

# Predicting Consumption Expenditure for the Analysis of Health Care Financing Equity in Low Income Countries: a Comparison of Approaches

Gemini Mtei · Josephine Borghi · Kara Hanson

Accepted: 16 October 2014 / Published online: 7 November 2014  
© Springer Science+Business Media Dordrecht 2014

**Abstract** The analysis of equity in the distribution of health care payments requires nationally representative income and expenditure surveys, containing information on health care payments and ability to pay. Such national household surveys in developing countries collect limited information on out-of-pocket payments for health care but comprehensive information on household consumption expenditure (a proxy of income). There are also limited nationally representative health surveys to conduct equity analyses requiring an administration of small health-specific surveys to collect detailed information on health care payments. However, collecting household expenditure is expensive and time . This study compares quantile regression to Ordinary Least Square in predicting consumption expenditure. Split sample method and cross validation tests are used to evaluate the prediction methodology. Unlike OLS, the quantile model does not distort the values of, the Gini index, the concentration index and the Kakwani index and is the preferred method for predicting consumption expenditure for financing incidence analysis.

**Keywords** Consumption expenditure · Predict · Equity · Financing incidence · Out-of-pocket · Quantile regression · Developing countries

## 1 Introduction

Financing incidence analysis is an important approach to assessing resource allocation equity in the health sector and assessing the impact on equity of financing reforms.

---

G. Mtei (✉)  
Ifakara Health Institute, Plot 463, Kiko Avenue, Mikocheni, Off Mwai Kibaki Road,  
P.O. Box 78373, Dar es Salaam, Tanzania  
e-mail: gmtei@ihi.or.tz

J. Borghi · K. Hanson  
London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK

Incidence analysis explores the distribution of health care financing burdens across socio-economic groups to find out whether payment arrangements are in accordance with the principal of fairness (Wagstaff et al. 1999; O'Donnell et al. 2008a). Out-of-pocket financing is considered to be the most inequitable health care financing source whereby the poorest individuals bear the highest burden (Wagstaff and Vandoorslaer 1992; Wagstaff et al. 1999). In some low income countries, out-of-pocket payments account for more than 70 % of total health care financing (Wagstaff and van Doorslaer 2003; O'Donnell et al. 2008b). It is estimated that about 150 million of the world's population incur catastrophic out-of-pocket expenditures and another 80 million fall into poverty because of out-of-pocket payments for health care (Xu et al. 2007). Frequent analyses need to be undertaken in order to evaluate whether health care financing reforms enhance equity and protect individuals against financial risks. Representative information on both health care payments and consumption expenditure is needed to analyze health care financing incidence (O'Donnell et al. 2008a). However, there are challenges in undertaking such analyses because national surveys such as Household Budget Surveys (HBS), which usually provide good measures of consumption, frequently contain incomplete data on health care expenditures. Collecting household consumption expenditure information is expensive and time consuming (Grosh and Glewwe 1998) and typically not feasible within small surveys designed to capture detailed information on health care payments. Where a measure of living standard is required, small surveys typically rely on wealth indices, which cannot be used to quantify the distribution of the burden of health care financing payments through the calculation of progressivity or redistributive indices (O'Donnell et al. 2008a). The analysis of the distribution of health care payments in this case is limited to the distribution of shares of health care payments across wealth groups. However, a number of empirical studies showed that it is possible to predict consumption expenditure by linking data from two surveys, for example, a national income and expenditure survey that has detailed consumption expenditure data and a survey of health expenditures that does not contain information on consumption expenditure, providing there are variables common to both surveys (Skinner 1987; Abeyasekera and Ward 2002; Blundell et al. 2006; Sumarto et al. 2007; Akazili et al. 2011; Nguyen et al. 2011). But, these previous studies have predicted consumption expenditure using the Ordinary Least Squares (OLS) regression model which imposes the constant variance assumption, and an assumption that the relationship between consumption expenditure and explanatory variables is the same irrespective of socio-economic status, an assumption that might not always be valid. Experience shows that household income and consumption expenditure are characterized by extreme outlying values and a skewed or tailed distribution (Schluter and Trede 2006). Removing extreme variables when predicting consumption would risk altering the degree of inequity embedded in its distribution. At the same time, using Ordinary Least Squares (OLS) regression on a skewed distribution like consumption runs the risk of over predicting consumption for households with extremely low consumption while under predicting it for those with extremely high consumption levels, resulting in biased coefficients and intercept (Koenker and Hallock 2001; Cameron and Trivedi 2005). According to Cameron and Trivedi (2005), distributional issues such as behavior of the lower tail of a distribution are well dealt with by quantile regression. This paper compares the performance of the traditional OLS model with the quantile regression model (Koenker and Bassett 1978; Koenker and Hallock 2001) in the prediction of consumption expenditure for the analysis of the incidence of out-of-pocket payments.

## 2 Methods

### 2.1 Data

Tanzania Household Budget Survey (HBS) 2007 data were used to predict consumption expenditure into a survey administered by the Strategies for Health Insurance for Equities in Less Developed Countries (SHIELD) project in 2008. The HBS, which is nationally representative, collected information from a total of 10,466 households drawn from 447 clusters between January and December 2007. Information collected includes individual and household characteristics, economic activities, health status, household consumption and income, ownership of assets, housing characteristics, household access to services and facilities and food security. The SHIELD survey was conducted in June 2008 in six districts (four rural and two urban) and collected information on household and individual characteristics, illness incidence and utilization of health services, out-of-pocket payments, health insurance membership and willingness to pay for insurance, economic activities, ownership of assets and housing characteristics from a sample of 2,234 households. The SHIELD survey was weighted for national representation (see Appendix 1). The SHIELD survey was deemed to be more appropriate for analyzing equity in out-of-pocket payments because it collected information on health service utilization rates and transport costs and insurance coverage which were not collected in the HBS data. A number of variables including asset ownership, housing, source of utilities and demographic characteristics had similar distributions in the two datasets (see Appendix Table 1). The major difference between the two surveys was that the SHIELD survey did not collect information on household consumption expenditure to facilitate the quantification of health care financing incidence, despite having detailed information on health care payments; the HBS contained detailed information on household consumption expenditure but limited data on out-of-pocket expenditure.

