor scale parameters but potentially only shape parameters (Klein, Kneib, Lang, & Sohn, 2015).

second modification concerns the cutoff that classifies households as vulnerable. Often this classification is based on whether their probability of being poor in the future is equal or greater than 0.5 or, alternatively, above the observed poverty rate (e.g. Chaudhuri et al., 2002; Pritchett et al., 2000). As an alternative, we construct a vulnerability cutoff employing the Receiver Operating Characteristic (ROC) curve. The main advantage compared to the original cutoff is that it allows to take account of accuracy metrics in terms of the true positive rate (TPR) or the false positive rate (FPR). In the context of vulnerability to poverty, the ROC curve has been already used to compare predictions of a range of vulnerability measures (Celidoni, 2013). Our contribution, however, substantially differs from this kind of analysis. Instead of using the ROC curve in order to *assess* performance, we construct a new vulnerability cutoff to *improve* performance. Lastly, we make use of the time length of our data set, and assess if including more information on the income history is able to improve the results.

As we wish to retrospectively observe for many years whether a household did become poor or not, the analysis relies on a high-quality long-term panel and is conducted using 15 years of the German Socio-Economic Panel (SOEP)[2]. We find that our new cutoff method significantly improves predictive performance. Placing the income regression model into the distributional regression framework allows modeling vulnerability as expected poverty in one step since variance and mean effects are simultaneously estimated but does not yield additional benefits in terms of predictive performance.

The structure of this paper is as follows: Section 2 briefly reviews the empirical approach to measure vulnerability as expected poverty, and highlights its drawbacks this paper addresses. Section 3 presents modifications to this measure. The modifications' performance is discussed in Section 4 while Section 5 concludes.

2 Measuring vulnerability as expected poverty

The literature on the empirical assessment of vulnerability is traditionally divided into three strands: vulnerability as expected poverty (VEP), vulnerability as expected utility (VEU), and vulnerability as exposure to risk (VER).[3] The latter one, VER, retrospectively measures if an observed shock reduced welfare (for an application see e.g. Skoufias & Quisumbing, 2005). The second strand, VEU, accounts for risk preferences and defines vulnerability as the difference between a utility derived from a certainty equivalent at which the household would not be vulnerable and the expected utility derived from possible states in the future (e.g. Ligon & Schechter, 2003). Besides being difficult to interpret, this approach has been criticized for being dependent on the choice of a utility function and risk aversion parameter (Celidoni, 2013; Christiaensen & Subbarao, 2005; Gaiha & Imai, 2008). Finally, VEP considers vulnera-

[2]Socio-Economic Panel (SOEP), data of the years 1993-2008, version 26, SOEP, 2010, doi: 10.5684/soep.v26.
[3]See Klasen and Waibel (2013) for a comprehensive review.

bility as the probability of a household falling into poverty in a future period. Most applications of this approach draw on Chaudhuri et al. (2002) who estimate a regression model with consumption as the dependent variable and a covariate-dependent variance of the error term. We will base our analysis on the expected poverty concept as it is easily comprehensible, interpretable, forward looking (in contrast to VER), and has been widely applied (in contrast to VEU). Easy implementation is also the reason why we stay close to the original VEP approach and propose modifications rather than presenting yet another vulnerability measure.

Vulnerability as expected poverty defines vulnerability of an individual or a household h at time t as the probability that some measure of welfare, usually income, expenditures, or consumption, y falls below the poverty line z at time $t+1$. That is

$$V_{ht} = \Pr(y_{h,t+1} < z) \tag{1}$$

To empirically estimate this probability, most applications follow Chaudhuri et al. (2002). Using cross-sectional data, it is assumed that consumption is generated by

$$\ln y_h = X_h\beta + e_h \tag{2}$$

where y_h is consumption expenditure, X_h are household characteristics, and e_h is the error term capturing idiosyncratic shocks under the assumption of being identically and independently distributed over time. Its variance is allowed to vary with covariates across households implying a relationship between higher volatility in consumption and poverty risk. The variance of e_h is given by

$$\sigma^2_{e,h} = X_h\theta \tag{3}$$

Estimating β and θ via a three-step feasible generalized least squares (FGLS) procedure (Amemiya, 1977), yields expected consumption and variance

$$\hat{\mathrm{E}}[\ln y_h|X_h] = X_h\hat{\beta} \tag{4}$$

$$\widehat{\mathrm{Var}}[\ln y_h|X_h] = \hat{\sigma}^2_{e,h} = X_h\hat{\theta} \tag{5}$$

Under the assumption that (log) consumption is normally distributed, the probability of being poor, i.e. the vulnerability level, will be

$$\widehat{\Pr}(\ln y_h < \ln z|X_h) = \Phi\left(\frac{\ln z - X_h\hat{\beta}}{\sqrt{X_h\hat{\theta}}}\right) \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The household is then classified as vulnerable if this probability is equal or greater than 0.5 (e.g. Chaudhuri et al., 2002; Günther & Harttgen, 2009; Nguyen, Jolly, Bui, Chuong T. P. N., & Le, 2015; Novignon, Nonvignon, Mussa, & Chiwaula, 2012; Zereyesus, Embaye, Tsiboe, & Amanor-Boadu, 2016). The brief sketch of the VEP approach shows its easy interpretation as it is expressed in monetary terms.[4]

