Robust Estimates of Vulnerability to Poverty Using Quantile Estimators

Christopher Oconnor

April 2022

Abstract

I propose an alternative empirical strategy—based on estimated medians with quantile regression—to standard strategies that identify individuals/households that are vulnerable to future episodes of poverty. This methodology is robust to measurement error and outliers; more accurate and rely on fewer assumptions than current mean-based methodologies and so facilitate more efficient use of scarce resources; and easy to implement which allows easy adoption by policymakers. Moreover, this methodology is shown to provide more dynamic information about household welfare than the static poverty rate. Applying this novel strategy to data from Uganda to illustrate its usefulness, I find that it more accurately identifies the future poor among the general population. The accuracy is highlighted in the fact that more than 2 in 3 households identified as vulnerable using the median strategy were poor within 1-2 years after being identified as vulnerable compared with less than 1 in 2 households using mean-based empirical strategies.

JEL Codes: C21, I32, D31

Keywords: Vulnerability, Poverty, Quantile Regression, Uganda

1 Introduction

In every country, there exists a non-trivial subset of the population that is at high risk of future episodes of poverty. In Ugandan, for example, 21.8% of households not poor in 2010/2011 became poor within 1-2 years and approximately a half became poor within 4-6 years. Static welfare measures— such as the prevalence of poverty, the severity of poverty and the depth of poverty—cannot ex-ante identify these households. Moreover, public policies based on ex-post measures of welfare may not be relevant to what is required to prevent future shortfalls in welfare. For example, policies targeting improvements in human and physical capital might be more appropriate for future shortfalls in welfare and poverty, while cash transfers and other consumption augmenting policies might be more important for current shortfalls. Ex-ante measures of welfare can be used to identify the subset of the population that are vulnerability to welfare shortfall and poverty, which can then help to inform important policy decisions.

The literature has not settled on the best technique for identifying vulnerable individuals, but the most popular way this is done is through the Vulnerability as Expected Poverty (VEP) approach (Gallardo, 2018).[1] This approach emphasizes expected future state of poverty (which can be affected or determined by exposure to risk). Under this approach, an individual is considered vulnerable if their probability of future state of poverty is above some threshold (often 50%). To characterize this probability, the probability distribution of future consumption/income will have to be determined or assumed. This is often assumed to be normal, with its parameters (the mean and variance) estimated by Ordinary Least Squares (OLS), or in some cases Maximum Likelihood (ML). Current estimation strategies used in identifying the vulnerable have been open to many criticisms such as incomparability across different proposed strategies, validity, and robustness concerns (Hardeweg et al., 2013).

[1] See Gallardo (2018) for a review of other approaches.

This paper will illustrate how, by using an estimation technique that is robust to measurement error and outliers to estimate the conditional median of consumption (Chesher, 2017), errors of identification inherent in the standard empirical methodologies are minimized and this results in more accurate identification of those who are vulnerable to poverty. This estimation technique, Quantile Regression (QR), 1) does not rely on the assumption of normality and 2) determine the underlying distribution (conditional on observed characteristics) empirically from the data. Moreover, this paper will show that using the median-based methodology provides more accurate identification and more information about the future poor than simply looking at the poverty rate. To preview the results, more than 2 in 3 households became poor within 1-2 years after being identified as vulnerable with the median-based methodology, compared to less than 1 in 2 with standard vulnerability methodologies. Using different statistical measures of prediction accuracy, the median-based methodology is shown to be more accurate than standard vulnerability measures. Moreover, among those currently non-poor that were identified as vulnerable, the median-based methodology identified a higher proportion of "poverty switchers" (non-poor who went on to become poor in the near future) compared to the proportion identified with standard vulnerability methodologies. Before outlining how median estimation using QR can be used to more accurately identify the vulnerable compared to standard methodologies, in the next section I will first provide an overview of the standard methods of vulnerability estimation.

2 Current methods of estimating vulnerability

A person is considered vulnerable to poverty if they are at significant risk of becoming poor or at significant risk of remaining poor. There are two key elements that must be considered when identifying individuals that are vulnerable to poverty: 1) a future well-being that has some possibility (risk) of being below the poverty line in the future, and 2) an estimation of

this risk of falling below the poverty line in the future. Vulnerability to poverty approaches can be classified under 4 groups (Gallardo, 2018): 1) those that emphasize vulnerability as exposure to risk which cause deviations from smooth consumption. A household is then considered vulnerable due to their inability to have smooth consumption, 2) those that emphasize expected future state of poverty. A household is then considered vulnerable if they are expected to be poor. This is often referred to as the Vulnerability as Expected Poverty (VEP) approach (Hoddinott and Quisumbing, 2003; Gallardo, 2018), 3) those that emphasize vulnerability as a utility gap, where the gap is the difference between the expected utility derived from different states of nature and that derived from a reference point (such as the poverty line), and 4) those that use a mean-risk criterion to calculate risk. Under this last category, vulnerable individuals are identified based on how far structural consumption deviates from its expected value, where deviation is estimated based on the mean/expected deviation (risk).

The VEP approach is the most popular method of identifying vulnerable individuals (Gallardo, 2018).[2] Important early papers in this line of research includes Christiaensen and Boisvert (2000), Pritchett et al. (2000), Chaudhuri et al. (2002), Chaudhuri and Christiaensen (2002), Kamanou and Morduch (2002), Chaudhuri (2003), Suryahadi and Sumarto (2003), and Christiaensen and Subbarao (2005). The idea is to first estimate the probability that future consumption is below the poverty line, that is $p(c_{t+1} < z)$, where c_{t+1} is a random variable corresponding to the distribution that future consumption is drawn from, and z is the poverty line. It can be argued that everyone observed in the current year have some non-zero risk of welfare shortfall and poverty in the future. Some criterion is thus needed to identify individuals who are vulnerable to poverty separately from the general population. In the VEP literature, a probability threshold is used for the cut off point for classifying a

[2]See Gallardo (2018) and Hoddinott and Quisumbing (2003) for a review of other approaches.

household as being vulnerable to poverty. The probability threshold most often used is 50%, such that an individual is considered vulnerable to poverty if the criterion $p(c_{t+1} < z) > 50\%$ is satisfied. This threshold level is a natural choice as an individual who is identified as vulnerable according to this threshold is more likely than not to be poor if faced with a random downside shock to their current level of consumption. Other threshold values have been used in the literature; for example, Lopez-Calva and Ortiz-Juarez (2014) used a probability of poverty cutoff of 10% as a dividing line between economic security and vulnerability, while one of the thresholds used by Chaudhuri et al. (2002) is to define an individual as vulnerable to poverty if her level of vulnerability is greater than or equal to the current poverty rate.

