OPHI
OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE
www.ophi.org.uk

UNIVERSITY OF OXFORD

Regression Analysis with AF measures

Paola Ballón

Managua, 6 de Septiembre de 2013

El camino recorrido..

Resumamos los análisis que hemos efectuado con M_0:

- Descomposición en sus dos componentes H y A
- Descomposición por grupo y región
- Desglose por dimensión
- Análisis de asociación y similitud
- Cálculo de los errores estándar e intervalos de confianza
- Análisis de cambios en el tiempo
- Análisis de robustez

Que nos queda aún por analizar?

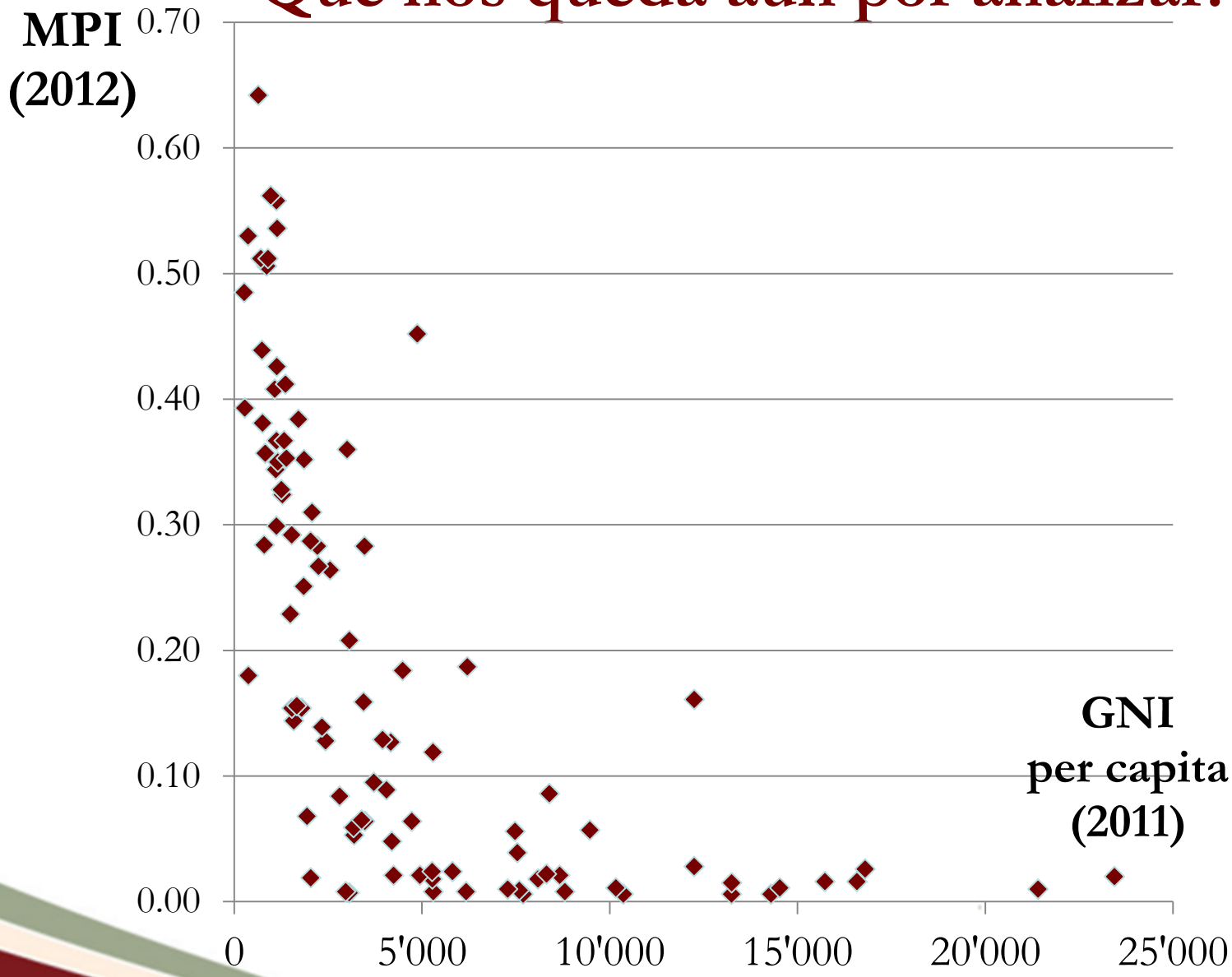
El siguiente ejemplo de Indonesia (1993) nos muestra las características de los hogares que son multidimensionalmente pobres (M_0=0.133) (Ballón & Apablaza, 2013)