The need for an enhanced approach results from three major drawbacks of the standard approach: First, as Celidoni (2013) points out, the welfare measure is always assumed to be (log)normally distributed. While this simplifies estimation and inference, in many applications income or expenditures do not behave (log)normally (e.g. McDonald & Ransom, 2008; Sohn, Klein, & Kneib, 2015). Second, once we depart from the normality assumption, parameters other than mean and variance could be modeled to capture the full effect of covariates on the whole conditional income or consumption distribution. Third, setting the vulnerability cutoff at 0.5, neglects the variability a household faces. If the expected income or consumption equals the poverty line on the log scale, the probability in equation (6) is 0.5 independent of the standard deviation (McCarthy, Brubaker, & La Fuente, 2016). Additionally, this classification does not always perform well in terms of prediction (Bergolo et al., 2012; Celidoni, 2013). Since in practice it is beneficial to classify a large amount of households correctly, this cutoff should be optimized. We tackle all of these drawbacks by two modifications: Introducing distributional regression is aiming at the first and second point, and partly at the third point. Within this flexible framework, different distributions can be chosen to model income, and all parameters of this distribution are related to a structured additive predictor which can incorporate nonlinear effects. Using a distribution other than the (log) normal also circumvents the problem of identical probabilities irrespective of the variance when expected incomes equal the poverty line. Our approach to use the ROC curve to determine the vulnerability cutoff focuses on the third drawback of the traditional method. We thus directly address recent criticism of the traditional vulnerability threshold raised by Bergolo et al. (2012) and McCarthy et al. (2016) and propose an endogenous cutoff which improves targeting of social policy programs.

3 Estimation strategy

Measuring vulnerability as expected poverty consists of two major steps: The first step yields the estimates of an income regression model and the second step translates these estimates into a measure of vulnerability. While the next two subsections discuss the first step by presenting the dataset in Section 3.1 and the model in Section 3.2, Section 3.3 takes care of the second step where the ROC curve is used to

[4]Some modifications to this model are available that are not of relevance for this work. These modifications include for example differentiating between covariate and idiosyncratic shocks on household and community level (Günther & Harttgen, 2009), accounting for depth of poverty (Hoddinott & Quisumbing, 2003), allowing for different risk sensitivity (Calvo & Dercon, 2013), and using an individual reference line which depends on the current living standard instead of a general poverty line (Dutta, Foster, & Mishra, 2011).

model the vulnerability cutoff.

3.1 Data and variables

To demonstrate the modifications, we use the Socio-Economic Panel (SOEP), a panel study of German households starting in 1984 that is carried out by the German Institute for Economic Research (DIW), Berlin. We used the version "Soepv26" and included observations of private households covering each year from 1993 to 2008. The income model in our analysis uses equivalence income as dependent variable and several household characteristics as covariates. These covariates include variables at the household level as well as characteristics of the household head. Households with incomplete information of the household head were excluded. We used household-level cross-sectional weights and inverse staying probabilities provided by the SOEP.[5] Equivalence income is computed using the modified OECD equivalence scales[6] and adjusted for inflation using 2005 as basis year.

Special care is taken to account for the different timing structure of retrospective income and prospective household characteristics (see e.g. Frick, Jenkins, Lillard, Lipps, & Wooden, 2008). Income for year t is extracted from the records of year $t+1$ since the income reported by household members in the SOEP survey year in fact refers to the income in the previous year. However, household composition can change from year to year and can thus affect the equivalence income. We therefore use the household composition of the year in which the income accrued, and not of the survey year.[7] We assume that households with a real annual income per adult equivalent exceeding 100,000 EUR are not at risk of becoming poor in the immediate future and excluded them from our analysis. The remaining data set includes between 5000 and 8000 households per year. Table 1 provides a descriptive overview of variables used.

[Place Table 1 about here]

The information extracted at the household level includes age structure of the household, i.e. the number of children, the number of household members between 18 and 34 and between 35 and 60 years old, and the number of elderly in the household. Further included are variables related to the ownership of residence and the employment situation of the household members, namely the number of full-time employees and its quadratic term. A higher number of full-working household members is generally associated with a higher household equivalence income. However, if the number is unusually large, this might be due to a low household income forcing some members to work that would study or stay at home were they living

[5] Staying probabilities are the product of contact probability and response probability given contact. Weights were used throughout the analysis both for descriptive statistics and regression analysis as well as for calculating the true positive rate and false positive rate.

[6] These scales take the number of household members and their age into account. Weights of 1, 0.5, and 0.3 are assigned to the household head, other household members above the age of 14, and children below the age of 14, respectively (see e.g. Atkinson, 2002; Krause & Ritz, 2006; Stauder & Hüning, 2004).

[7] This introduces some bias if individuals, who accrued income in the previous year, have joined or left the household but if we chose to consider the current year household composition, a similar and arguably more problematic bias would occur as household composition and income-earning do not refer to the same year.

in richer households. The selected covariates related to the household head include age and its quadratic term, sex, marital status, education, industry (or unemployed and inactive). The fraction of unemployed and inactive appears quite large but considering the household heads' average age of about 53 years, this originates from a large fraction of retired household heads. Due to using panel data, we also included a household's past income as a covariate in all models.[8]