To characterize $p(c_{t+1} < z) > 50\%$, we must know the probability distribution of the random variable c_{t+1}. This probability distribution is almost always assumed to be normal.[3] The normal probability distribution is characterized by its mean ($E(c_{t+1}|X_t)$) and variance ($\sigma(c_{t+1}|X_t)$), both of which must be estimated. These parameters are typically assumed to be a function of observable characteristics and shocks, and are estimated via ordinary least squares or maximum likelihood (eg. Christiaensen and Boisvert, 2000; Pritchett et al., 2000; Chaudhuri et al., 2002; Christiaensen and Subbarao, 2005; Gunther and Harttgen, 2009; Zereyesus et al., 2017). After estimating the variance and expected value of future consumption, individual *i's vulnerability level* can be defined as

$$v_{i,t} = \Phi\left(\frac{z - E(c_{t+1}|X_t)}{\sigma(c_{t+1}|X_t)}\right) \tag{1}$$

Individuals who are *vulnerable to poverty* can be identified using the probability threshold and indicator function as follows:

[3]One exception to this is Lopez-Calva and Ortiz-Juarez (2014) who estimated whether future poverty status equals 1 (poor) or 0 (not poor) given current household characteristics and shocks. This estimation strategy requires the assumption of a logistic or probit link function for the distribution of the latent variable determining future poverty status. Another exception is Hohberg et al. (2018) who use the Receiver Operating Characteristic (ROC) curve in a distributional regression.

$$v_{i,t} = \mathbb{1}\left(\Phi\left(\frac{z - E(c_{t+1}|X_t)}{\sigma(c_{t+1}|X_t)}\right) > 0.5\right) \tag{2}$$

Where $\mathbb{1}(.)$ equals 1 when the condition in the bracket is true.

An issue arises where the assumption of normal distribution of c_{t+1} may lead to inclusion and exclusion errors. To illustrate this point, suppose that the poverty line is \$100 and there are four possible states of nature with equal probability and a skewed distribution. In three of the four states an individual's consumption is \$95, and in the fourth it is \$120. In this case the individual would not be expected to be poor even though they face a 75% chance of being in poverty. By assuming that the distribution of future consumption is normal, the expected value of future consumption will erroneously imply that the individual has less than a 50% probability of future poverty since it is above the poverty line. Therefore, accurately characterizing the true conditional distribution can improve vulnerability identification and improve potential policy intervention. Moreover, the assumption of normality may not hold even on the log scale (Koenker Hallock, 2001), and even if it holds the results of log-transformed data may lead to biased estimates of the untransformed raw variable (Manning, 1998; Manning and Mullahy, 2001). These transformations of data should be used parsimoniously, and generalized estimation models should be preferred (Feng et al, 2014)

The inaccuracies of current empirical strategies used to estimate vulnerability is also likely due to their reliance on OLS, in most applications, to estimate the conditional mean and variance as much as their reliance on the assumption that the underlying distribution of future consumption is normal. OLS assumes normally distributed errors that are uncorrelated with independent variables. The errors in consumption expenditure or income data are rarely homoskedastic, and papers in the vulnerability literature tries to take this into account by using the Feasible Generalized Least Squares (FGLS) approach developed by Amemiya (1977). FGLS is done by assuming some form of the covariance of the error

matrix, obtaining a consistent estimator of that matrix and then conducting a weighted OLS using the estimated covariance of the errors matrix. Examples of vulnerability papers using this technique are Chaudhuri et al. (2000), Christiaensen and Subbarao (2005), Chiwaula et al. (2011), and Zereyesus et al. (2017). There are several issues with the FGLS approach: 1) the FGLS estimator is based on asymptotics, so it may be even less efficient than OLS for small or medium sized samples (Greene, 2002); 2) the FGLS estimate still uses OLS, where it adjusts how the estimated coefficients relate to the mean by taking into consideration heteroskedastic errors. The estimates do not inform on how the errors are distributed about the mean;[4] 3) FGLS is reliant on the assumption that the covariance of the errors matrix is correctly assumed (Woolridge, 2002).

3 Proposed empirical methodology

Due to the limitations of current methodologies identified in the previous section, this paper proposes the use of a more general estimating method based on the median: Quantile Regression for the 50th percentile. This paper will show how this strategy leads to improved accuracy in the estimates of household vulnerability to poverty.[5] After estimating the conditional median of consumption expenditure ($Q_{0.5}(c_{i,t+1}|X_t)$), households vulnerable to poverty can now be identified by the updated condition:

$$v_{i,t} = \mathbb{1}(p(c_{i,t+1} < z) > 0.5) = \mathbb{1}(Q_{0.5}(c_{i,t+1}|X_t) < z) \tag{3}$$

There is no reason to separate, in time, a random variable from its effective realization (Gallardo, 2018). For example, in a die roll experiment, if the die is rolled in the current period and lands on a 3, this does not change the distribution of the random variable ($X \in$

[4]In other words, they do not make the errors normal about the mean, merely adjust the estimates to take into consideration heteroskedasticity.

[5]The unit of analysis in this paper will be the household and not the individual level.

$(1,6)$) at all. If a roll of the die is made again in another period, the distribution from which the result of the die throwing experiment is drawn from will remain the same as long as the characteristics of the die remains the same. Therefore, assuming this sort of time stationarity, I can rewrite equation (3) to identify the vulnerable:

$$v_{i,t} = \mathbb{1}(Q_{0.5}(c_(i,t)|X_t) < z) \tag{4}$$

This paper will show that estimating the median of the conditional distribution of consumption expenditure will help to minimize inclusion and exclusion errors and provide more accurate information on the future poor when compared with the standard empirical methodologies in the vulnerability to poverty literature. It will achieve this by 1) not relying on any assumptions about the distribution of future consumption, and 2) determining the underlying distribution (conditional on observed characteristics) empirically from the data. Moreover, using QR to estimate the conditional median will help to overcome the empirical drawbacks of OLS based empirical strategies like those outlined in the previous section of this paper; specifically: 1) it will not rely on the assumption about the conditional distribution of the errors and there is no need to adjust estimates to take it into account if non-normal, 2) no need to assume the form of the covariance of errors matrix, 3) no need to log transform the data to force (unconditional) normality of the response variable, 4) it is more robust to measurement error and outliers which are features commonly seen in household survey data.