### 2.2 Consumption Expenditure Prediction Methods

#### 2.2.1 Variables

We compared the use of quantile regression (Koenker and Bassett 1978; Koenker and Hallock 2001) to OLS to predict consumption expenditure in the SHIELD survey using the same set of explanatory variables in both models. A model of annual adult equivalent consumption was estimated from the sum of household expenditures on food, non-food and durable items. The explanatory variables included: log of the wealth index constructed from a number of durable assets and housing characteristics<sup>1</sup> estimated using the polychoric principal components analysis (PCA) approach proposed by Kolenikov and Angeles (2004)<sup>2</sup>; household size, marital status, gender, age, education and employment of the household head; and locality (whether the household lives in an urban or rural area). A constant positive term was added to the wealth index to eliminate negative values hence

---

<sup>1</sup> The index was comprised of radio, TV, tables, watches, iron, vehicle, flow material, roof material, source of cooking fuel, and source of light fuel. These were the variables with high positive significant correlation with consumption expenditure.

<sup>2</sup> Polychoric PCA was used in this case because the index comprised a number of categorical variables which hinders the application of normal PCA proposed by Filmer and Pritchett (2001)

facilitating its conversion to logarithm (Bollen et al. 2002). The log of adult equivalent consumption was used as the dependent variable.

### 2.2.2 Regression Models

The OLS is the most commonly used regression model to predict consumption expenditure in previous studies (Skinner 1987; Abeyasekera and Ward 2002; Blundell et al. 2006; Sumarto et al. 2007). However as explained before, this model has limitations in predicting skewed distributions, such as consumption expenditure or income. In this study we compare prediction using OLS regression against quantile regression for predicting consumption expenditure.

The main difference between OLS and quantile regression models is in the specification of the loss function (Cameron and Trivedi 2005). The Ordinary Least Squares model minimizes the sum of squared errors in Eq. 1 below,

OLS regression equation

$$\sum e^2 = (y - X\beta)'(y - X\beta) \tag{1}$$

where  $y$  = log of adult equivalent consumption expenditure = dependent variable,  $X$  = Explanatory variables,  $\beta$  = regression coefficients,  $e$  = residuals.

The quantile regression minimizes the sum of absolute deviation from quantile  $q$  in Eq. 2 below,

Quantile regression equation

$$Q_N(\beta_q) = \sum_{i:y_i \geq x'_i\beta} q|y_i - x'_i\beta| + \sum_{i:y_i < x'_i\beta} (1 - q)|y_i - x'_i\beta| \tag{2}$$

where  $q$  = quantile,  $y_i$  = log of adult equivalent consumption expenditure = dependent variable,  $x_i$  = vector of explanatory variables,  $\beta_q$  = coefficient estimate (slope) at a respective quantile  $q$ ,  $Q_N$  = Quantile function.

A non-parametric simultaneous quantile regression model (Gould 1992; Cameron and Trivedi 2005) was used to estimate the household consumption expenditure model (Eq. 3) at the 20th percentile (lower tail), the median (50th) and 80th percentile (upper tail) of the consumption expenditure distribution using 400 iterations.

Quantile consumption expenditure estimation model

$$y_{qhs} = \alpha_q + \sum_{i=1}^N \beta_q X_{ihbs} + \varepsilon; i = 1, 2, \dots, n; q = 0.20, 0.50, 0.8 \tag{3}$$

where  $X_{ihbs}$  is a vector of explanatory variables (wealth index and household demographic characteristics) in the HBS data,  $y_{qhs}$  = log of adult equivalent household consumption in the HBS data at quantile  $q$ ,  $\alpha_q$  and  $\beta_q$  are regression coefficients at quantile  $q$ ,  $q$  = household quantile, For the OLS model, the follow household consumption expenditure model was estimated (Eq.4),

OLS consumption expenditure estimation model

$$y_{hs} = \alpha + \sum_{i=1}^N \beta X_{ihbs} + \varepsilon; i = 1, 2, \dots, n \tag{4}$$

where  $X_{ihbs}$  is a vector of explanatory variables in the HBS data (wealth index and household demographic characteristics),  $y_{hbs} = \log$  of adult equivalent consumption expenditure in the HBS data,  $\alpha$  and  $\beta$  are regression coefficient estimates.

### 2.2.3 Analysis Methods

STATA 11 (StataCorp. 2009) was used in the estimation of both models. The test for multicollinearity was conducted using the variance inflation factor (VIF) method while heteroscedasticity was explored using Breusch–Pagan and Cook–Weisberg tests. We adjusted for clustering during the model estimation. It was not possible to use survey weights when estimating the quantile regression. Therefore, for comparison purposes, consumption estimates for both the quantile and the OLS models were derived without weights. Household survey weights were used in all other analyses including the calculation of the means, poverty incidence, and financing incidence. The Wald test was used to test for the significance of the differences (independence) between the coefficients estimated across quantiles, in the quantile model. It is important to observe that there is a potential for endogeneity between the consumption and employment variables which might call for the use of Instrumental Variable (IV) or Two Stage Least Squares (2SLS) models. However the use of either IV or 2SLS will not address the limitations embedded in the OLS model when predicting consumption expenditure. Since this is a comparative study that compares the commonly used OLS methodology against quantile regression, we don't expect endogeneity to affect the comparison because its effect will appear in both models. Also we wanted to maintain similar variables used in previous studies in order to be able to compare the findings. Previous studies, e.g. Sumarto et al. (2007) included employment variables in their consumption prediction models.