3.2 A distributional regression model for income

Chaudhuri et al. (2002) formulate a model in which both the mean and variance are covariate dependent. To facilitate inference, log income is often assumed to follow a normal distribution. However, often other distributions such as the Generalized Beta, the Singh-Maddala, or the dagum distribution can provide a better fit (e.g. McDonald & Ransom, 2008; Sohn et al., 2015). These distributions can have more or other parameters than location and scale. In the distributional regression framework, all of these distributional parameters can vary with covariates allowing us to not only model the expected mean but the whole conditional income distribution. That is, the conditional income distribution is given by a density conditioned on parameters θ_k, $k = 1, \ldots, K$, of which each of the K parameters is itself dependent on the explanatory variables. We thus write

$$g_k(\theta^{(k)}) = \eta^{(k)} = X^{(k)}\beta^{(k)} + \sum_{j=1}^{J^{(k)}} s_j(z_j^{(k)}) \tag{7}$$

where g_k is a link function, $\eta^{(k)}$ the predictor for the kth parameter, the matrix X_k contains covariates described in Section 3.1 which are assumed to have a linear effect and $s_j(z_{j,k})$ are smooth functions of J continuous covariates z which have non-linear effects.[9] More precisely, for the covariate *past income* we relax the restrictive assumption of a linear effect and use P(enalised)-splines (Eilers & Marx, 1996) to flexibly model its relationship to the dependent variable. As conditional distributions, we use in addition to the normal distribution of log incomes, the Singh-Maddala distribution which has been shown to provide a good fit to the SOEP income data (Biewen & Jenkins, 2002; Selezneva & van Kerm, 2016). For the two parameters of the normal distribution, we thus have

$$\begin{aligned} \hat{\mu} = \eta^{(\mu)} &= X^{(\mu)}\beta^{(\mu)} + \sum_{j=1}^{J^{(\mu)}} s_j(z_j^{(\mu)}) \\ \log(\hat{\sigma}) = \eta^{(\sigma)} &= X^{(\sigma)}\beta^{(\sigma)} + \sum_{j=1}^{J^{(\sigma)}} s_j(z_j^{(\sigma)}) \end{aligned} \tag{8}$$

[8] In an earlier version of this paper, we tested different income models with and without past income. Without past income, the model performed much worse than models including past income.

[9] The covariates do not model systemic nationwide risk directly. Our approach builds year specific models (see Section 3.3) whose differences in coefficients reflect changes in macroeconomic effects over time.

with a log link employed for σ to ensure positivity and a multiplicative connection. For the three parameter Singh-Maddala distribution we have

$$\log(\hat{a}) = \eta^{(a)} = X^{(a)}\beta^{(a)} + \sum_{j=1}^{J^{(a)}} s_j(z_j^{(a)}) \tag{9}$$

$$\log(\hat{b}) = \eta^{(b)} = X^{(b)}\beta^{(b)} + \sum_{j=1}^{J^{(b)}} s_j(z_j^{(b)})$$

$$\log(\hat{q}) = \eta^{(q)} = X^{(q)}\beta^{(q)} + \sum_{j=1}^{J^{(q)}} s_j(z_j^{(q)})$$

This model formulation differs from the traditional approach in assuming a different response distribution, hence modeling three instead of two parameters. While the common approach uses a three step Feasible Generalized Least Squares (FGLS) estimator to estimate the model (8), distributional regression models are estimated via a back-fitting algorithm that maximizes the penalized likelihood. In this way, parameters are estimated simultaneously in contrast to the step wise FGLS approach. The methodology is implemented in the `gamlss` package in R, and described in Stasinopoulos and Rigby (2007). The model in equation (8) that assumes normally distributed log incomes and includes only linear and quadratic effects serves as our comparison model. This model is similar to the one formulated in Chaudhuri et al. (2002) but we also apply the `gamlss` algorithm to this model and include past income as covariate.

After comparing the two income models, we exploit the time dimension of our data set and assess if including a household's history of past incomes improves predictive performance.

3.3 Constructing a vulnerability cutoff using the ROC curve

The ROC curve is a well-established instrument to quantify and compare the accuracy of binary diagnostic techniques in many fields (e.g. Egan, 1975; Thompson & Zucchini, 1989). In the case of vulnerability, we assess how well different probability cutoffs predict the household's future poverty status.

The approach proceeds by first fitting an income model with income in $t-1$ as the dependent variable and to predict each household's probability of being poor in t based on this model. We then order the predicted probabilities and use each one as a hypothetical vulnerability cutoff. Households with probabilities above the hypothetical vulnerability cutoff are categorized as vulnerable, the others as not vulnerable. For each possible cutoff, this classification is compared with the actual poverty status in time t leading to four different diagnosis-outcome combinations: true positives, false positives, true negatives,

and false negatives.[10]

The ROC curve is then simply a graph in which the false positive rate (FPR) is plotted on the x-axis against the true positive rate (TPR) on the y-axis. The result is a non-decreasing function with start point (0,0) and end point at the point (1,1). Roughly speaking, the faster the curve approaches the level TPR=1 the better the diagnostic method; the method is perfect if its ROC curve reaches TPR=1 straight away.

By varying the cutoff point, we can balance the TPR and FPR according to some pre-specified criteria. As the vulnerability cutoff decreases, both the TPR and FPR will increase. This characteristic of the ROC curve can be used to construct a vulnerability cutoff that satisfies some targeting measure and is less arbitrary then a vulnerability cutoff set at 0.5 probability of becoming poor.[11]

The criteria used to select the optimal cutoff can be chosen according to the aim of the policy. Popular targeting measures that balance between TPR and FPR are the targeting differential (Ravallion, 2009) and the total error rate. The former is simply defined as the difference between TPR and FPR while the latter is the sum of wrongly classified individuals (false positives and false negatives) divided by the total population. A greater value of the targeting differential and a smaller value of the total error rate indicate better targeting. Klasen and Lange (2016) interpret these targeting measures as welfare functions defined over TPRs and FPRs and place them into the unifying ROC framework. In this framework, the policymaker chooses a combination of TPR and FPR where the slope of the ROC curve equates the marginal rate of substitution between TPR and FPR that depends on the underlying welfare function. In case of the total error rate, often greater weight is attached to the FPR and the optimal TPR/FPR combination results in a more narrowly targeted program. The optimal combination is then given by

$$\frac{dTPR}{dTPR} = \frac{1 - H_0}{H_0}, \tag{10}$$

where H_0 and $-(1 - H_0)$ are weights attached to the TPR and FPR, respectively.