The QR estimator is a special type of m-estimator (Woolridge, 2002). An m-estimator is the parameter that solves the following minimization problem: $\min_{\beta \in \mathbb{B}} \frac{1}{N} \sum_{i=1}^{N} q(w, \beta)$, where w is a vector of random variables. If $q(w, \beta)$ is squared deviations, this gives the usual OLS estimate of β. If $q(w, \beta)$ was instead absolute deviation (eg. $q(w, \beta) = |y - x\beta|$), this gives us the estimate of β corresponding to the conditional median of y. The seminal paper by Koenker Bassett (1978) showed that by using a asymmetrically weighted (tilted) absolute value function, defined as $\rho(\tau)$, this provide estimates at every τ^{th} quantile, not just the

median. Each quantile estimate ($\beta(\tau)$) is then the solution of the following minimization problem $\min_{\beta(\tau)\in\mathbb{B}} \frac{1}{N}\sum_{i=1}^{N}\rho(\tau)|y - x\beta|$.

Lengthy panel data is most desirable in vulnerability estimation as it allows the linking of changes in household characteristics with changes in their welfare by controlling for household specific effects. However, in the literature vulnerability is most often estimated using cross-sectional or short panel data as lengthy panel data is often not available for developing countries.[6] Cross-sectional models are appropriate even when the panel length is short.[7] This paper will illustrate how estimating the conditional median using cross-sectional data provides better estimates of vulnerability compared to standard empirical strategies that estimate vulnerability in the cross-section.

4 Comparison models

The median-based (quantile) model will be compared with two other vulnerability models and in some instances to the static poverty rate. These two vulnerability models are mean-based and generally consistent with the empirical methodology of most Vulnerability as Expected Poverty measures. Comparing these two models to the median-based estimates of vulnerability will help to determine whether median-based models provide more and better information on predicting the future poor than standard methods. The two vulnerability models are based on papers by Zereyesus et al (2017) [ZETA] and Gunther and Harttgen (2009) [GH].

[6] If sufficient panel length is available, the quantiles-via-moments estimator developed by Santos Silva Machado (2019) provides reasonably good estimation of quantile models. There is some bias in the estimates using this technique, but as noted by the authors the bias can be removed using the split-panel jackknife procedure of Dhaene and Jochmans (2015).

[7] Using a short panel with variables that are fixed or slow changing will result in effects not being identified due to lack of variation (Vial Hanoteau, 2015).

Comparison between the results of the median and mean based methodologies will be done in different ways. First, the data used in this analysis—Uganda National Panel Survey (UNPS)—will be divided into a training wave (2010/2011) and test waves (2011/2012, 2013/2014 and 2015/2016).[8] Second, using the training wave I will estimate the vulnerability to poverty rate for Uganda using the Median, ZETA and GH methodology, along with the standard errors and confidence interval of each estimate. This will give an idea of how many households are being identified as vulnerable under each methodology. Third, I will test the predictive performance/internal validity of each methodology by comparing how well they identify the future poor. To do this, for the test waves of the data I look at the poverty status of households identified as vulnerable in the training wave. The idea is that the most effective measure of vulnerability will have a higher success rate in the ex-ante identification of the future poor. Fourth and finally, I will utilize different measures to evaluate the predictive capabilities of each empirical vulnerability measure. These measures include the rates of inclusion and exclusion errors, statistical accuracy, and a computation of the Matthews Correlation Coefficient of each empirical method.

4.1 Zereyesus et al. (2017) [ZETA]

The model used in ZETA is based on one of the most cited and influential paper in the vulnerability literature, Chaudhuri et al. (2002). The authors used three-step feasible GLS (FGLS) to model consumption. Consumption data is rarely homoskedastic, and therefore FGLS estimation is used to control for this. The first stage involves running an OLS model of consumption on its determinants. The second stage has two steps: 1) from this first stage, the residuals are collected, squared and used as the dependent variable in a second stage regression involving the same independent variables from the first stage. 2) the inverse of the

[8]The UNPS data is publicly available at `https://microdata.worldbank.org/index.php/catalog/lsms` (retrieved March 6, 2020).

predicted values from the first step of this second stage is used to weight the entire model in the first step. The prediction from this new model in the second step is the asymptotically efficient estimate of the variance of consumption. For the third stage, the inverse of the square root of the predicted values in the second stage (that is, the inverse of the estimated standard deviation of consumption) is then used to weigh the model from stage one. The final predicted value from this stage is the estimated mean of consumption.

Using the mean and standard deviation characterized from the three-stage procedure above, the vulnerability estimates of households is determined by assuming that the distribution of consumption is log normal. This is the procedure that is used to calculate all vulnerable to poverty estimates in the subsequent sections that are attributed to the ZETA empirical methodology.

4.2 Gunther and Harttgen (2009) [GH]

The independent variables that are used to determine vulnerability estimates are often based on household characteristics, community characteristics, and shocks. Gunther and Harttgen (2009) recognized that, due to the inclusion of community variables which are common to all households within a community, the i.i.d. assumption of OLS may be violated. To control for this, they modeled consumption using multi-level modelling. The estimation procedure is done in two steps: the first step involves running a multilevel regression of consumption on household characteristics, community characteristics, and shocks. Residuals due to household variation, those due to community level variation and those due to total variation are then collected and squared. In a second step, heteroskedaticity is then controlled for by running a regression of the squared total residuals in the first stage on the household characteristics, community characteristics, and shocks. This also allow the estimate of consumption variance to depend on observable characteristics. The predicted values from the second step represents the variance of consumption. The predicted values from the first step

will represent the mean of consumption. The estimated variance and mean of consumption are then used, along with the assumption that conditional consumption is log normally distributed, to calculate estimates of vulnerability. This is the procedure that is used to calculate all vulnerable values in the subsequent sections that are attributed to the GH empirical methodology.