Hogares identificados como pobres multidimensionales
Características del Jefe del Hogar

Promedio			*Proporcion*		
Años de educación	**Edad**	**Tamaño del hogar**	**Hombre**	**Musulman**	**Protestante**
2.1	25.5	5.1	80%	91%	2%

Este es un análisis **descriptivo,** aun **no conocemos** cuál es el « **efecto** » (“tamaño”) de cada una de estas características en la pobreza.

Que nos queda aún por analizar?



aun **no conocemos** el « **efecto** » de GNI en el MPI

Porqué es importante cuantificar estos "efectos"?

Desde un punto de vista de la política pública es útil comprender los mecanismos de transmisión de políticas macro en medidas de pobreza.

Cómo podemos cuantificar estos efectos?

El **análisis de regresión** permite cuantificar el "efecto/tamaño" de los determinantes micro y macro de la pobreza multidimensional.

Asimismo, nos permitirá analizar las interacciones entre $\mathbf{M_0}$ y variables *no* incluidas en el proceso de medición (escenario ideal)

Podemos diferenciar entre:

a) Regresiones 'micro' : unidad de análisis es el hogar o el individuo

b) Regresiones 'macro' : unidad de análisis es un agregado "espacial", como provincia, distrito o país.

Cuáles son los análisis vía regresión que nos interesaría estudiar con variables pertenecientes a la familia de medidas AF ?

Regresiones micro:

a) estudiar los determinantes de la pobreza a nivel del hogar

b) construir perfiles de pobreza;

Regresiones macro

a) estudiar la elasticidad de la pobreza al crecimiento económico

b) estudiar la relación entre variables macro como el ingreso promedio, gasto público, descentralización, en la pobreza multidimensional, o en su variación a traves del tiempo.

Cuales son las variables 'clave' que podemos regresionar?

Variable dependiente medida AF: Y	Rango de Y	Modelo de Regresión	Nivel	Distribución conditional $p_Y(y)$
Binaria $(c_i \geq k)$	0,1	Probabilidad	Micro	Bernoulli
M_0, H	[0,1]	Proportion	Macro	Binomial
$n_{00}, n_{01}, n_{01}, n_{11})$	0,1,2,..	Count	Micro	Poisson

El modelo de regresion Simple

$$E(y \mid x) = \beta_0 + \beta_1 x + u$$

variable explicada

variable explicativa

perturbación

E(y | x) es la media condicional de y dado x. Es importante saber que en este modelo $\mathbf{y \in \Re}$

β_0 es el intercepto, y β_1 es la pendiente

El supuesto principal es:
El valor promedio de *u no* depende de los valores de *x*.
El objetivo es estimar los parámetros β.

Regresiones "micro" con c_i

En el caso de regressiones micro, la variable central de análisis es el **score de privación censurado.** Recordemos que este score refleja la **distribución conjunta** de los hogares identificados como pobres multidimensionales.

El análisis más simple que podríamos considerar es el **model de probabilidad.** Este modelo cuantifica la probabilidad de un hogar de ser identificado como pobre multidimensional.

En la medición de pobreza con medidas AF, el modelo de probabilidad es equivalente a **comparar** el **score de privación** de un hogar con el umbral de pobreza multidimensional (k).