## 2.3 Prediction of Consumption Expenditure Within the SHIELD Survey

Both the OLS and quantile consumption expenditure estimates were used to predict consumption expenditure into the SHIELD survey and results were compared. Quantile regression coefficient estimates from Eq. 3 were applied to respective covariates in Eq. 5, to estimate consumption expenditure for households specified by the wealth index to be located in quantiles 20, 50 and 80, respectively, in the SHIELD dataset. We impose the assumption that both the HBS and the SHIELD survey come from a similar population and this assumption is established based on the pattern of mean distribution of a number of variables from both surveys. There is similarity in both surveys in the mean values for most of the variables as shown in Appendix Table 1. Further the time lapse between the two surveys is small hence we don't expect huge variations in most of the variables used in this prediction.

Quantile consumption expenditure prediction model

$$y_{qshield} = \alpha_{qhbs} + \sum_{i=1}^N \beta_{iqhbs} X_{ishield}; \quad i = 1, 2, \dots, n; \quad q = 0.20, 0.50, 0.80 \quad (5)$$

where  $X_{ishield}$  is a vector of explanatory variables in the SHIELD data (wealth index and household demographic characteristics),  $y_{qshield} = \text{predicted log adult equivalent consumption expenditure in the SHIELD data at quantile } q$ ,  $\alpha_{qhbs}$  &  $\beta_{iqhbs}$  = quantile coefficient estimates from Eq. 3 using HBS data.

For the OLS model, coefficient estimates from Eq. 4 were applied to respective covariates in Eq. 6 to predict consumption in the SHIELD data.

OLS consumption expenditure prediction model

$$\widehat{y}_{shield} = \widehat{\alpha}_{hbs} + \sum_{i=1}^N \widehat{\beta}_{ihbs} X_{ishield}; \quad i = 1, 2, \dots, n \quad (6)$$

where  $X_{ishield}$  is a vector of explanatory variables in the SHIELD data (wealth index and household demographic characteristics) is a vector of explanatory variables (wealth index and other household demographic characteristics),  $\widehat{y}_{shield}$  = predicted log adult equivalent consumption expenditure in the SHIELD data,  $\widehat{\alpha}_{ihbs}$  &  $\widehat{\beta}_{ihbs}$  = OLS coefficients estimated from Eq. 4 using HBS data.

Predicted consumption expenditure in the SHIELD data was inflated to 2008 prices using the annual inflation rate of 7 % (Bank of Tanzania 2008). The decision to estimate quantile regression consumption expenditure model at quantiles 20, 50 and 80 and use the estimates to predict consumption expenditure in the respective quantiles in the SHIELD data was based on a comparison of quantile plots and the slopes of consumption expenditure estimates when using OLS and quantile regression (see Fig. 1). The variation between slopes estimated using OLS and quantile regression was wider at the tails of the consumption expenditure distribution for most of the included explanatory variables (i.e. quantile 20 and quantile 80) while there was limited difference at the middle of the distribution (see Fig. 1). Coefficient estimates at quantile 20 and quantile 80 were therefore used to predict consumption expenditure at the lower tail (quantile 20 and below) and the upper tail (quantile 80 and above) of the wealth index distribution in the SHIELD survey, and the median quantile (quantile 50) was used to predict consumption expenditure in the middle quantiles. The major assumption imposed during prediction was that if both the HBS and SHIELD surveys had information on household consumption expenditure, the conditional mean distribution across consumption quantiles in the HBS would be similar to the conditional mean across consumption quantiles in the SHIELD survey because of similarities in the distribution of several variables in the two surveys (see Appendix Table 1). However, because there was no consumption information in the SHIELD survey, a wealth index was used to proxy consumption expenditure. It was then assumed that the conditional mean distribution across consumption expenditure quantiles in the HBS is similar to the conditional mean distribution across wealth index quantiles in the SHIELD survey. In this case, the effect of, for example, variations in household size on the variations in household consumption located at 20th quantile of the consumption quantile was assumed to be similar to the effect of household size on the variations of the wealth index (as a proxy of consumption) for households located in the 20th quantile of the wealth index. In this case, the predicted consumption for households located at the 20th quantile of consumption distribution would be similar to predicted consumption for households located at the 20th quantile of the wealth index. Validation of results in Sect. 3.2 proves this assumption to be true. Previous studies have also proposed that a wealth index may be used as a proxy measure of consumption expenditure (Sahn and Younger 2000; Filmer and Pritchett 2001; Moser and Felton 2007).

## 2.4 Reliability and Criterion Validity of Predicted Consumption Expenditure

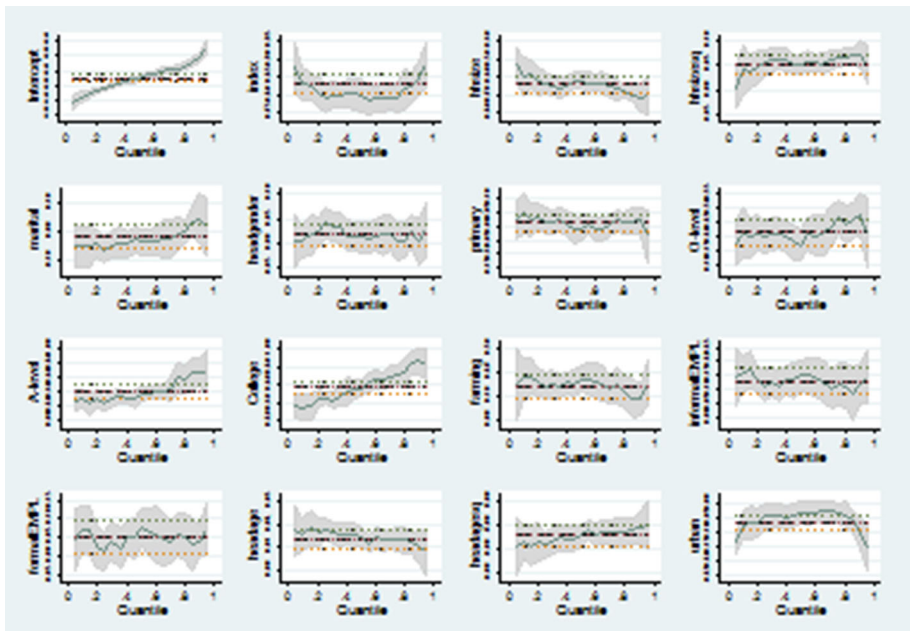
The split sample approach (Carmine and Zeller 1979) was used to test the reliability of predicted consumption expenditure. The HBS was divided into two random samples. The first sample was used to estimate the consumption expenditure model and predict into the

second sample, and vice versa. The consistency of predicted consumption expenditure was then examined as a measure of reliability.