5 Estimation Model and Data

Median Consumption for household i is modelled through the following process:

$$Q_{0.5}(c_{i,t}) = x_{i,t}\beta + x_{i,t}\theta_{i,t}\gamma + \theta_{i,t} + e_{i,t} \tag{5}$$

The deterministic part of household consumption, $x_{i,t}$, include household/individual characteristics (eg. education of household head, labor and housing characteristics) and the economic and social characteristic of the environment (eg. access to markets, roads and electricity). The variables $\theta_{i,t}$ and $e_{i,t}$ denote observed (eg. floods or death of income earner) and unobserved idiosyncratic shocks, while the interaction term $x_{i,t}\theta_{i,t}$ denote coping strategies.[9]

If shock data are available, they should be included in empirical models used to estimate vulnerability to poverty as they assist in the unbiased estimation of the marginal effects of included variables (Ward, 2016). However, based on specification in (5), the relationship between included household characteristics and shocks is related to the interaction term since shocks, even when observed, are otherwise completely random. If the effect of the interaction

[9]Some studies on vulnerability separated observed covariate (common) shocks from observed idiosyncratic shocks. However, the data used in the present paper allowed me to combine observed covariate shocks into observed idiosyncratic shocks. For instance, I can determine whether a household in a locality was affected by a flood or some other covariate shock, regardless of whether the community overall experienced this shock. Since communities are drawn along sometimes trivial boundaries, it is possible that some households in certain parts of the community did not experience a particular shock while households in other section of the community did.

term is negligible, then not including shocks should still provide reasonably good estimates of the marginal effects of the other variables.

The dataset that is being used to implement the quantile estimation methodology is the Uganda National Panel Survey (UNPS). This is a nationally representative panel that began in 2005 (the baseline year) and is collected with technical assistance by the World Bank. From this panel there are 5 total waves conducted between 2009 and 2016[10]. Four of these waves (between 2010 and 2016) will be utilized to conduct the analysis that will be carried out in this paper[11]. The 2010/2011 wave will be used as a *training dataset*, and the 2011/2012, 2013/2014 and 2015/2016 waves will be used as the *test datasets*. The UNPS include information on individual and household characteristics. It also collects information on the communities where the households are located, a comprehensive agriculture questionnaire for households engaging in this activity, and it includes data on a range of shocks that the households and the community faced. The coverage of the sample is around 3000 households across Uganda for the test and training years. Attrition rate between the training year (2010/2011) and the first test year (2011/2012) was 7.6%. After every 3 waves, a third of the sample is randomly refreshed to ensure the overall sample keeps representativeness with the general Ugandan population. Therefore, for the 2013/2014 wave, which was the 3rd wave after the 2009 wave, approximately a third of the sample were randomly dropped (total attrition was around 40% due to the combination of planned and natural attrition). The Ugandan poverty line used in this paper is from Levin (2012), which was 516,546.7 UGX per adult equivalent in 2010 (this equated to $1.67 per person per day in purchasing power parity (PPP) in 2010)[12].

[10] Baseline data was collected in 2005 which could be considered as a 6^{th} wave.

[11] The first wave was excluded since it was conducted during the great recession of 2008-2009, and statistical noise associated with that event may be strong.

[12] Levine (2012) generated the poverty line for Uganda using 1997 as the base year, and for this paper the poverty line was scaled up to 2010 prices using CPI from the World Bank World Development Indicators website. The PPP exchange rate (private consumption) was also retrieved from the same website.

Table 1 provide summary statistics. Average annual adult equivalent consumption expenditure in 2010 was 819,885.9 Ugandan Shillings (UGX), which is approximately PPP $981.4. As with many low-income countries, the employment rate in Uganda is high due to high levels of self-employment, and so it is meaningless in the estimation of standard of living. Instead, I use formal employment and a measure of whether a household head is self-employed (enterprise owner) as a way of estimating the effect of labor market participation on consumption expenditure[13]. On average, 10.7% of individuals in household are in formal employment and 30.4% of household heads are enterprise owners. The average household size is 6.4 individuals, while about 3.1 children 0-14 years old are in each household. Some 12.1% of households use electricity, while 72.3% of households have access to "protected" water sources. About 41.4% of households obtained credit either from a financial institution or from friends or relatives.

At the community level, 60% of households are in communities with some educational institution, and 19% are in communities with tarmac roads. About 26.9% of households experienced drought or irregular rain conditions during the year the data covers. Less than 1% of households had an income earner that lost their source of income during the year, while about 2.6% of households faced some shock from pest/livestock disease.

6 Results and discussion

The final model used in the estimation of median consumption is based on a subset of variables from the list of variables in Table 1. The process of final model selection is inspired by the procedure used by Gunther and Harttgen (2009). Quantile regression is computationally

[13]Formal employment is defined as individuals who work in businesses that are registered for Value Added Tax, Income tax or where the workers in that business are in the Pay As You Earn (PAYE) system.

intensive, and a large design matrix can make estimation fail, so having a manageable set of the most important covariates is useful. The process of model selection involves first running two separate quantile regressions: one on household characteristics including interaction with shocks, and another with community characteristics including interaction with shocks. Insignificant interactions from these 2 models were removed from analysis. In a second stage, a quantile regression was run with all household and community characteristics, shocks and the significant interaction terms from the first stage. A third stage further reduced the number of interaction terms by removing the insignificant terms from the second stage and running a final quantile regression with all household and community characteristics, shocks, and the significant interaction terms from the second stage. The results of this model are illustrated in Table 2.

The predicted values from the regression in Table 2 were then used to predict the median consumption of households in the dataset. The median consumption was then compared to the poverty line to identify households who were vulnerable to poverty. For the ZETA and GH measures of vulnerability, the mean and the variance of consumption were first estimated. These values, along with the poverty line were used as parameters inside the normal cumulative distribution function to estimate the probability that future consumption will be below the poverty line. If this probability is above 50.0%, then the ZETA and GH measures will identify that household as containing individuals vulnerable to poverty. Table 3 provide estimates of the percentage of Ugandan households that were identified as vulnerable to poverty by the Median, ZETA and GH empirical methods.

ZETA identified a higher proportion of households as containing vulnerable persons than the GH or Median method (see Table 3). Specifically, ZETA identified 46.7% of households as vulnerable, GH identified 39.2% as vulnerable and the Median measure identified 40.4% of households as vulnerable.