On the other hand, if the policymaker aims at maximizing the targeting differential, the optimal TPR/FPR combination is given by

$$\frac{TPR}{FPR} = 1, \tag{11}$$

i.e. where the slope of the ROC curve is unity. This results in a program that is more widely targeted.

[10] Regarding the choice of a poverty line, it is common practice in Germany to use a relative poverty line, set at 60 percent of the median per capita income (see e.g. Celidoni, 2013; Stauder & Hüning, 2004). As vulnerability refers here to the probability of poverty in the next period, it is necessary to specify the value of the poverty line in the next period. To avoid the additional source of uncertainty that arises from forecasting the poverty line, we use the relative poverty line of the current year, which ranges from about 10,000 to 11,700 EUR annual income and yields poverty estimates between 10 to 14 percent in each year.

[11] See Landau (2012) for an extensive introduction to using the ROC curve to measure vulnerability to poverty including profiles of vulnerable households, extensions to an n-year period and interval income data, and analyses of macroeconomic variables.

See Figure 1 to demonstrate this point.

[Place Figure 1 about here]

Suppose we aim to identify a large share of vulnerable households, we would rely on the targeting differential as the targeting measure. After calculating the TPR and FPR for each year, the optimal combination of TPR is found where the slope of the ROC curve equals unity. For all our years, the optimal point is close to TPR=0.8 but differs in FPR. Given the similar TPR for each year, and in order to provide a simpler comparison across the years, we decided to use a constant criterion to find the optimal vulnerability cutoff. For demonstration purposes, we aim at achieving a prescribed TPR of about 80 percent but other targeting criteria can be chosen here. This adaptation to the targeting mechanism is a strong advantage of the new cutoff method compared to arbitrary set values.

With the method described so far, a cutoff for the current year t is determined such that a TPR of 80 percent is reached when comparing predicted and observed incomes. However, the aim is to estimate vulnerability as a forward looking perspective on poverty. Hence, the vulnerability cutoff must be fixed *ex ante*. For this analysis, probabilities for one year in the future, $t+1$, are calculated based on the model that uses incomes in t as the dependent variable. The vulnerability line of the previous year t is used to identify vulnerable households. Alternatively, moving averages of vulnerability cutoffs of previous years can be used as well. Figure 2 shows the calculated vulnerability cutoff for each year and two different models. In contrast to the fixed 0.5 probability, the new cutoffs are lower and differ over time though for both models in a very similar way. Whether households classified as vulnerable in the current year did actually become poor in the next year is assessed using the TPR and FPR.

[Place Figure 2 about here]

To summarize our modifications, the step- by- step procedure is as follows:

1. Under different distributional assumptions, we estimate the distributional parameters of a household's income distribution in $t-1$ for each individual household.
2. We then use these estimates to get each household's predicted probability of being poor in t.
3. The probabilities are sorted in a decreasing order. Then, each of these probabilities is taken separately as a potentially vulnerability to poverty line. By comparing the classifications made under this vulnerability cutoff with the actual poverty status in t, the corresponding true positive rates and false positive rates are derived.
4. The vulnerability line that yields a true positive rate of 80 percent is adopted as the current vulnerability cutoff.
5. To estimate vulnerability to poverty in the next year, this vulnerability line is used as a cutoff to

divide households according to their predicted probabilities for next year, $t+1$, into vulnerable or not vulnerable.

6. Finally, we assess if household classified as vulnerable did actually become poor. We compare the performance of our new cutoff with the alternative 0.5 probability.

4 Results

This sections presents the predictive performance of our modifications. We start with the new cutoff and then check if further improvements can be achieved by the distributional income model. Lastly, the time dimension of our data set is exploited by analyzing if including more years of past income can improve predictions and we analyze predictions more years ahead using our modified approach.

4.1 The new cutoff

To make predictions about poverty status in the next year, we use the calculated vulnerability cutoffs (of the previous year). More precisely, households with a predicted probability for $t+1$ above the most recent probability cutoff are considered as vulnerable. Related TPR and FPR are presented in Figure 3b for the case of a model with normally distributed log incomes. Each prediction for each year is represented by a tuple (FPR, TPR).[12] The better a model performs the more tuples should lie in the 4th quadrant which corresponds to a FPR below 20 percent and a TPR of (at least) 80 percent. Note that the predictions do not exactly meet a TPR of 80 percent but are close to it as we use the previous vulnerability cutoff and not the current one. In terms of FPR, the model yields good results for all years except two that have a high FPR of about 30 percent.

The new cutoff is compared against the 0.5 approach using the same income model. Figure 3a shows that the "traditional" cutoff method does not perform acceptably. The method classifies only few households as vulnerable. That is, while yielding a low FPR, it is not able to identify those households that will be poor in the next period. This is in contrast to Zhang and Wan (2009) who found that the 0.5 probability performs well in predicting poverty status in rural China but it is in line with Bergolo et al. (2012) and Celidoni (2013) who found weak predictive performance for data from Argentina and Chile, and from Germany, Great Britain and Italy, respectively.