Predicted consumption expenditure was tested for criterion validity (Carmine and Zeller 1979) by comparing the Gini inequality index of actual and predicted consumption expenditure across both sample splits. The proportion of the population identified as poor using predicted compared to actual consumption expenditure across both samples was another criterion used to examine validity. The two were used as measures of external validity. A household was considered poor if per day consumption expenditure was less than 1.25 USD per capita (Ravallion et al. 2008).

### 2.5 Analysis of the Incidence of Out-of-Pocket Payments

Analysis of the incidence of total out-of-pocket health care payments in the previous year was first performed. A disaggregated analysis was then conducted to explore variations in incidence by service components (payments for drugs, transport, consultation fee, registration fee and laboratory fees). Progressivity was analyzed using the Kakwani



**Fig. 1** Quantile plots of the comparison of the magnitude of consumption model coefficient estimates. Other include consultation fee, registration fee and diagnosis fee. The *solid dashed line* in the middle of each small graph in this figure gives the coefficient estimate when using OLS which is the mean effect on consumption of a unit change in explanatory variable. The *dotted lines* above and below the mean value give the confidence intervals of the estimates. The *green solid lines* give the magnitude of effect of the explanatory variables across household quantiles with the *grey shading* surrounding this line showing the confidence intervals. Index, wealth index; hhsize, household size; hhsize<sup>2</sup>, household size square; marital, marital status of the head; headgender, gender of the head; primary, primary education; Olevel, ordinary level education; Alevel, advance level education; college, college education; farming, working in farming activities; informalEMPL, working in informal employment; formalEMPL, working in formal employment; headage, age of the household head; headagesq, household head square; urban, living in urban locality



progressivity index (Kakwani 1977). Graphs of the distribution of the consumption share of out of pocket payments across wealth groups were also constructed.

### 3 Results

#### 3.1 Consumption Expenditure Estimates

The quantile regression results showed variation in the magnitude of the effect of the explanatory variables across household consumption expenditure quantiles (Table 1; Fig. 1). For example, a unit increase in household size reduced consumption expenditure by 62 % in the 20th quantile and the median quantile (q50), while the same reduced consumption by 68 % in the 80th quantile (Table 1). The difference was only significant between estimates at q50 and q80. Using OLS, a unit increase in household size reduced consumption expenditure by 64 %. Similarly the effect of the wealth index on household consumption expenditure varied across wealth groups (see Fig. 1). Significant differences between OLS estimates and quantile regression estimates were observed in the magnitude of effect of advanced level and college level education, farming, informal employment, urban–rural location, and household size, especially near the tails of the distribution (see Fig. 1).

There was similarity between the OLS estimates compared to quantile estimates for the effect of marital status of the household head, gender of the household head, and primary education. The coefficients from the quantile regression for these variables in all three quantiles were within the OLS confidence interval (see Fig. 1).

Predicted total adult equivalent consumption expenditure in the SHIELD survey using quantile regression was 2.6 billion Tanzania shillings (Table 2a) equivalent to 80 % of the total actual consumption expenditure estimated in the HBS. Total predicted consumption expenditure when using OLS was 73 % of actual consumption expenditure. The predicted mean consumption expenditure for the poorest 20 % in the SHIELD survey using quantile regression was 4 % larger than actual consumption expenditure in the HBS while the OLS method over-predicted this by 35 %. Mean predicted consumption for households in the least poor quintile in the SHIELD survey was 95 % of the actual consumption expenditure in the HBS when using quantile regression, while predicted consumption expenditure using OLS was 74 % of actual consumption expenditure (Table 2a).

The poorest 20 % of the population accounted for about 5.9 % of total actual consumption expenditure in the HBS while they accounted for approximately 6 and 8.6 % of total consumption expenditure as predicted by the quantile and OLS models, respectively, in the SHIELD survey (Table 2b).

#### 3.2 Reliability and Validity of Predicted Consumption

Predicted consumption expenditure was consistent in terms of the mean distribution across household quintiles and total consumption expenditure in both sample splits (Table 2a). The degree of inequality of predicted consumption expenditure as measured by the Gini index (Table 2b) was also consistent across both sample splits. Comparison between quantile predicted consumption expenditure and actual consumption expenditure also showed similarity in their level of inequality across both samples as measured by the Gini index (Table 2b) while prediction using the OLS model reduced the level of inequality in the distribution of consumption expenditure by about 23 %.