El modelo de probabilidad con c_i

Sabemos que si **ci** está por **encima** del umbral de pobreza multidimensional (k), el hogar es pobre multidimensional.

Esto se representa por medio de una variable **aleatoria binaria** (Y) que toma el valor de 1 si el hogar es pobre multidimensional, 0 si no.

El modelo de regresion con variable dependiente binaria

$$Y_i = \begin{cases} 1 & \text{si y solo si } c_i > k \\ 0 & \text{si no} \end{cases}$$

Está basado en la distribución de probabilidad Bernoulli.

Los valores que toma la variable binaria ocurren con probabilidad π_i . Esta probabilidad es condicional a las variables explicativas.

Los modelos que tradicionalmente se consideran en este caso son: Probit y Logit

Los modelos Probit y Logit

Porque usamos estos modelos, y no el modelo de regresión lineal simple?

El modelo de regresión lineal simple **no es adecuado** porque supone que el rango de variación de la variable dependiente puede tomar cualquier valor $(-\infty, +\infty)$

La **probabilidad sólo** puede tomar valores **entre 0 y 1**. Por lo tanto se requiere un modelo que garantice éste rango de variación acotado.

Los modelos de regresión Probit y Logit

Para ello se requiere una **función** que permita una correspondencia de los valores de *Y* al intervalo comprendido entre 0 y 1.

Para este propósito podemos emplear cualquier **distribución de probabilidad acumulada.**

En general uno emplea:

la distribución acumulada normal estándar (Modelo Probit),
la distribución acumulada logística (Modelo Logit).

El modelo Logit

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

El "logit" de la probabilidad π, es el logaritmo de las "chances" que la variable binaria tome el valor de 1 en lugar de 0. En nuestro caso, el logit nos da las "**chances relativas**" que el hogar sea pobre multidimensional.

Es interesante ver que el modelo **logit** es un modelo **lineal** y aditivo para el logaritmo de las chances, pero también podemos expresarlo como un **modelo multiplicativo** para las "chances":

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

El interés del análisis esta en estimar π_i.

Interpretación de los parámetros del modelo

β_j denota el coeficientes de regresión parcial correspondiente a la variable x_j. Cada β_j se interpreta ya sea como la **variación marginal** en el logit, o como el **efectos multiplicativo** en las "chances".

De esta manera, β_j indica el cambio en el logit debido a una variación unitaria en x_j , y e^{β_j} indica el efecto multiplicativo en las chances ante una variación unitaria de x_j. Por ello e^{β_j} es conocido como el "**ratio de chances**" (**odds ratio**) resultante de una variación unitaria en xj.

En ambos casos estos efectos se interpretan asumiendo que las otras variables explicativas no sufren cambio alguno.

Ejemplo

Está basado en una sub-muestra de la encuesta de hogares IFLS de Indonesia (Indonesian Family Life Survey (IFLS)) analizada por Ballon and Apablaza, (2012).

Los resultados de regresión que mostramos a continuación corresponden a la regresión logística estimada para la provincia West Java en 2007.

West Java es una provincia de Indonesia localizada en la parte oeste de la isla de Java que se caracteriza por su alta densidad poblacional.

Para la estimación del perfil de pobreza de los hogares en West Java asumimos un valor de k=33% y consideramos como variables explicativas, las características socio-demográficas de estos hogares.

Variables explicativas

•Años de educación del jefe del hogar: definido como el número de años de educación;

•Si el jefe del hogar es mujer: representado con una variables dummy que toma el valor de 1 si es el caso, y 0 si no,

•Tamaño del hogar: definido como el número de miembros del hogar;

Variables explicativas

•Área de residencia del hogar: representada con una variable dummy que toma el valor de 1 si el hogar reside en el área urbana de West Java y 0 si no;

•Si la religión principal del hogar es Musulmana: representada con una variable dummy variable que toma el valor de 1 si es el caso y 0 si no.

Estas características socio demográficas se escogieron de manera a restringir una posible endogeneidad entre el vector ci y las variables explicativas.