The main comparisons are to evaluate the effectiveness of the Median empirical method with the ZETA and GH empirical methods. A relatively effective measure of vulnerability will have a higher success rate in the ex-ante identification of the future poor compared with alternative measures. I exploit the test datasets to do this evaluation. The test datasets (test years) are 2011/2012 (henceforth year $t+1$), 2013/2014 (henceforth year $t+2$) and 2015/2016 (henceforth year $t+3$). Table 4 gives the proportion of households identified as vulnerable in year t that became poor in year $t+1$, year $t+2$ or year $t+3$. Of those identified in year t as vulnerable to future poverty according to the ZETA and GH method, less than a half of them became poor in year t+1 and less than three-fourths became poor by year $t+3$ (4-6 years later). In comparison, 69.5% of households identified as vulnerable according to the Median methodology became poor in year $t+1$, and 9 in 10 of them became poor by year $t+3$. For those who are identified as vulnerable in year t according to the ZETA and GH measures, a lower proportion of them will ever be poor in either the training year or test years compared to the Median measure, where more than 9 in 10 (92.1%) are poor or will eventually become poor by year $t+3$.

About 21.8% of households not poor in year t became poor in year $t+1$, and 50.6% became poor by year $t+3$. Static poverty cannot ex-ante tell us which households these are, but vulnerability measures can provide information about the future poor from among the currently non-poor. Table 5 shows the poverty status in test years for those vulnerable to poverty but not poor in training year (year t). For those predicted to be vulnerable according to the Median methodology, 42.7% of them became poor in year $t+1$ and 74.1% of them became poor by year $t+3$. For households identified as vulnerable according to ZETA/GH methodology, 27.3%/28.7% of them became poor in year $t+1$ and 47.0%/49.0% of them became poor before year $t+3$. This illustrates the superiority of the Median methodology over standard methodologies in identifying the future poor from among the current non-poor.

The third and final type of comparison I will make between the Median methodology and the standard methodologies is in regards to the predictive capabilities of each method according to 1) inclusion error, 2) exclusion errors, 3) accuracy and 4) Matthews Correlation Coefficient. Evaluating predictive capabilities using these metrics is difficult since *true* vulnerability is unobservable. Moreover, observed future episodes of poverty does not guarantee that the household *truly* had a greater than 50% chance of future poverty. If a household is observed as poor in any of the test rounds of the data after being identified as vulnerable in the training round, this could be due to mean reversion and not due to them truly being vulnerable to poverty. With this drawback in mind, I will introduce two arguably policy relevant definitions for households that are *truly* vulnerable to poverty: i) a household is *truly* vulnerable to poverty if they were poor in at least 2 of the 3 test waves of data after being identified as vulnerable; ii) a household is *truly* vulnerable to poverty if they were poor in any of the 3 test waves of data after being identified as vulnerable. Table 6A provides information on the predictive capabilities according to the first definition of the *truly* vulnerable, and Table 6B does this according to the second definition of the *truly* vulnerable.

When the truly vulnerable is defined according to the first definition, inclusion and exclusion errors are higher in the GH and ZETA methodology compared to the Median methodology (see Table 6A). Moreover, these errors are higher than the Median methodology even if we only look at those who are currently poor and consider them as vulnerable. For GH inclusion/exclusion is 31.8%/50.9%, while for ZETA it is 50.5%/49.4%. In comparison, for the Median methodology, inclusion/exclusion errors are 24.1%/33.6%.

Ranking vulnerability measures according to inclusion and exclusion can also be supplemented by using a confusion matrix to derive 1) a measure of statistical accuracy, and 2) the Matthews Correlation Coefficient. Accuracy is the proportion of households that are

correctly classified as vulnerable + non-vulnerable (true positives and negatives). True positives (negatives) are the total number of households that were (were not) poor at least twice in the 3 test waves. False positives [negatives] are the total number of households counted as vulnerable under each methodology but were not [were] poor at least twice in the 3 test waves. For the Median vulnerability measure, accuracy is 69.2%, while it is below 50% for the other vulnerability measures (see Table 6A). Next, I can calculate the Matthews Correlation Coefficient (cf. Matthews, 1975). The Matthews Correlation Coefficient gives an estimate of the predictive accuracy of a (usually) bivariate classifier. The correlation coefficient ranges from -1 to 1, where 1 indicates that the empirical method (vulnerability classifier) provides perfect classification, and 0 indicates that there is no difference between the classifier and random assignment. The Matthews Correlation Coefficient requires 4 inputs (the number of false and true positives and negatives) and considers false classifications, while the statistical accuracy measure does not. The Matthews Correlation Coefficient for each empirical method is reported in Table 6A. The results show that the Matthews correlation coefficient for the Median vulnerability method is higher than the corresponding value for the ZETA and GH empirical methods. Moreover, the ZETA method only marginally improved upon random assignment with a correlation coefficient value close to 0, while for the Median based methodology it is 0.42527. Furthermore, the correlation coefficient is higher for the Median methodology than for the currently poor if the currently poor are classified as vulnerable to poverty

When the truly vulnerable is defined as households that were poor in any of the 3 test waves of data, the results of the comparison are consistent with the first definition of the truly vulnerable (see Table 6B); inclusion and exclusion errors are lower for the Median methodology compared with the ZETA and GH methodologies, accuracy is higher for the Median methodology, and lastly the Matthews Correlation Coefficient is higher for the median-based methodology. Overall, the results show that regardless of the measure of predictive perfor-

mance, the median-based method outperforms those that are mean-based.

7 Conclusion

This paper proposes an empirical strategy to estimate vulnerability to poverty that is robust to measurement error and outliers, and more accurate and useful compared to standard methods of vulnerability estimation. This strategy involves the use of quantile regression to estimate the median of the conditional distribution of consumption expenditure. The standard way in the literature of defining vulnerability is based on the probability (or expectation) of future poverty (Vulnerability as Expected Poverty approach). Specifically, in this paper households are identified as vulnerable if their probability of future poverty is greater than 50.0%. That is, they are considered vulnerable to poverty if they are more likely than not to be poor. Using this definition of vulnerability, estimates were computed based on three empirical strategies. The main strategy is the one being proposed by this paper; a quantile regression based empirical strategy that computes the conditional median. The other two strategies are based on the studies of Zereyesus et al. (2017) [ZETA] and Gunther and Harttgen (2009) [GH] that are mean-based methods. These two papers were chosen as the empirical strategies they employed are consistent with most mean-based empirical strategies in the Vulnerability as Expected Poverty literature. ZETA employs a feasible GLS type method that adjusts the estimates for heteroskedasticity, while GH employs a multi-level/mixed regression that takes into consideration violations of the i.i.d. assumption of OLS.