[Place Figure 3 about here]

Nonetheless, the low TPRs for some years are conspicuous and we put further effort in investigating why we obtain these striking differences in predictive performance. The reason is that the income model

[12] The first year we can make a prediction for is 1995 as we require the vulnerability cutoff of the previous year and the first one available is of 1994.

underlying both approaches has difficulties in predicting the vulnerability for very low incomes as our dataset only comprises about 10-14 percent poor households. Hence, observations from 'poor' households only have a small impact on the overall model's prediction ability and the model predicts overly optimistic resulting in too few households being classified as vulnerable. Our new cutoff method is able to mitigate this weakness of the income model since the cutoff is determined endogenously and directly aiming at predictive performance. In contrast, the traditional method relies heavily on the model specification and its prediction abilities that can be weak especially at the lower end of the income distribution. Since the 0.5 probability was advocated for developing countries, where the share of the poor is generally higher than in Germany, it is possible that the traditional cutoff performs better in other settings. This means that the old cutoff method cannot be readily applied to every country context. We tested the alternative of using the poverty rate as a lower vulnerability cutoff which yielded satisfactory TPRs but very high FPRs. However, both cases, using a 0.5 probability or the poverty rate, are arbitrary decisions. In contrast, our cutoff constructed on the basis of the ROC curve can be chosen to satisfy some prescribed targeting criterion and better predicts poverty status. It mitigates weak prediction abilities of the underlying income model but is still in line with the existing vulnerability as expected poverty measure.

4.2 The effects of distributional regression

In addition to the improved vulnerability cutoff, we check if further improvement can be achieved by changing the underlying income model. As an alternative to the normal distribution of log incomes, the Singh-Maddala distribution is applied and nonlinear effects of past incomes are included. None of the specifications clearly outperforms the other. Using a different distribution with parameters beyond mean and variance, seems to yield only minimal differences.[13] This might be due to using data from an industrialized country where social safety nets likely reduce shocks on households and household income distribution are relatively stable. Another reason is that including past income already accounts for a large part of the prediction such that further parameters and covariates hardly have an influence. We will examine in Section 4.3 if this also applies to including more years of past income.

It has to be highlighted that distributional regression is an attractive alternative to estimate vulnerability to poverty as it estimates mean and variance simultaneously. This has the advantage of not relying on a step wise procedure that complicates uncertainty quantification. Due to the flexibility of distributional regression, it can easily adapt to situation with non normal distributed incomes. This is likely to be the case in developing countries where more variation in income data is expected and modeling scale and shape parameters becomes more important than in our example.

[13] With a fixed cutoff, changing the underlying model is unlikely to significantly affect the vulnerability classification as the classification only changes for the households with a probability of around 0.5. Our cutoff, however, is endogenously determined meaning that it changes according to the calculated probabilities. We thus expected some differences in vulnerability classification with our new cutoff method when changing the underlying model.

4.3 Exploring the time dimension

In addition to the two main modification, we extend our analysis in two ways to take account of the rich dataset of 15 panel waves. We first include more past income data to check if this improves our predictions further. Secondly, predictions for three years ahead are evaluated. That is, a household is classified as vulnerable if it is likely to be poor in three years' time.

Regarding the first point, including more than one past income does not significantly improve our predictions. We checked the differences between including the past income of last year, of the two last years, of the three last years and of the last five years. Predictive performance of the models are similar and we thus conclude that including the last past income already explains a large part of the current income. One reason could be that incomes in Germany are relatively stable over time such that including more past incomes do not yield much further information.

So far, we only assessed if a household is vulnerable to be poor in the next year. We extend this analysis by checking the performance of our two modifications to a three years time horizon. Since we do not know the poverty line in $t+3$, we rely on the one of year t. Figure 4 shows the results for the traditional approach using the normal distribution and the 0.5 probability cutoff compared to the new cutoff method and the distributional income model using the three parameter Singh-Maddala distribution. The modified approach better identifies vulnerable household than the original one. Compared to the one year horizon, the FPR increases and is now around 30 percent for all years but still reaches a relatively high TPR.

[Place Figure 4 about here]

5 Conclusion

This paper discusses vulnerability as an important concept for policymakers that aim at preventing households to fall into poverty in the future. The majority of the empirical literature follows the concept of vulnerability as expected poverty which defines a household as vulnerable if its probability of earning an income less than the poverty line is higher or equal than 0.5. This value was arbitrarily set and only little attention has been paid to the predictive performance of this measure.

Several modifications are proposed: Firstly, we use distributional regression to model all parameters of an income distribution instead of estimating mean and variance separately as introduced by Chaudhuri et al. (2002). Secondly, to address recent criticism of the traditional vulnerability threshold, we propose a different cutoff method to differentiate between vulnerable and non-vulnerable households. Lastly, last year's income and more information on the income history is included. All suggestions are implemented using household panel data from Germany to be able to evaluate predictions for many years and to ensure

data quality.

Distributional regression for vulnerability to poverty estimation is convenient since it models mean and variance simultaneously. Comparing the three parameter Singh-Maddala distribution to the normal distribution of log incomes does not change predictive performance in our example. Similarly, once past income is included, incorporating even more information on the income history does not yield further benefits. We suggest both effects are negligible for our data set as past income already explains much of the variation and due to the use of data from an industrialized countries where social safety nets allow households to cope with idiosyncratic shocks. Future research can certainly contribute by investigating under which circumstances modeling parameters beyond mean and variance becomes necessary.

More improvement is achieved by the new cutoff. We find that in terms of predictive performance our new cutoff method outperforms the traditional approach. Due to its endogenous construction, it can mitigate weaknesses in the income generating model specification. An important advantage, in contrast to arbitrary set values, is that it can be specified according to the targeting scheme. It is thus a useful tool if researchers or policymakers have a panel data set at hand and are particularly interested in correctly identifying vulnerable households rather than in measuring overall vulnerability.