Resultados de la regresión logistica – 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

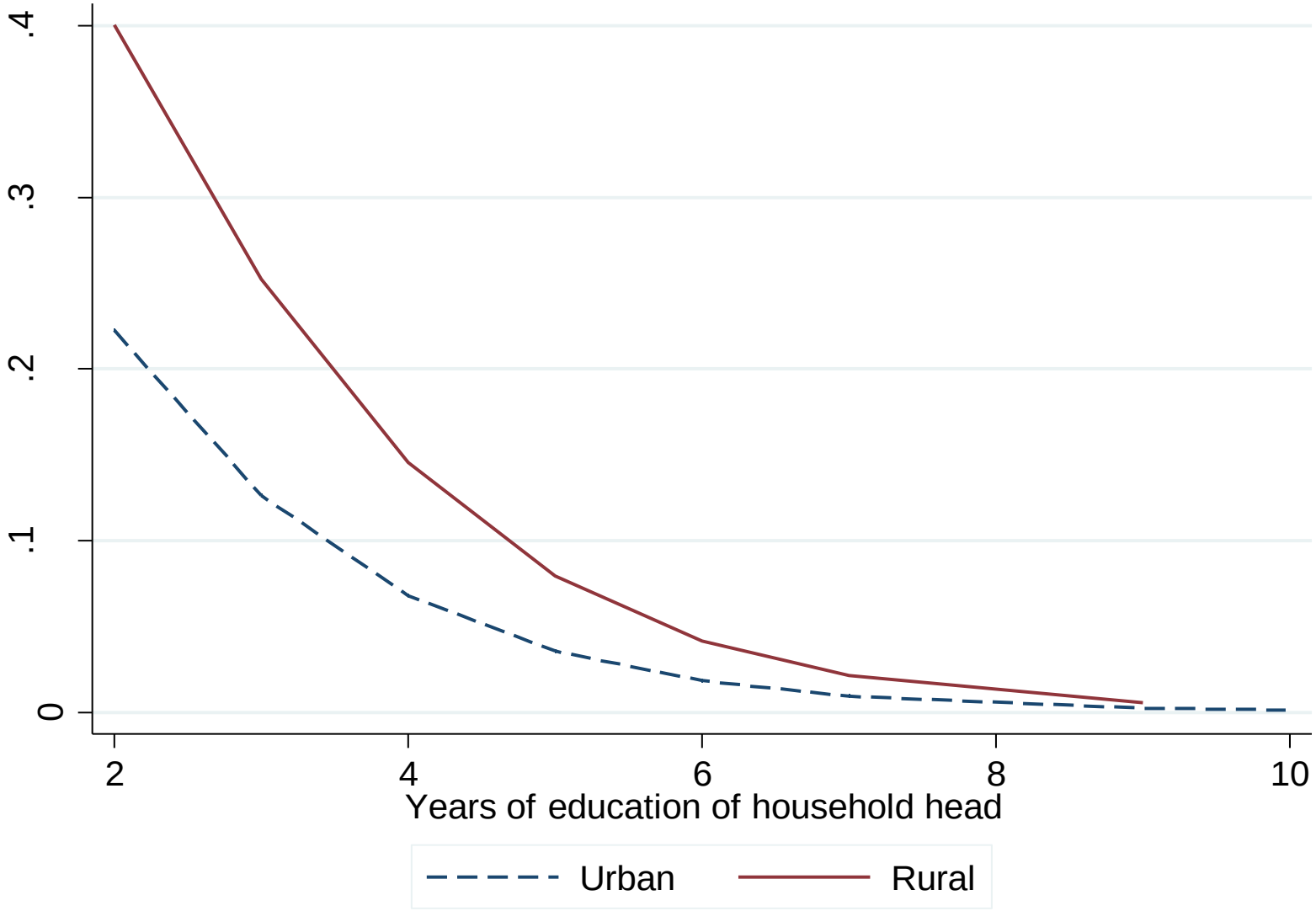
Los parámetros estimados con signo negativo indican una disminución en las chances (odds). La disminución se calcula como (1-odds ratio)*100.