The results show that 40.4% of households contained vulnerable individuals according to the quantile methodology. The ZETA and GH methods identified 46.7% and 39.2% of households as vulnerable to poverty, respectively. The difference in the number of persons identified as vulnerable across strategies likely has to do with the fact that the median-based

quantile regression method explicitly finds the center of the conditional distribution of the welfare measure (household consumption), while the mean-based models assume that the conditional distribution is (log) normal and finds the first moment.

To evaluate the advantage that the median-based method has over the mean-based measures, out of sample analysis was conducted using waves of the Uganda National Panel Survey not used for estimation. A relatively effective measure of vulnerability will have a higher success rate in the ex-ante identification of the future poor compared with alternative measures. Of those who are identified as vulnerable in the current year, according to the ZETA and GH methodology less than a half were poor a year later and less than 3 in 4 became poor within the next 4-6 years later. For the Median methodology, more than 2 in 3 households identified as vulnerable were poor a year later and about 9 in 10 were poor within the next 4-6 years. The main benefit of the vulnerability measures over static poverty measures is that it can provide dynamic welfare information about households that are not currently poor. About 50.6% of households switched their poverty status from not-poor to poor between 2010 and 2016. Of the households that were vulnerable but not currently poor according to the ZETA and GH methodology, less than half of them switched their poverty status to poor within 4-6 years later. The Median methodology performed better in this regard as more than two-thirds of households that were not poor but vulnerable became poor within 4-6 years after the current year.

To further evaluate the advantage that the median-based vulnerability method has over the mean-based methods, various measures of predictive capabilities/quality were examined under the assumption that truly vulnerable households (that each empirical method wish to identify) will be poor in at least 2 of the 3 test waves. Results show that errors of inclusion (incorrectly counting those not vulnerable to poverty as such) and exclusion (excluding those who are rightly vulnerable to poverty) were lowest in the median-based method com-

pared to the two other empirical methods. Statistical accuracy, which is the proportion of all households correctly classified as vulnerable + non-vulnerable, was also higher in the Median vulnerability measure. The last measure of predictive capability analyzed was the Matthews Correlation coefficient (MCC). This number ranges from -1 (worst) to 1 (best). The MCC was higher for the Median method than the mean-based ones. Furthermore, using these measures of predictive capabilities/quality, the Median method performed better than simply identifying those who are currently poor as vulnerable to poverty.

References

Amemiya, T. (1977). "The maximum likelihood estimator and the non-linear three stage least squares estimator in the general nonlinear simultaneous equation model". Econometrica, 45. 955–968.

Chaudhuri, S. (2003). "Assessing vulnerability to poverty: concepts, empirical methods and illustrative examples". Department of Economics, Columbia University

Chaudhuri, S. and Christiaensen, L. (2002). "Assessing household vulnerability to poverty: illustrative examples and methodological issues". Presentation at the IFPRI-World Bank Conference on 'Risk and Vulnerability: Estimation and Policy Applications', September 23–24, 2002, Washington, DC.

Chaudhuri, S., Jalan, J. Suryahadi, A. (2002). "Assessing household vulnerability to poverty from cross-sectional data: a methodology and estimates from Indonesia". Department of Economics Discussion Paper Series 0102-52, Columbia University.

Chernozhukov, V., Fernández-Val, I. Galichon, A. (2010). "Quantile and probability curves without crossing". Econometrica, 78(3). 1093–1125.

Chesher, A. (2017). "Understanding the effect of measurement error on quantile regressions". Journal of Econometrics. 223–237.

Chiwaula, L., Witt, R. Waibel, H. (2011). "An asset-based approach to vulnerability: the case of small-scale fishing areas in Cameroon and Nigeria". Journal of Development Studies. 47(2). 338–353.

Christiaensen, L. Boisvert, R. (2000). "On measuring household food vulnerability: case evidence from Northern Mali". Department of Applied Economics and Management Working Papers WP-2000-05, Cornell University.

Christiaensen, L. Subbarao, K. (2005). "Towards an understanding of household vulnerability in rural Kenya". Journal of African Economies. 14(4). 520–558.

Dhaene, G., Jochmans, K. (2015). "Split-panel jackknife estimation of fixed-effect models".

Review of Economic Studies. 82, 991–1030.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y. Tu, X. M. (2014). "Log-transformation and its implications for data analysis". Shanghai Arch Psychiatry, 26(2). 105–109.

Gallardo, M. (2018). "Identifying Vulnerability to poetry: A critical survey". Journal of Economic Surveys, 32(2). 1074–1105.

Greene, W. H. (2002). "Econometric Analysis" (5th ed., pp 221). Upper Saddle River, NJ: Prentice Hall.

Gunther, I. Harttgen, K. (2009). "Estimating households' vulnerability to idiosyncratic and covariate shocks: a novel method applied in Madagascar". World Development, 37(7). 1222– 1234.

Gunther, I. Maier, J. (2014). "Poverty, vulnerability, and reference-dependent utility". Review of Income and Wealth, 60(1). 155–181.

Hardeweg, B., Wagener, A. Waibel, H. (2013). "A distributional approach to comparing vulnerability, applied to rural provinces in Thailand and Vietnam". Journal of Asian Economics, 25. 53–65

Hoddinott, J. Quisumbing, A. (2003). "Methods of microeconometric risk and vulnerability assessments". Social Protection Discussion Paper Series 0324, World Bank.

Hohberg, M., Landau, K., Kneib, T., Klasen, S. Zucchini, W. (2018). "Vulnerability to poverty revisited: Flexible modeling and better predictive performance". The Journal of Economic Inequality, 16(3). 439–454.

Koenker, R. Bassett Jr., G. (1978). "Regression Quantiles". Econometrica, 46(1). 33–50.

Koenker, R. Hallock, K. F. (2001). "Quantile Regression". Journal of Economic Perspectives, 15(4). 143–156.

Levine, S. (2012). "Exploring Differences in National and International Poverty Estimates: Is Uganda on Track to Halve Poverty by 2015?". Social Indicators Research, 107(12). 331-

349.