References

Amemiya, T. (1977). The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model. *Econometrica*, *45*(4), 955.

Atkinson, A. B. (2002). *Social indicators: The EU and social inclusion.* Oxford: Oxford University Press.

Bergolo, M., Cruces, G., & Ham, A. (2012). Assessing the predictive power of vulnerability measures:evidence from panel data for argentina and chile. *Journal of Income Distribution*, *21*(1), 28-64. Retrieved from `http://EconPapers.repec.org/RePEc:jid:journl:y:2012:v:21:i:1:p:28-64`

Biewen, M., & Jenkins, S. P. (2002). *Accounting for poverty differences between the United States, Great Britain, and Germany* (No. 311). Berlin: Berlin, Germany : DIW Berlin, German Institute for Economic Research.

Calvo, C., & Dercon, S. (2013). Vulnerability to individual and aggregate poverty. *Social Choice and Welfare*, *41*(4), 721–740.

Celidoni, M. (2013). Vulnerability to Poverty: An Empirical Comparison of Alternative Measures. *Applied Economics*, *45*, 1493–1506.

Chaudhuri, S. (2003). *Assessing vulnerability to poverty: concepts, empirical methods and illustrative examples* (Mimeo). Columbia University.

Chaudhuri, S., Jalan, J., & Suryahadi, A. (2002). *Assessing Household Vulnerability to Poverty from Cross-sectional Data: A Methodology and Estimates from Indonesia* (Tech. Rep. No. 0102-52). New York: Department of Economics, Columbia University.

Christiaensen, L. J., & Subbarao, K. (2005). Towards an Understanding of Household Vulnerability in Rural Kenya. *Journal of African Economies*, *14*(4), 520–558.

Dutta, I., Foster, J., & Mishra, A. (2011). On measuring vulnerability to poverty. *Social Choice and Welfare*, *37*(4), 743–761.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York, NY: Acad. Pr.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, *11*(2), 89–121.

Feeny, S., & McDonald, L. (2016). Vulnerability to multidimensional poverty: Findings from households in melanesia. *The Journal of Development Studies*, *52*(3), 447-464.

Frick, J. R., Jenkins, S. P., Lillard, D. R., Lipps, O., & Wooden, M. (2008). Die internationale Einbettung des Sozio-oekonomischen Panels (SOEP) im Rahmen des Cross-National Equivalent File (CNEF). *Vierteljahrshefte zur Wirtschaftsforschung*, *77*(3), 110–129.

Gaiha, R., & Imai, K. (2008). *Measuring Vulnerability and Poverty* (No. 40/2008).

Günther, I., & Harttgen, K. (2009). Estimating Households Vulnerability to Idiosyncratic and Covariate Shocks: A Novel Method Applied in Madagascar. *World Development*, *37*(7), 1222–1234.

Hoddinott, J., & Quisumbing, A. (2003). *Methods for Microeconometric Risk and Vulnerability Assessments* (No. 0324). Washington, DC.

Jha, R., & Dang, T. (2010). Vulnerability to Poverty in Papua New Guinea in 1996. *Asian Economic Journal*, *24*(3), 235–251.

Klasen, S., & Lange, S. (2016). *How Narrowly Should Anti-poverty Programs Be Targeted? Simulation Evidence from Bolivia and Indonesia* (Courant Research Centre: Poverty, Equity and Growth - Discussion Papers No. 213). Courant Research Centre PEG. Retrieved from `https://ideas.repec.org/p/got/gotcrc/213.html`

Klasen, S., & Waibel, H. (2013). *Vulnerability to Poverty*. London: Palgrave Macmillan UK.

Klein, N., Kneib, T., Lang, S., & Sohn, A. (2015). Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany. *Annals of Applied Statistics*, *9*(2), 1024–1052.

Krause, P., & Ritz, D. (2006). EU-Indikatoren zur sozialen Inklusion in Deutschland. *Vierteljahrshefte zur Wirtschaftsforschung*, *75*(1), 152–173.

Landau, K. (2012). *Messung der Vulnerabilität der Armut: Eine statistische Analyse mit deutschen Paneldaten* (Dissertation). Universität Göttingen, Göttingen.

Ligon, E., & Schechter, L. (2003). Measuring Vulnerability. *The Economic Journal*, *113*(March).

Ligon, E., & Schechter, L. (2004). *Evaluating different approaches to estimating vulnerability* (No. 30159).

McCarthy, N., Brubaker, J., & La Fuente, A. d. (2016). *Vulnerability to Poverty in rural Malawi* (No. WPS7769). Washington, DC. Retrieved from `http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP//`

McDonald, J. B., & Ransom, M. (2008). The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation of Related Measures of Inequality. In D. Chotikapanich (Ed.), *Modeling Income Distributions and Lorenz Curves* (Vol. 5, pp. 147–166). New York, NY: Springer-Verlag New York.

Moser, C. O. (1998). The asset vulnerability framework: Reassessing urban poverty reduction strategies. *World Development*, *26*(1), 1–19.

Nguyen, K. A. T., Jolly, C. M., Bui, Chuong T. P. N., & Le, T. H. (2015). Climate change, rural household food consumption and vulnerability: The case of Ben Tre province in Vietnam. *Agricultural Economics Review*, *16*(2), 95–109.

Novignon, J., Nonvignon, J., Mussa, R., & Chiwaula, L. S. (2012). Health and vulnerability to poverty in Ghana: evidence from the Ghana Living Standards Survey Round 5. *Health economics review*, *2*, 11.