**Table 1** Consumption model estimation results for OLS and quantile regression

Variables	OLS	Quantile regression		
		q20	q50	Q80
Log wealth index	0.830*** (0.0262)	0.808*** (0.0230)	0.797*** (0.0183)	0.815*** (0.0237)
Log household size	-0.642*** (0.0289)	-0.620*** (0.0352)	-0.621*** (0.0286)	-0.679*** (0.0346)
Log household size squared	0.0530*** (0.0118)	0.0482*** (0.0137)	0.0535*** (0.0114)	0.0664*** (0.0141)
Marital status of household head	-0.0163 (0.0216)	-0.0386 (0.0256)	-0.0391* (0.0221)	0.00128 (0.0269)
Gender of household head	0.0168 (0.0134)	0.0246 (0.0156)	0.0126 (0.0143)	0.00828 (0.0160)
<i>Education</i>				
Primary education	0.0103 (0.0200)	0.0338 (0.0245)	-0.0130 (0.0195)	0.0115 (0.0224)
Secondary O-level	0.217*** (0.0267)	0.211*** (0.0353)	0.172*** (0.0255)	0.249*** (0.0308)
Secondary A-level	0.404*** (0.0655)	0.305*** (0.0793)	0.368*** (0.0696)	0.575*** (0.0944)
University/college	0.578*** (0.0624)	0.455*** (0.0627)	0.570*** (0.0512)	0.743*** (0.0622)
<i>Head employment</i>				
Farmer	0.144*** (0.0278)	0.160*** (0.0403)	0.171*** (0.0266)	0.126*** (0.0366)
Informal employment	0.129*** (0.0263)	0.138*** (0.0398)	0.147*** (0.0248)	0.114*** (0.0351)
Formal employment	0.0519* (0.0280)	0.0257 (0.0397)	0.0767*** (0.0241)	0.0554 (0.0352)
Age of household head	0.0140*** (0.00231)	0.0168*** (0.00372)	0.0141*** (0.00292)	0.0134*** (0.00307)
Age of household head squared	-0.0001*** (0.0001)	-0.0002*** (0.0001)	-0.0001*** (0.0001)	-0.0001*** (0.0001)
Locality	-0.0644* (0.0342)	-0.0651*** (0.0205)	-0.0258* (0.0145)	-0.0370 (0.0229)
Constant	12.26*** (0.0650)	11.82*** (0.0833)	12.26*** (0.0671)	12.66*** (0.0721)
R-squared/pseudo R-squared	0.613	0.361	0.391	0.401
Observations	10,102	10,102	10,102	10,102

The coding of categorical variables was as follows, head marital status (0 = never married 1 = ever married), head gender (0 = female 1 = male), education (1 = no education 2 = primary and adult education 3 = secondary-ordinary level 4 = secondary advanced level 5 = college or university), head employment (1 = not working 2 = farming 3 = self employed 4 = formal employment), location (0 = rural 1 = urban). Standard errors in brackets

\* Significant at 10 %, \*\* significant at 5 %, \*\*\* significant at 1 %

**Table 2** Comparison between predicted and actual consumption expenditure using quantile and OLS regressions

Quantile	HBS		SHIELD		Sample 1		Sample 2	
	Actual CE	Predicted CE (quantile)	Predicted CE (OLS)	Actual CE	Predicted CE (quantile)	Predicted CE (OLS)	Actual CE	Predicted CE (quantile)
Table 2a. Mean annual household adult equivalent consumption across quintiles								
Poorest 20 %	121,910	127,239	164,624	123,322	104,387	149,095	120,582	102,311
2nd Poorest 20 %	204,694	234,882	246,746	205,995	176,335	216,258	203,692	179,024
Middle	289,389	326,361	315,913	291,985	264,756	285,257	286,933	271,146
2nd least poorest 20 %	426,940	440,509	420,451	431,924	379,072	403,100	422,006	380,958
Least poorest 20 %	1,037,135	981,789	769,782	1,029,200	954,280	792,210	1,045,867	932,116
Total CE (bill Tsh.)	3,260	2,620	2,390	1,630	1,470	1,440	1,460	1,460
Table 2b. Mean annual household adult equivalent consumption as a percentage of total annual consumption across quintiles								
Poorest 20 %	5.86	6.04	8.61	5.92	5.56	8.14	5.81	5.50
2nd Poorest 20 %	9.84	11.12	12.91	9.91	9.4	11.64	9.81	9.58
Middle	13.91	15.56	16.43	14.01	14.1	15.48	13.78	14.56
2nd least poorest 20 %	20.53	20.75	21.89	20.76	20.15	21.88	20.29	20.39
Least poorest 20 %	49.86	46.53	40.15	49.4	50.79	42.87	50.31	49.97
Gini index	0.43	0.39	0.31	0.43	0.44	0.34	0.44	0.43
Percentage poor jointly poor in both actual CE and	36.33	36.32	31.57	37.36	41.38	35.23	38.61	39.95
jointly not poor in both actual CE and					71.17	64.98		66.90
CE consumption expenditure					74.23	80.35		77.00
								81.03

Predicted consumption expenditure using quantile regression classified 41 % of the population as poor individuals in sample 1 while actual consumption expenditure gave an estimate of 37 % (Table 2b). The OLS prediction gave an estimate of 35 %. A similar pattern was also observed for predictions in sample 2 (Table 2b).

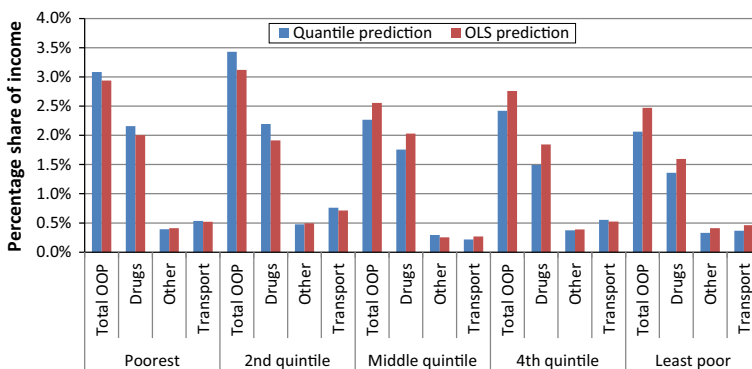
A relatively higher proportion (71 and 66 % in samples 1 and 2, respectively) of individuals were jointly classified as poor by both actual and quantile predicted consumption expenditure compared to OLS prediction (65 and 62 % in samples 1 and 2, respectively).

### 3.3 Incidence of Out-of-Pocket Payments

Using quantile predicted consumption expenditure in the SHIELD survey, the analysis of progressivity of total out-of-pocket healthcare payments showed that the poorest pay a higher proportion of their income out-of-pocket compared to the least poor (Fig. 2). A similar pattern of distribution was observed using the OLS predicted consumption expenditure. However, analysis using quantile predicted consumption expenditure shows that the poorest 20 % pay about 3 % of their income as total out-of-pocket payments, while for the least poor 20 % total out-of-pocket payments account for about 2 % of their income. Comparison using the OLS predicted consumption expenditure indicates that the poorest 20 % also pay about 3 % of their income total out-of-pocket expenditure while for the least poor 20 % out of pocket payments account for about 2.5 % of their income.