Lopez-Calva, L. Ortiz-Juarez, E. (2014). “A vulnerability approach to the definition of the middle class”. Journal of Economic Inequality, 12. 23-47.

Manning, W. G. (1998). “The logged dependent variable, heteroscedasticity, and the retransformation problem”. Journal of Health Economics, 17(3). 283–295.

Manning, W. G. Mullahy, J. (2001). “Estimating log models: to transform or not to transform”. Journal of Health Economics, 20(4). 461–494.

Matthews, B. W. (1975). ”Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. Biochimica et Biophysica Acta (BBA) - Protein Structure, 405(2): 442–451.

Pritchett, L., Suryahadi, A. Sumarto, S. (2000). “Quantifying vulnerability to poverty: a proposed measure with application to Indonesia”, Policy Research Working Paper 2437 (The World Bank, Washington DC).

Santos Silva, J. M. C. (2019, September). “Quantile regression: Basics and recent advances”. Paper presented at the 2019 UK Stata Conference, London, UK.

Santos Silva, J. M. C. Machado, J. A. F. (2019). “Quantiles via moments”. Journal of Econometrics. 145–173.

Suryahadi, A. Sumarto, S. (2003). “Poverty and vulnerability in Indonesia before and after the economic crisis”. Asian Economic Journal, 17(1). 45–64.

Vial, V. Hanoteau, J. (2015). “Returns to Micro-Entrepreneurship in an Emerging Economy: A Quantile Study of Entrepreneurial Indonesian Households’ Welfare”. World Development, 74. 142–157.

Ward, P. S. (2016). “Transient poverty, poverty dynamics, and vulnerability to poverty: An empirical analysis using a balanced panel from rural China”. World Development, 78. 541–553.

Woolridge, J. M. (2002). "Econometric Analysis of Cross Section and Panel Data". Upper Saddle River, NJ: Prentice Hall.

Zereyesus, Y. A., Embaye. W. T., Tsiboe, F. Amanor-Boadu, V. (2017). "Implications of Non-Farm Work to Vulnerability to Food Poverty-Recent Evidence From Northern Ghana". World Development, 91. 113–124.

Tables

Table 1
Summary Statistics of dependent variable and covariates

		Mean	Std. Dev
Dependent Variable	Adult equivalent consumption expenditure (in 2010 local currency unit)	819885.9	1008070
Household Characteristics	Formal Sector employment all members (% of HH members)	0.107	0.367
	Household head enterprise owner (1=yes)	0.304	0.460
	Household Size	6.394	3.336
	Area (1=urban, 0 rural)	0.225	0.418
	Sex of HH head (1=female)	0.304	0.460
	# rooms per adult equivalent	0.610	0.511
	Fertilizer use (1=yes)	0.119	0.323
	Pesticide use (=1 if household use pesticide)	0.087	0.281
	Dependency ratio	1.296	1.097
	% of persons aged ¿14 in HH who can read and write with understanding (literacy)	0.744	0.330
	# cattle owned	1.627	5.912
	Use electricity (1=yes)	0.121	0.326
	Age of HH head	45.949	15.262
	Highest education of HH head		
	Some primary (=1 if true)	0.509	0.500
	Some secondary (=1 if true)	0.158	0.365
	Post secondary (=1 if true)	0.073	0.260
	Number of Children (0-14 years old)	3.112	2.227
	Accessed credit (1=yes)	0.414	0.493
	Flush toilet access (1=yes)	0.018	0.132
	Latrine use as main toilet (1=yes)	0.763	0.426
	Motor Bike ownership (1=yes)	0.068	0.251
	Mobile phone ownership (1=yes)	0.559	0.497
	Television ownership (1=yes)	0.118	0.323
	Access to protected water sources (1=yes)	0.723	0.448
Community Characteristics	% of employed in comm. who own their own enterprise (avg % across all comm)	0.529	0.330
	Human capital of comm: comm wide avg of # of persons in HH completing		
	Some primary	1.185	0.523
	Some secondary	0.383	0.329
	Post secondary	0.141	0.225
	Regions (% of HH from these regions)		
	Kampala	0.068	-
	Central	0.242	-
	Eastern	0.238	-
	Northern	0.255	-
	Western	0.197	-
	Financial services in community (% of HH in comm with access)	0.053	0.224
	Education services in community (% of HH in comm with access)	0.600	0.490
	Health services in community (% of HH in comm with access)	0.303	0.460
	Market services in community (% of HH in comm with access)	0.344	0.475
	Road infrastructure services in community: Tarmac (% of HH in comm with access)	0.190	0.392
Shocks	Drought/Irregular Rains	0.269	0.444
	Floods	0.038	0.192
	Crop Pests & Disease or Livestock Disease	0.026	0.160
	Loss of Employment of HH member (except ill/inj)	0.004	0.064
	ill/inj Income Earning or ill/inj other HH member	0.108	0.310
	Death of Income Earner(s)	0.006	0.080
	Conflict/Violence	0.010	0.100
	Fire	0.008	0.091

Note: Author calculations from 2010 UNPS data. HH=Household, comm=community. The mean under each shock variable refers to the proportion of HH in each community reporting the shock.

Table 2
Quantile regression results for the conditional median adult equivalent expenditure
Dependent variable: Adult equivalent consumption expenditure