Los parámetros estimados con signo positivo indican un aumento en el odds. El aumento se calcula como (odds ratio-1)*100.

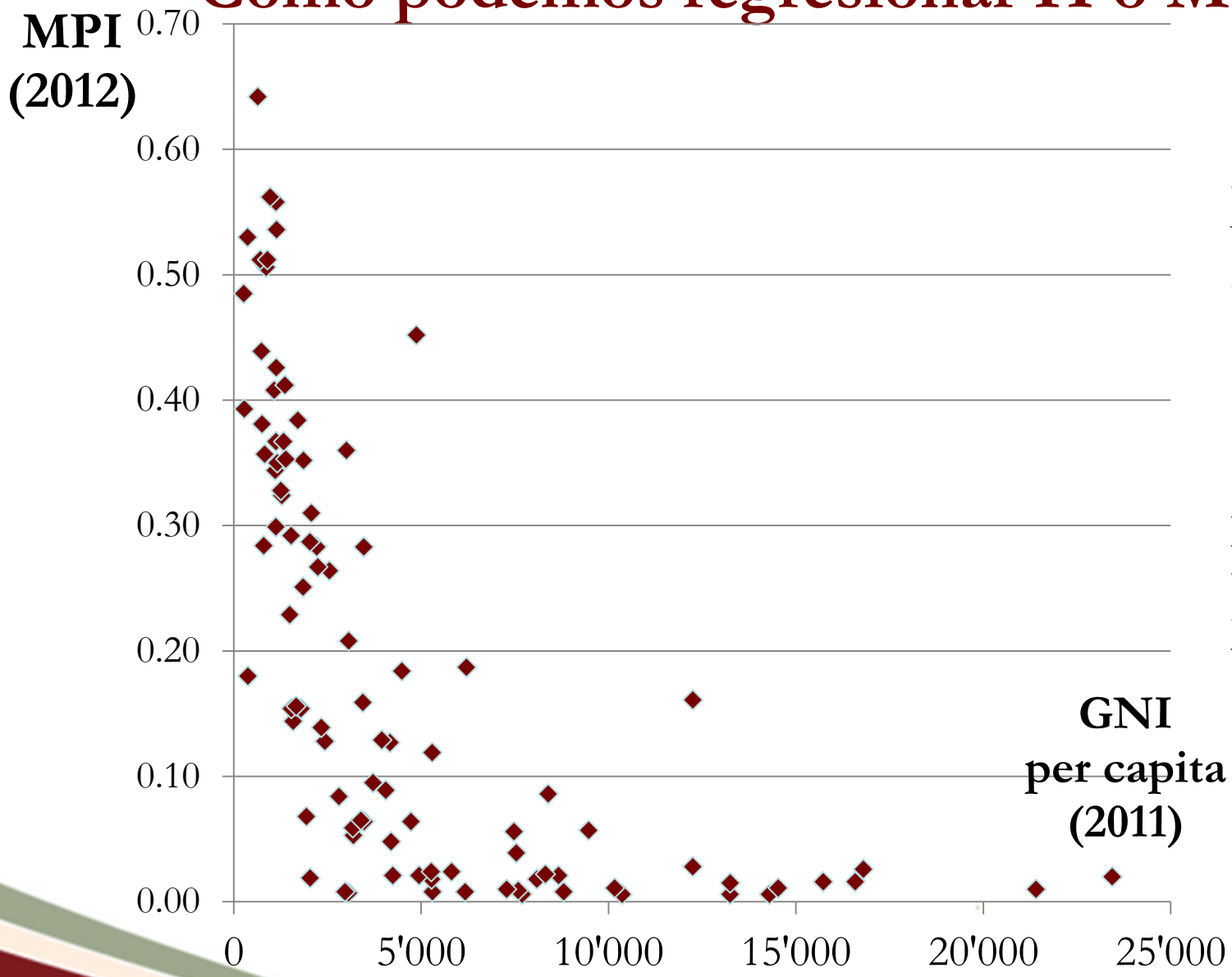
Para el efecto de la educación: (1-0.51)*100,

Para el efecto del género: (1.28-1)*100%.