Disaggregated analysis shows similarities in the share of income spent on drugs, transport and other out-of-pocket payments between quantile and OLS predicted consumption expenditure among the poorest households, while the difference between the two prediction methods increases among the higher wealth groups, particularly for total expenditure and expenditure on drugs (Fig. 2).

Comparison with the Kakwani indices shows that total out-of-pocket payments together with the individual components were consistently regressive when using either quantile or OLS predicted consumption expenditure. However, quantile prediction gave a more regressive index of total out-of-pocket payments (Kakwani index  $-0.10$ ) compared to the OLS (Kakwani index  $-0.04$ ). A similar observation applies to the individual components of out-of-pocket payments (Table 3). Comparison of standard errors generated from the use of the two approaches indicates that Kakwani indices derived using OLS predicted



**Fig. 2** Comparison of the distribution of out-of-pocket payments across household between quantile and OLS prediction models

**Table 3** Comparison of Kakwani indices between quantile and OLS predicted consumption expenditure

	Total out-of-pocket	Drugs	Transport	Other
Consumption prediction using quantile regression				
Concentration index	0.290 (0.032)	0.281 (0.039)	0.304 (0.067)	0.316 (0.059)
Kakwani index	-0.103 (0.047)	-0.112 (0.052)	-0.089 (0.070)	-0.078 (0.105)
Consumption prediction using Ordinary Least Squares (OLS)				
Concentration index	0.267 (0.035)	0.262 (0.042)	0.266 (0.065)	0.291 (0.069)
Kakwani index	-0.044 (0.056)	-0.049 (0.065)	-0.045 (0.068)	-0.019 (0.107)

Standard errors in bracket; Other OOPs include payments for consultation fee, registration fee and diagnosis fee 5.3

consumption were all insignificantly regressive while quantile predicted consumption expenditures shows that out-of-pocket payments were significantly regressive (Table 3). In addition, the OLS model under-estimated the magnitude of the concentration indices of health care payments (Table 3).

#### 4 Discussion

The main objective of this paper was to propose a methodology for predicting consumption expenditure which would help to address the challenge of collecting household consumption data when using a small household survey to undertake equity analyses of health system financing or provision. Results showed that such prediction is indeed possible, confirming findings from Ghana (Akazili et al. 2011; Nguyen et al. 2011). The second objective was to examine the appropriateness/performance of quantile regression compared to OLS in the prediction of consumption expenditure. The study found that there were variations in the magnitude of the effect of explanatory variables across quantiles which justified the use of quantile regression. The quantile prediction model also performed better in the identification of poor households than the OLS prediction model, where 71 and 67 % were correctly identified using quantile regression in samples 1 and 2, compared to 65 and 62 % using OLS regression. Other studies have indicated even lower performance of the OLS in predicting the poor, with about 30–52 % of the population being correctly predicted in the study by Abyeekera and Ward (2002). Similarly, findings from a study by Sumarto et al. (2007) indicated that only 50 and 47 % of the poor were correctly predicted using the consumption correlate model in urban and rural populations respectively. In addition, the Gini index of predicted consumption expenditure using the quantile model was similar to the index of actual consumption expenditure whereas prediction using OLS model significantly under-estimated the degree of inequality embedded in the consumption distribution. This suggests that quantile prediction could overcome previous concerns that the Gini coefficient of predicted consumption will be under-estimated due to the shrinkage of predicted values towards its mean (Abeyasekera and Ward 2002; Sumarto et al. 2007; Matsaganis et al. 2008) as it reduces the degree of over prediction of consumption expenditure among the poor households and under prediction among the rich.

This study showed that the choice of prediction method did not alter the pattern of the incidence of health care payments. However, the choice did affect the magnitude of the measures of inequity and the significance of results. The use of the OLS model under-estimated the degree of inequality embedded in the distribution of out-of-pocket payments,

which could be a result of under-estimation of inequality in the distribution of consumption expenditure as shown by the Gini index above. The Kakwani index values generated using OLS predicted consumption were not significantly different from zero, indicating that out-of-pocket payments were not regressive (in other words, not significantly different from proportional). Results using quantile prediction model indicated that out-of-pocket payments were significantly regressive, findings which are consistent with previous analyses conducted in Tanzania using the 2000/01 Household budget survey (Mtei et al. 2012) and the comparative analysis conducted using actual HBS data for years 2000/01 and 2007 (Mtei 2012).

It is important to note that the Gini index plays a significant role when it comes to the analysis of inequities in the distribution of health care payments. Therefore, the consumption expenditure prediction approach needs to maintain the degree of inequality in the distribution of actual consumption. The use of the quantile regression model seems to better satisfy this requirement than the OLS model.

This study is the first attempt to conduct a comparative analysis of different approaches to predicting consumption and to explore the implications for the distribution of health care payments. It adds to a previous contribution by Filmer and Pritchett (2001) in proposing an alternative approach to constructing a measure of living standards or wealth in the absence of actual consumption expenditure or income.

Although the validation in this study has shown that the quantile prediction approach is a better option than the OLS, it is important that this proposed methodology be tested with other datasets and incidence analyses be applied to types of health care payments before it can be generalized in Tanzania and other countries. The present study has some limitations. First, the variables used in both prediction methodologies were limited to those available in the SHIELD survey, and were not an exhaustive list of all variables with positive correlation with consumption. Since the initial objective of this survey was not to collect variables that will be used for predicting consumption expenditure, information on some easy to collect variables that have higher correlation with consumption was not collected. For the purpose of improving accuracy of the prediction equation, variables such as the number of meals per day, number of times per week a particular food item (rice, meat, etc.) has been consumed, total spending on food per week, expenditure on transport and other highly frequently purchased items could be included in future analyses. In addition, the prediction methodology used an asset index to classify households into lower, middle and upper quartiles in the SHIELD survey, and used these classifications to link the SHIELD survey with the HBS in the prediction process. Previous studies have shown some concerns that asset or wealth indices are not good proxies of consumption (Howe et al. 2009), which might reduce the validity of their use in linking the quantiles in this study. However, the positive sign and the significance of the wealth index coefficient in the consumption model estimated in this study gives confidence that wealth indices and consumption expenditure have a positive relationship.

