OPHI

OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE

www.ophi.org.uk

UNIVERSITY OF OXFORD

Summer School on Multidimensional Poverty Analysis

Oxford Poverty & Human Development Initiative, (OPHI), University of Oxford

3–15 July 2017

Marrakech, Morocco

Tabita, Kenya — Rabiya, India — Stéphanie, Madagascar — Agathe, Madagascar — Dalma, Kenya — Ann-Sophia, Kenya — Valérie, Madagascar

Regression Models for the AF Measures

Bouba Housseini

12 July 2017

Tabita, Kenya Rabiya, India Stéphanie, Madagascar Agathe, Madagascar Dalma, Kenya Ann-Sophie, Kenya Valérie, Madagascar

Where we are:

Post-estimation analyses of M_0 comprise:

- Decompositions into H and A, or by group or region
- Breakdown by dimension
- Robustness analysis to parameters selection in measurement design.
- Computation of standard errors for statistical inference - confidence intervals and hypothesis testing.
- Analysis of distributions and dynamics over time.

What are we missing?

Using data for Indonesia 1993, the following **characterisation** (**descriptive**) of multidimensional poverty (M0=0.133) (Ballon & Apablaza, 2013)

MD poor households
characteristics of the household head

Average			*Proportion*	
Years of education	**Age**	**Household size**	**Male head**	**Muslim**
2.1	25.5	5.1	80%	91%

we **still need to isolate** the « **effect** » (size) of each of these characteristics on overall poverty in a multivariate framework.

Why is this important?

From a policy perspective, in addition to measuring poverty we must perform **some vital** analyses regarding the transmission mechanisms between policies and poverty measures.

This is to assess how **poverty** is **explained** by non-$\mathbf{M_0}$ related variables

How can we account for this?

Through **regression analysis** we can **account** for the "effect/size" of micro and macro **determinants** of multidimensional poverty.

We can differentiate between:

- **'micro'** regressions: unit of analysis is the household or the person

- **'macro'** regressions: unit analysis is some "spatial" aggregate, such as a province, a district or a country.

Micro and Macro Regressions

What are some vital regression analysis we may wish to study with AF measures?

Micro regressions:

a) explore the determinants of poverty at the household/individual level

b) create poverty profiles;

Macro regressions:

a) explore the elasticity of poverty to economic growth and economic performance in general,

b) understand how macroeconomic variables (e.g. average income, public expenditure, decentralization, infrastructure density, information technology) relate to multidimensional poverty levels or changes across groups or regions—and across time.

Which are some 'focal' variables to regress?

Dependent variable AF measure: Y	Range of Y	Regression Model	Level	Conditional Distribution $p_Y(y)$
Binary $(c_i \geq k)$	0,1	Probability	Micro	Bernoulli
M_0, H	[0,1]	Proportion	Macro	Binomial

The classic regression model

$$y_i = \underbrace{E\left[Y_i \mid \mathbf{x}_i\right]}_{\text{deterministic component}} + \underbrace{\varepsilon_i}_{\text{error component}}$$

where: $E\left[Y_i \mid \mathbf{x}_i\right]$ denotes the conditional expectation of the random variable Y_i given $\mathbf{x}_i$, and ε_i is a disturbance or random error.

This model is a general representation of regression analysis. It attempts to explain the variation in the dependent variable through the conditional expectation **without imposing** any functional form on it.

The linear regression model

If we specify a *linear* functional form of the conditional expectation $E[Y_i \mid \mathbf{x}_i]$ denoted as $\mu_{Y_i|\mathbf{x}_i}$

$$\mu_{Y_i|\mathbf{x}_i} = E[Y_i \mid \mathbf{x}_i] = \eta_i = \beta_0 + \sum_j \beta_j x_{ij}$$

we obtain the classic linear regression model (LRM)

$$y_i = \eta_i + \varepsilon_i.$$

η_i is referred to as the predictor in the generalized linear model.

Generalised Linear Modelling

The GLM family of models involves **predicting** a *function* (g) of the conditional mean of a dependent variable as a *linear combination* of a set of explanatory variables η_i (**the linear predictor**). This function is referred to as the **link function** .

A GLM takes the form:

$$g\left(\mu_{Y_i|\mathbf{x}_i}\right) = \eta_i = \beta_0 + \sum_j \beta_j \mathbf{x}_{ij}$$

Classic linear regression is a specific case of a GLM in which the conditional expectation of the dependent variable is modelled by the identity *function*.