Covariate	Coefficient	Covariate	Coefficient	Covariate	Coefficient
HH head fomally employed	598319.4**	Credit from family or friends	69416.53***	Eastern Uganda	-522182.6***
	(289413.4)		(14740.35)		(85208.81)
HH head enterprise owner	65925.98***	Latrine for main toilet use	6585.332	Northern Uganda	-508248.2***
	(17444.88)		(19905.39)		(84769.12)
HH size	-18777.9***	Flush toilet for main toilet use	893888.2	Western Uganda	-562575.6***
	(5272.609)		(615736.4)		(83430.7)
Urban HH	21888.43	Bike ownership	230392.4***	Drought	-90588.31**
	(32000.42)		(38952.91)		(41576.64)
Female headed HH	-77790.68***	Mobile phone ownership	110969.1***	Floods	38344.34
	(14483.16)		(13356.67)		(34271.93)
# of rooms per adult equivalent person	260562.2***	Television ownership	147098***	Livestock and crop diseases	16926.81
	(42961.7)		(40480.46)		(28542.9)
Fertilizer use	54235.64***	Access to water potable water	38123.26***	Loss of emplyment	-165309.5**
	(18286.97)		(14185.83)		(78919.98)
Pesticide use	-156.8358	% enterprise ownership	71730.86***	ill/inj of HH members	62512.25***
	(26506.03)		(26780.92)		(17374.67)
Dependency ratio	-15655.68*	% primary educated	16498.06	Conflict/violence	-84793.68
	(8972.321)		(16291.21)		(78556.78)
Household literacy rate	-15843.98	% secondary educated	97548.17***	Fire	-62732.09
	(21547.7)		(35839.86)		(58855.69)
# of cattle owned	8401.348***	% tertiary educated	222693.4***	Death of income earner	60418.41
	(1640.029)		(78157.96)		(114154.8)
HH use Electricity	511068.1***	Educational facility in community	-15954.67	Drought*HH size	11847.46**
	(130224.3)		(15426.89)		(4640.026)
Age of HH head	-2027.1***	Health facility in community	-5194.49	Drought*dependency ratio	32886.77***
	(569.3593)		(17597.1)		(11576.35)
HH head primary education	-14074.5	Financial facility in community	-43836.65*	Central*drought	-150187.7***
	(19597.99)		(26508.71)		(47858.75)
HH head secondary education	22476.08	Market	-1555.281	Northern*drought	-89219.06***
	(32715.81)		(16926.95)		(33442.41)
HH head tertiary education	98992.41*	Tarmac	69754.83*	Constant	914411.9***
	(51866.73)		(36922.1)		(102101.4)
# of children 0-14	-12354.37	Central Uganda	-302921.1***		
	(8140.671)		(94152.73)		

Note: Author calculations from 2010 UNPS data. HH=Household, comm=community. *, **, and *** denote significance at the 0.10, 0.05, and 0.01 level. Robust standard errors are in brackets. The 3 columns of covariates and coefficients are from the same model, but the table is ordered in this way for brevity. All coefficients are in levels and in Ugandan Shillings (UGX); no variables are in logs or any other transformations.

Table 3
Proportion of housholds vulnerability to poverty in Uganda (%)

Empirical Method	Prop. Vulnerable (%)	Standard error	CI low	CI high
ZETA	46.7	0.045	37.5	55.5
GH	39.2	0.012	36.7	41.6
Median	40.4	0.021	36.3	44.4

Note: Author calculations from UNPS data. Setandard error computed via bootstrap set to 1000 replications.

Table 4
Cross-validation of future poverty status amongst those currently vulneble or poor in year t (%)

Empirical Method	vul/poor in t and poor in t+1	vul in t and poor in t or t+1	vul/poor in t and poor in t+1 or t+2	Vul in t and poor in t, t+1 or t+2	vul/poor in t and poor in t+1, t+2 or t+3	Vul in t and poor in t, t+1, t+2 or t+3
ZETA	42.3	58.0	65.2	71.9	68.0	73.4
GH	49.4	65.0	71.7	77.2	75.2	80.0
Median	69.5	86.3	86.3	91.1	88.4	92.1
Poor in year t	66.2	-	87.9	-	89.4	-

Note: Author calculations from 2010-2016 UNPS data. The training dataset corresponds to year t, and the 3 chronologically ordered test waves are year t+1 through t+3. Numbers indicate the propotion of those classified as vulnerable that actually became poor in the wave captioned in each column heading. For example, "vul in t and poor in t+1" indicates the proportion of those who were vulnerable in year t according to each empirical methodology (or, in the last row, the proportion of those poor in year t) that actually became poor in wave t+1 of the UNPS.

Table 5
Cross-validation amongst the year t non-poor (as a proportion of all vulnerable)

Empirical Method	vul in t and poor in t+1	vul in t and poor in t+1 or t+2	vul in t and poor in t+1, t+2 or t+3
ZETA	27.3	47.0	51.9
GH	28.7	49.0	57.9
Median	42.7	68.6	74.1

Note: Author calculations from 2010-2016 UNPS data. Each number in this table is a percentage. Numbers indicate what proportion of individuals currently classified (in year t) as vulnerable but non-poor (under each empirical method) actually became poor in any of the test waves.

Table 6A
Measures of the predictive capability of each empirical method

Empirical Method	Inclusion error (%)	Exclusion error (%)	Accuracy (%)	Matthews Correlation Coefficient
ZETA	50.5	49.4	50.0	0.00116
GH	31.8	50.9	60.6	0.17447
Median	24.1	33.2	72.3	0.42527
Poor in year t	26.5	33.6	70.7	0.39520

Note: Author calculations from 2010-2016 UNPS data. For this table, a person is considered *truly* vulnerable if they become poor in at least two of the next three waves of the UNPS. Inclusion error is the proportion of households that were not *truly* vulnerable but incorrectly identified as such after year t+3. Exclusion error is the proportion of households that were *truly* vulnerable but not identified as such after year t+3. Accuracy is the proportion of households that are correctly classified as *truly* vulnerable + truly non-vulnerable. Matthews Correlation Coefficient gives an estimate of the predictive accuracy of a (usually) bivariate classifier/measure. The correlation coefficient ranges from -1 (worst measure) to 1 (best measure).

Table 6B
Measures of the predictive capability of each empirical method

Empirical Method	Inclusion error (%)	Exclusion error (%)	Accuracy (%)	Matthews Correlation Coefficient
ZETA	50.5	49.5	50.1	0.00029
GH	29.8	57.1	51.7	0.12541
Median	14.8	46.5	63.7	0.36785
Poor in year t	13.7	43.5	66.3	0.40720

Note: Author calculations from 2010-2016 UNPS data. For this table, a person is considered *truly* vulnerable if they become poor in any of the three waves of the UNPS. Inclusion error is the proportion of households that were not *truly* vulnerable but incorrectly identified as such after year t+3. Exclusion error is the proportion of households that were *truly* vulnerable but not identified as such after year t+3. Accuracy is the proportion of households that are correctly classified as truly vulnerable + truly non-vulnerable. Matthews Correlation Coefficient gives an estimate of the predictive accuracy of a (usually) bivariate classifier/measure. The correlation coefficient ranges from -1 (worst measure) to 1 (best measure).