Pritchett, L., Suryahadi, A., & Sumarto, S. (2000). *Quantifying Vulnerability to Poverty: A Proposed Measure, Applied to Indonesia*. The World Bank.

Ravallion, M. (2009). How relevant is targeting to the success of an antipoverty program? *World Bank Research Observer*, *24*(2), 205-231.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Selezneva, E., & van Kerm, P. (2016). A distribution-sensitive examination of the gender wage gap in Germany. *The Journal of Economic Inequality*, *14*(1), 21–40.

Skoufias, E., & Quisumbing, A. R. (2005). Consumption insurance and vulnerability to poverty: A synthesis of the evidence from bangladesh, ethiopia, mali, mexico and russia. *The European Journal of Development Research*, *17*(1), 24-58.

Sohn, A., Klein, N., & Kneib, T. (2015). A Semiparametric Analysis of Conditional Income Distributions. *Schmollers Jahrbuch*, *135, Proceedings of the 11th International Socio-Economic Panel User Conference (SOEP 2014)*, 13-22.

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, *23*(7).

Stauder, J., & Hüning, W. (2004). *Die Messung von Äquivalenzeinkommen und Armutsquoten auf der Basis des Mikrozensus* (No. 13).

Suryahadi, A., & Sumarto, S. (2003). Poverty and Vulnerability in Indonesia Before and After the Economic Crisis. *Asian Economic Journal*, *17*(1), 45–64.

Thompson, M. L., & Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine*, *8*(10), 1277–1290.

Zereyesus, Y. A., Embaye, W. T., Tsiboe, F., & Amanor-Boadu, V. (2016). *Participation in non-farm work and vulnerability to food poverty of households in northern Ghana.* Boston, MA.

Zhang, Y., & Wan, G. (2009). How Precisely Can We Estimate Vulnerability to Poverty? *Oxford Development Studies*, *37*(3), 277–287.

Tables

Table 1: Summary statistics: mean and proportion of the included variables

	1994	2000	2006
Household characteristics:			
hh equ.income in EUR (sd)	19357.43 (10014.40)	20779.96 (10184.69)	20637.90 (11398.48)
under 18 in abs. no.	0.45 (0.85)	0.42 (0.83)	0.36 (0.74)
18-34 years old in abs. no.	0.50 (0.74)	0.41 (0.66)	0.33 (0.61)
35-59 years old in abs. no.	0.78 (0.85)	0.76 (0.83)	0.80 (0.83)
over 60 in abs. no.	0.55 (0.74)	0.57 (0.76)	0.60 (0.78)
full working in abs. no.	0.75 (0.75)	0.68 (0.71)	0.60 (0.66)
owner	41.10	42.80	43.60
main tenant	55.60	55.00	54.50
sub tenant	3.30	2.10	1.90
Information on household head:			
age of hh head in years (sd)	52.24 (17.47)	52.20 (17.09)	53.62 (16.72)
male	59.10	56.40	56.10
education: no degree	1.40	1.20	1.50
education: 9/10th	19.40	19.00	13.80
education: vocational or high school	50.60	46.60	48.90
education: high school and vocational	3.30	5.30	5.30
education: higher vocational	8.40	10.70	9.70
education: higher education	16.80	17.20	20.90
married	54.40	50.30	47.90
single	16.20	20.90	22.10
widowed	16.80	14.60	13.10
divorced	10.70	12.20	14.40
separated	1.90	2.00	2.50
industry: inactive or unemployed	45.80	45.20	46.80
industry: agriculture	0.90	0.90	0.50
industry: energy	1.00	0.60	0.80
industry: construction and mining	8.80	8.50	6.30
industry: manufacturing	13.10	10.30	9.70
industry: trade	7.20	7.90	8.00
industry: transport	3.10	3.60	3.00
industry: bank, insurance	2.30	2.50	2.50
industry: services	17.70	20.60	22.40
n	5159	5470	7875

Note: Numbers are given in percent unless otherwise stated. Statistics are weighted using the cross sectional weights and staying probabilities provided by the SOEP.

Figures

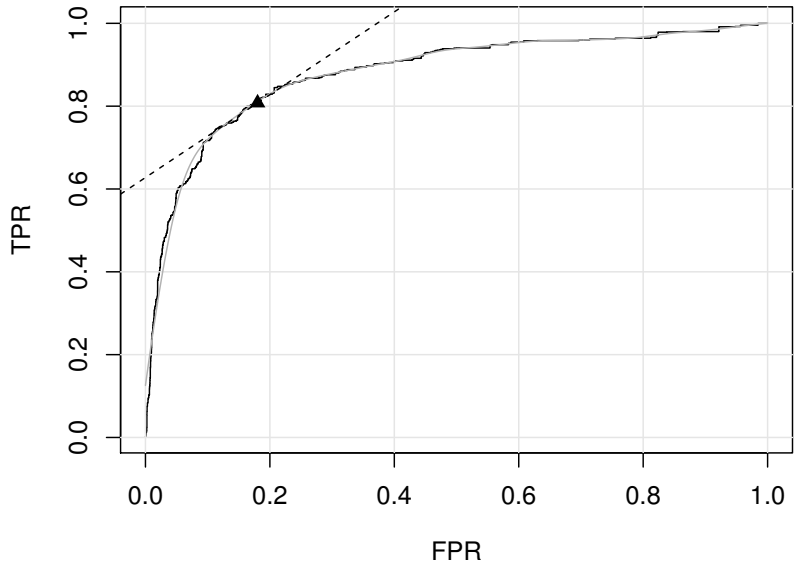


Figure 1: ROC curve based on SOEP data, year 1995. The solid black (grey) line represents the (smoothed curve of) TPR/FPR-combinations for each possible cutoff. Using the targeting differential as targeting mechanism, the dashed tangential line determines the optimal TPR/FPR combination that is marked with a triangle. Illustration is based on Klasen and Lange (2016).