Generalized Linear Regression Models with AF Measures

Dependent variable AF measure: Y	Range of Y	Regression Model	Level	Conditional Distribution $p_Y(y)$	Link $g(\mu_i) = \eta_i$		Mean function $\mu_i = G(\eta_i)$
Binary $(c_i \geq k)$	0,1	Probability	Micro	Bernoulli	Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\Lambda(\eta_i)$
M_0, H	[0,1]	Proportion	Macro	Binomial	Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$

Note: $\Phi(\cdot)$ and $\Lambda(\cdot)$ are the cumulative distribution functions of the standard-normal and logistic distributions, respectively. For the binary model, the conditional mean μ_i is the conditional probability π_i.

A binary model in the GLM framework

$$Y_i = \begin{cases} 1 & \text{if and only if } c_i \geq k \\ 0 & \text{otherwise} \end{cases}$$

The outcomes of this binary variable occur with probability $\boldsymbol{\pi}_{\mathrm{i}}$ which is a **conditional probability** given the explanatory variables:

$$\pi_i \equiv \Pr\left(Y_i \mid \mathbf{x}_i\right) = \mu_{Y_i|\mathbf{x}_i}$$

For a **binary model** the **conditional distribution of the dependent variable**, or random component in a GLM, is given by a **Bernoulli distribution.**

A binary model in the GLM framework

To ensure that the **π_i** stays between 0 and 1, a GLM commonly considers two alternative link functions (g): **probit link -** quantile function of the standard normal distribution function, and the **logit link** – quantile of the logistic distribution function.

The logit model (log of the odds) of π gives the **relative chances** of being multidimensionally poor.

$$\log_e \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

The logit model

$$\log_e \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

The logit model is a linear, **additive** model for the log odds, equation , but it is also a **multiplicative** model for the odds:

$$\frac{\pi}{1-\pi} = e^{\beta_0}(e^{\beta_1})^{x_{1i}} \ldots (e^{\beta_k})^{x_{ki}}$$

Our interest lies on conditional mean π_i.

Interpretation of Model Parameters

The partial regression coefficients β_j are interpreted as **marginal changes** of the logit, or as **multiplicative** effects on the odds.

Thus β_j indicates the change in the logit due to a one-unit increase in x_j , and e^{β_j} is the **multiplicative effect** on the odds of increasing x_j by 1, while holding constant the other explanatory variables.

For this reason e^{β_j} is known as the **odds ratio** associated with a one-unit increase in x_j .

Example

Poverty profile for West Java, Indonesia in 1993
(Ballon & Apablaza, 2013)

They regress the **log of the odds of being multidimensionally poor** (with k=33%) on demographics, and socio-economic characteristics of the household head.

These have been selected on the grounds of '**restraining**' any 'possible' **endogeneity issue** that may arise in the construction of this poverty profile.

Logistic regression results – West Java, 1993

Variable	Parameter Estimate	Robust Std. Err.	*t* ratio	Significance level	Odds ratio
Years of education of household head	-0.68	0.03	-19.65	***	0.51
Female household head	0.24	0.09	2.71	***	1.28
Household size	0.09	0.01	7.02	***	1.10
Living in urban areas	-0.85	0.07	-11.40	***	0.43
Being Muslim	-0.02	0.32	-0.07	n.s.	0.98