## 5 Conclusion

This study proposes quantile regression as a better option than OLS when predicting consumption for poverty classification and the analysis of the incidence of health care payments. The quantile model retains the degree of inequality embedded in actual consumption expenditure; hence it does not distort the value of the Gini index of consumption distribution, the concentration index of health care payments or the Kakwani indices. The Gini index calculated using actual consumption is 0.43 while predictions using quantile

regression and OLS gave the Gini index values of 0.39 and 0.31 respectively. Consequently, financing incidence results calculated using quantile predicted consumption expenditure were more reliable than those derived from predictions derived using the OLS. The out-of-pocket Kakwani index obtained using quantile regression predicted consumption is  $-0.10$  while the OLS predicted consumption gave the Kakwani index value of  $-0.04$ . The Kakwani value from the quantile regression is similar to the one calculated using actual consumption in the HBS.

The analysis in this paper has a limitation that variables used in the prediction model were confined to those collected in the SHIELD survey. Since the initial objective of this survey was not to collect data on variables that will be used for predicting consumption expenditure, information on some easy to collect variables that have higher correlation with consumption was not collected. Future researches need to test the proposed methodology with other datasets as a way of further validating the prediction model for its generalization. Where possible surveys should be used to collect information on easy to capture consumption expenditure items, such as the number of meals consumed per day, number of times per week a particular food item (rice, meat, etc.) has been consumed and use them as explanatory variables in the prediction models. These variables have higher correlation with total consumption expenditure and hence will help to improve the prediction power.

**Acknowledgments** The SHIELD survey was administered under the SHIELD project funded by International Development and Research Centre (Grant Number 103457) and the European Commission (Sixth Framework Programme; Specific Targeted Research Project No: 32289).

## Appendix 1: Calculation of Survey Weights in the SHIELD Survey

A two level weight was generated. The first weight ( $W_1$ ) aims to make the sample of insured and the uninsured representative of the population at the district level. Eq. 7 was therefore used to weight both insured and uninsured at the district level separately. The second weight ( $W_2$ ) aims to achieve national representation by making the insured and uninsured populations representative of the national population. The final weight ( $W_f$ ) was obtained by multiplying  $W_1$  and  $W_2$ . Population data for urban and rural areas were derived from the 2002 national population census. Information on membership of Community Health Fund and National Health Insurance Fund was secured from the Ministry of Health and Social Welfare and the National Health Insurance Fund office, respectively.

The weights were derived as follows;

$$W_1 = \frac{\sum_{i=1}^{n^d} d_{ij}}{\sum_i^{m^d} S_{dij}}, \quad m^d \in n^d \quad (7)$$

where  $d$  = district,  $i$  = individual in district  $d$ ,  $j$  = insurance status of an individual in district  $d$ ,  $sd$  = district sample.

The second weights was calculated as,

$$W_2 = \frac{\sum_{i=1}^{n^L} L_{ij}}{\sum_i^{m^d} d_{ij}}, \quad m \in n \quad (8)$$

where  $L$  = locality (urban or rural).

The final weight was calculated as,

$$W_f = W_1 * W_2$$

**Appendix Table 1** Mean distribution of variables used in the prediction process

Variables	Category code number (categorical variables)	Data source	
		HBS	SHIELD
<i>Variables used in wealth index construction</i>			
Assets			
Radio	1	0.66	0.67
TV		0.08	0.12
Tables	1	0.70	0.73
Watches	1	0.45	0.50
Iron	1	0.26	0.39
Vehicle	1	0.01	0.01
Floor material			
Earth	1	0.68	0.63
Concrete	2	0.32	0.37
Roof material			
Grass, leaves, bamboo	1	0.44	0.34
Metal sheets, asbestos	2	0.56	0.65
Concrete, tiles	3	0.00	0.01
Cooking fuel			
Firewood, farm residuals	1	0.74	0.69
Kerosine, coal, charcoal	2	0.26	0.30
Electricity, gas, solar	3	0.01	0.01
Light fuel			
Firewood, kerosene	1	0.88	0.86
Electricity, solar, gas	2	0.12	0.14
<i>Variables used in the consumption expenditure model estimation</i>			
Wealth index (first component)		-0.12	-0.01
Household size		4.77	5.34
Age of the head		44.45	44.50
Head marital status			
Never married <sup>#</sup>	0	0.07	0.10
Married/divorced/widow	1	0.93	0.90
Head gender			
Female <sup>#</sup>	0	0.24	0.22
Male	1	0.76	0.78
Education level			
No education <sup>#</sup>	1	0.24	0.16
Primary	2	0.67	0.72
Form four	3	0.07	0.09
Form six	4	0.01	0.01
University/college	5	0.02	0.02
Employment			
Not working <sup>#</sup>	1	0.06	0.06
Farming	2	0.62	0.74



**Appendix Table 1** continued

Variables	Category code number (categorical variables)	Data source	
		HBS	SHIELD
Informal employment	3	0.18	0.11
Formal employment	4	0.14	0.10
Locality			
Rural <sup>#</sup>	0	0.72	0.79
Urban	1	0.28	0.21
Observations		10,102	2,115

<sup>#</sup> Reference category in the consumption model estimation

## Appendix 2: Similarities and Differences Between the SHIELD and the HBS Data

The two surveys had similarities and differences in the distribution of a number of variables (Appendix Table 1). For example, the HBS 2007 estimated about 72 % of the Tanzanian population as living in a rural locality while this was approximated to be 75 % in the SHIELD survey. In addition, about 76 and 77 % of the household heads were male in the HBS and SHIELD surveys, respectively. The mean household size was 4.8 in the HBS and 5.1 in the SHIELD survey. There were also similarities in the ownership of assets like radio, tables, watches, and housing characteristics across the two surveys.