Regresión Logística



Como podemos regresionar H o M0?



H o M0 son variables que toman valores entre 0 y 1

Cuál es el peligro de usar un modelo lineal?

Regresión con H o M0 como variables dependientes

H y M0 son índices, que representan **fracciones** o proporciones, que toman valores en el intervalo comprendido entre 0 y 1 (incluyendo los extremos)

Por lo tanto, un modelo econométrico para estas variables dependientes debe tomar en cuenta la **forma** de su distribución, que está **acotada** en el intervalo comprendido entre 0 y 1.

Podemos especificar un modelo lineal de regresión para H oM0?

Si especificamos un modelo de regresión lineal estamos **asumiendo** que la variable dependiente (y su esperanza condicional) pueden tomar **cualquier valor**.

Esto **no es adecuado** para variables de tipo fracción como H o M0, pues tal especificación ignora la forma de su distribución.

Sin embargo, si el **interés** del análisis no está en modelar las tasas de incidencia sino su **variación absoluta** entre dos periodos de tiempo, entonces uno puede **especificar** un modelo de **regresión lineal**, ya que la variación ya no esta acotada y puede tomar cualquier valor.

Enfoques para modelos donde la proporción es la variable dependiente

Existen dos enfoques que **difieren** en el **tratamiento** que se aplica a las observaciones en los **extremos de la distribución** (cuando la fracción toma el valor de 0 o 1).

Estos enfoques se denominan:

a) Una-etapa **(one-step).** Existe un **solo modelo** para el conjunto de observaciones que toma la proporción.

b) Dos-etapas **(two-step).** Las observaciones en los extremos se modelizan por separado. (Wagner, 2001; Ramalho and Ramalho, 2011).

La decisión de cual modelo es conveniente, en general, esta basada en criterios teóricos (Ballon, 2013).

Enfoques de una-etapa

Se pueden resumir en cuatro tipos:

a) **Ignorar** la naturaleza acotada de la proporción especificando un modelo lineal. Lo cual ya vimos que no es adecuado

b) Usar la **transformación log** de la proporción, después de añadir una constante arbitraria pequeña a cada observación. Esta transformación modifica la distribución de la proporción de manera *ad hoc* y por lo tanto tampoco es adecuada.

Four one-step approaches

c) Usar un modelo **Tobit, asumiendo que la proporción es una variable censurada en 1 y en 0.**

Como lo menciona Maddala (1991), este enfoque también tiene desventajas, pues los modelos Tobit son apropiados para modelizar variables que son censuradas por definición. Conceptualmente los modelos Tobit no son adecuados cuando la censura no es una consecuencia natural de la variable, que es le caso de las tasas de incidencia H o M0.

d) Asumir una **distribución condicional especifica** para la proporción. Por ejemplo la distribución Beta distribución. La distribución Beta, es una distribución para una variable continua definida en el intervalo abierto (0,1). (Wagner, 2001; Ramalho and Ramalho, 2011) .

Papke and Wooldridge approach

Para modelizar H o M0 usamos el enfoque propuesto por Papke and Wooldridge (1996).

Papke y Wooldridge proponen un método de estimación particular basado en la **quasi-likelihood**

propuesto por Gourieroux, Monfort y Trognon(1984), McCullagh y Nelder (1989) que

usa la función log-likelihood de la Bernoulli.

Algunas referencias

Para una comparación **teórica y conceptual** entre los enfoques de una vs dos partes pueden referirse a : Ballon (2013)

Para modelos tipo **Bi-probit**, pueden referirse a: Chatterjee, M. (2013)

Para una **estudio empírico** de los **determinantes** macro de H and M0 en Indonesia, pueden referirse a: Ballon and Apablaza(2012)

Gracias ☺