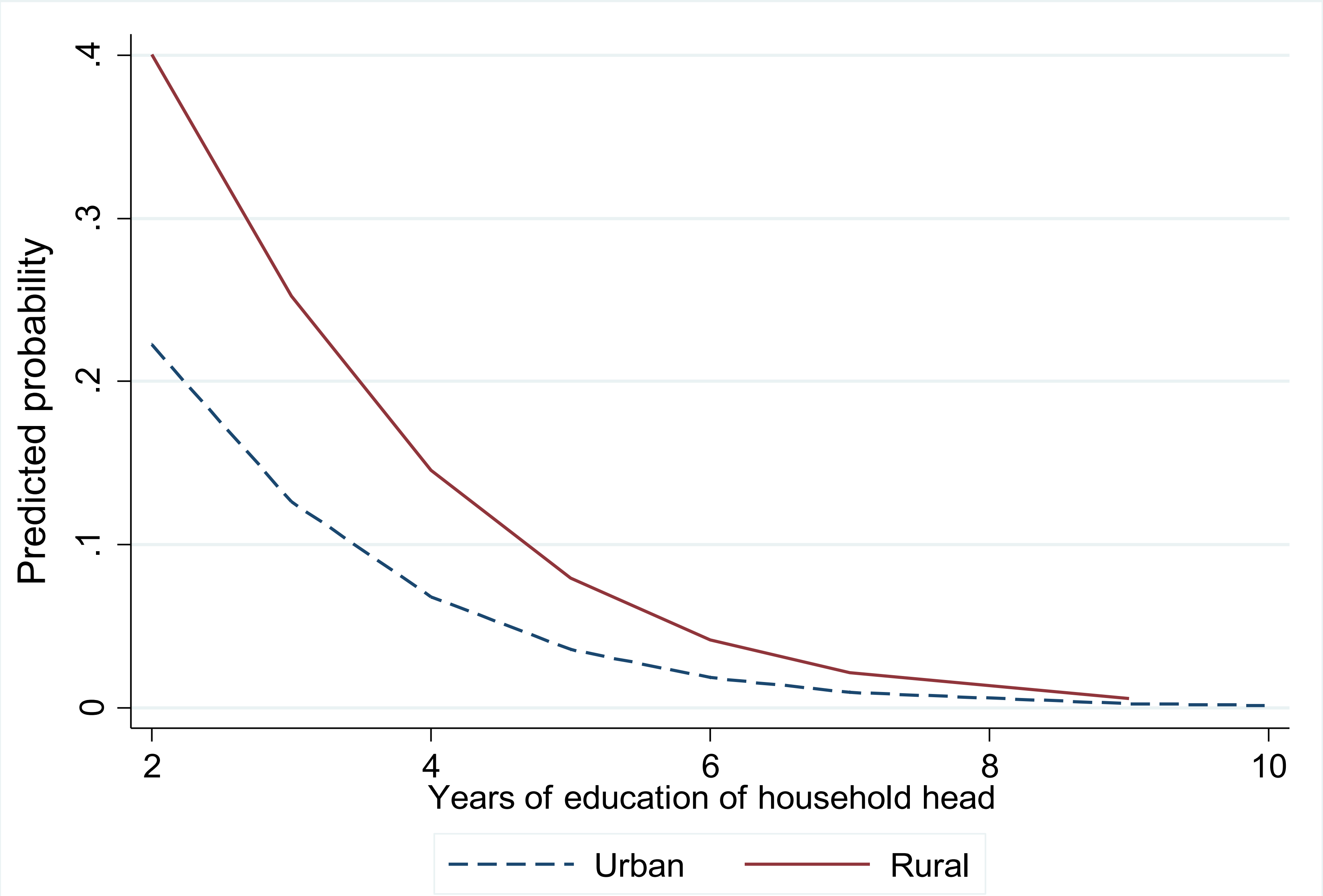
*** denotes significance at 5% level; n.s. denotes non-significance

Estimated parameters exhibiting a **negative** sign denote a **decrease** in the **odds**, this is obtained as (1-odds ratio)*100.

For the effect of **education** (1-0.51)*100 ↓49%,

For the effect of **gender** (1.28-1)*100% ↑ 28%.

Logistic regression



Macro Regression Models for M_0 and H

H and M0 are indices, **bounded** between zero and one

Thus an econometric model for these endogenous variables must account for the **shape** of their distribution, which has a **restricted range** of variation that lies in the unit interval.

H and M0 are therefore **fractional** (proportion) variables bounded between zero and one with the possibility of observing values at the boundaries.

Papke and Wooldridge (1996) Approach

To model H or M0 we follow the modeling approach proposed by Papke and Wooldridge (1996).

Papke and Wooldridge propose a particular **quasi-likelihood** method to estimate a proportion.

The method follows Gourieroux, Monfort and Trognon(1984) and McCullagh and Nelder (1989) and is based on the **Bernoulli log-likelihood function**

Econometric issues

The aim of most econometric regressions is to get a credible estimate of a relationship between two variables or phenomena. Sources of bias:

- **Endogeneity** can result from reverse causality (Y affects X and X affects Y) and confounding factors (Z affects both Y and X).
- **Measurement errors** can result from i) conceptual errors, ii) data collection errors, etc.
- **The specification problem:** Limits to inference without theory (*see* Kenneth Wolpin, 2013).

Dealing with Econometric Issues

- **The underlying theory:** using or building on theories to derive the regression model
- **Panel data estimation**: Fixed and Random Effects models: unobserved heterogeneity, omitted variables, degree of freedoms, etc.
- **Instrumental variables approach:** lagged variables, terms of trade, colonial origin, ecological and climatic variables, etc…
- **Robustness analysis:** with respect to the specification, the inclusion or exclusion of variables and observations, etc.