The two surveys collected information on payments for consultation fees, drugs, and diagnosis. While this information was collected separately for inpatient and outpatient care in the SHIELD survey, expenditure on drugs was collected as a separate category in the HBS and not linked to the type of care sought. Payments in the SHIELD survey were linked to the visit which a household member made in the past 1 year for inpatient care or past 1 month for outpatient care while this information was not linked to visits in the HBS. It is argued that having a more detailed question on the breakdown of health payments has a greater prompting effect, implying less memory loss (Heijink et al. 2011) a fact that may lead to more accurate reporting of out of pocket payments. In addition, the SHIELD survey collected information on transport costs associated with health care utilization while this was not collected in the HBS.

## References

- Abeyasekera, S., & Ward, P. (2002). *Models for predicting expenditure per adult equivalent for AMMP surveillance sentinel sites*. Adult Morbidity and Mortality Project, Ministry of Health, Tanzania: Dar es Salaam.
- Akazili, J., Gyapong, J., et al. (2011). Who pays for health care in Ghana? *International Journal for Equity in Health*, 10(26). doi:10.1186/1475-9276-10-26.
- Bank of Tanzania (2008). Annual Report 2007/08. Ministry of Finance & Economic Affairs. Dar es Salaam, Bank of Tanzania.
- Blundell, R., Pistaferri, L., et al. (2006). Imputing consumption in the PSID using food demand estimates from the CEX. Program of research, ESRC Centre for the Microeconomic Analysis of Public Policy, IFS.
- Bollen, K. A., Glanville, J. L., et al. (2002). Economic status proxies i studies of fertility in developing countries: Does the measure matter? *Population Studies*, 56, 81–96.
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics methods and applications*. Cambridge: Cambridge University Press.

- Carmine, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment California*. USA: Sage Publications Inc.
- Filmer, D., & Pritchett, L. (2001). Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of India. *Demography*, 38(1), 115–132.
- Gould, W. W. (1992). Quantile regression with bootstrapped standard errors. *Stata Technical Bulletin*, 9, 19–21.
- Grosh, M., & Glewwe, P. (1998). Data watch: The World Bank's living standards measurement study household surveys. *The Journal of Economic Perspective*, 12(1), 187–196.
- Heijink, R., Xu, K., et al. (2011). Validity and comparability of out of pocket health expenditure from household surveys: a review of the literature and current surveys instruments. WHO discussion paper no. 1-2011. Geneva, WHO.
- Howe, L. D., Hargreaves, J. R., et al. (2009). Is the wealth index a proxy for consumption expenditure? a systematic review. *J of Epidemiology and Community Health*, 63, 871–880.
- Kakwani, N. C. (1977). Measurement of tax progressivity: An international comparison. *The Economic Journal*, 87(345), 71–80.
- Koenker, R., & Bassett, G. J. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50.
- Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4), 143–156.
- Kolenikov, S., & G. Angeles (2004). The use of discrete data in PCA: Theory, simulations and applications to socioeconomic indices, Working paper no. wp-04-85 of MEASURE evaluation project, Carolina Population Centre, University of North Carolina. Download at <http://www.cpc.unc.edu/measure/publications/wp-04-85>.
- Matsaganis, M., T. Mitrakos., et al. (2008). Modelling household expenditure on health care in Greece. Working paper no. 68. Athens, Economic research department, Bank of Greece.
- Moser, C., & Felton, A. (2007). The construction of an asset index measuring asset accumulation in Ecuador CPRC Working paper 87, The Brookings Institution, USA.
- Mtei, G. J. (2012). Health care financing progressivity and household risk protection in the context of health system financing reforms in Tanzania. Global Health Department. London, London School of Hygiene and Tropical Medicine, University of London. Doctor of Philosophy: 392.
- Mtei, G., Makawia, S., et al. (2012). Who pays and who benefits from health care? An assessment of equity in health care financing and benefit distribution in Tanzania. *Health Policy and Planning*, 27(suppl 1), i23–i34.
- Nguyen, H. T., Rajkotia, Y., et al. (2011). The financial protection effect of Ghana National Health Insurance Scheme: evidence from a study in two rural districts. *International Journal for Equity in Health*, 10(4). doi:10.1186/1475-9276-10-4.
- O'Donnell, O., Van Doorslaer, E., et al. (2008a). Who pays for health care in Asia? *Journal of Health Economics*, 27(2), 460–475.
- O'Donnell, O., Van Doorslaer, E., et al. (2008b). *Analyzing health equity using household survey data: a guide to techniques and their implementation*. Washington: DC, The World Bank.
- Ravallion, M., Chen, S., et al. (2008). Dollar a day revisited. Policy Research Working Paper 4620, The World Bank.
- Sahn, D. E., & Younger, S. D. (2000). Expenditure Incidence in Africa: microeconomic evidence. *Fiscal Studies*, 21(3), 329–347.
- Schluter, C., & Trede, M. (2006). Identifying multiple outliers in heavy-tailed distributions with an application to market crashes. *Journal of Empirical Finance*, 15(4), 700–713.
- Skinner, J. (1987). A superior measure of consumption from the panel study of income dynamics. *Economic Letters*, 23(2), 213–216.
- StataCorp. (2009). Stata Statistical Software: Release 11. College station TX, StataCorp LP.
- Sumarto, S., Suryadarma, D., et al. (2007). Predicting consumption poverty using non-consumption indicators: experiments using Indonesia data. *Social Indicators Research*, 81, 543–578.
- Wagstaff, A., & van Doorslaer, E. (2003). Catastrophe and impoverishment in paying for health care: with applications to Vietnam 1993–1998. *Health Economics*, 12(11), 921–934.
- Wagstaff, A., van Doorslaer, E., et al. (1999). Equity in the finance of health care: some further international comparisons. *Journal of Health Economics*, 18(3), 263–290.
- Wagstaff, A., & Vandoorslaer, E. (1992). Equity in the finance of health care: Some international comparison. *Journal of Health Economics*, 11(4), 361–387.
- Xu, K., Evans, D. B., et al. (2007). Protecting households from catastrophic health spending. *Health Affairs*, 26(4), 972–983.