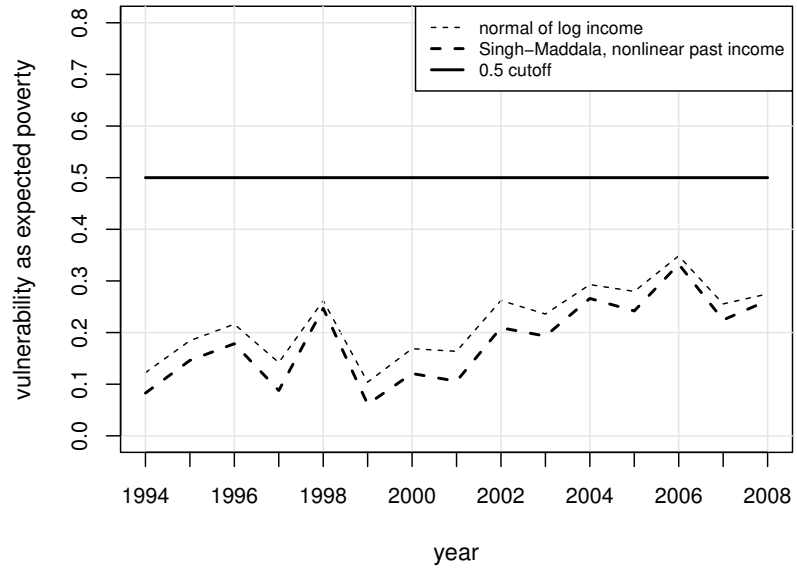
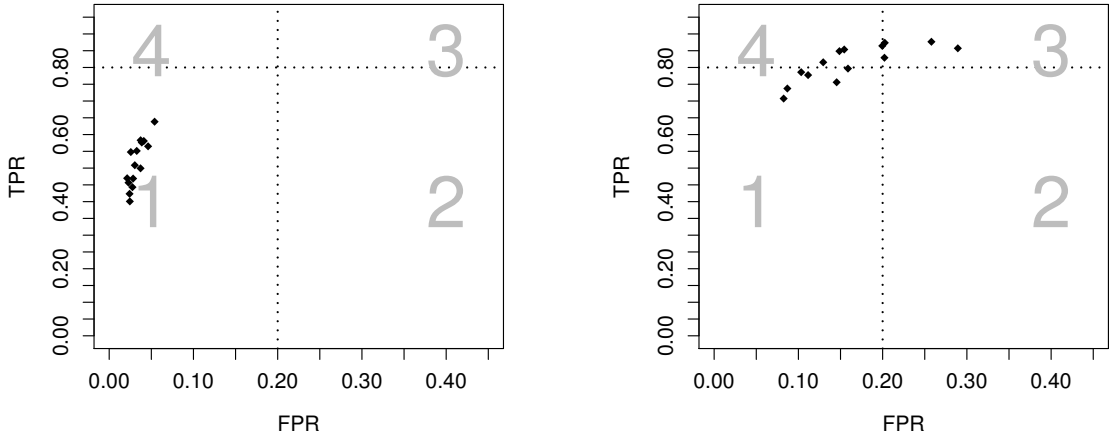


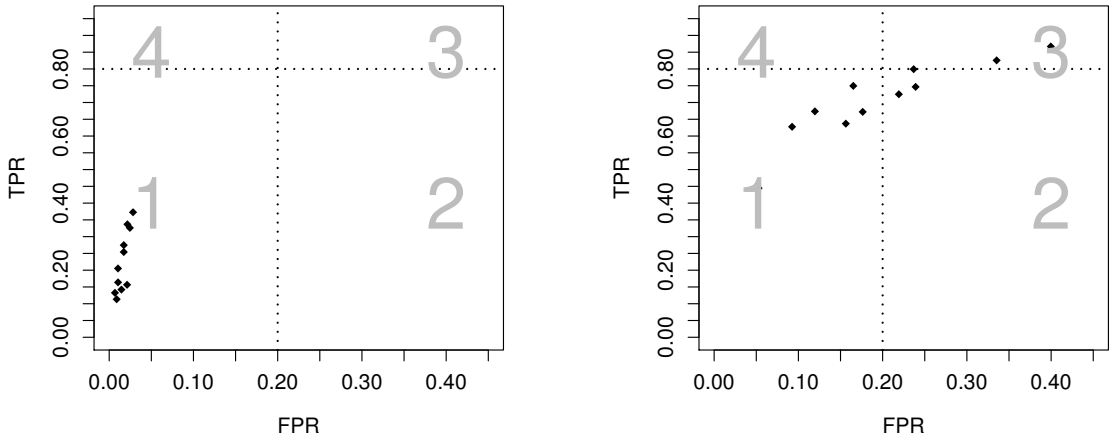
Figure 2: Varying cutoff points based on the ROC and at TPR = 0.8.



(a) 0.5 cutoff

(b) new cutoff

Figure 3: Plots of accuracy for two different cutoffs and same underlying model assuming normal log incomes. Best predictions lie in the 4th quadrant, worst predictions in the 2nd quadrant.



(a) 0.5 cutoff

(b) new cutoff

Figure 4: Plots of accuracy for two different cutoffs evaluating vulnerability for three years ahead using the same underlying model assuming normal log incomes. Best predictions lie in the 4th quadrant, worst predictions in the 2nd quadrant.