Survey nonresponse and the distribution of income

Emanuela Galasso*
Development Research Group, World Bank

Module 1. Sampling for Surveys

1: Why are we concerned about non response?

2: Implications for measurement of poverty and inequality

3: Evidence for the US

  – Estimation methods

  – Results

4: An example for China

1: Why do we care?

Types of nonresponse

- **Item-nonresponse**
- (participation to the survey but non-response on single questions)
  - Imputation methods using matching
    - Lillard et al. (1986); Little and Rubin (1987)

Types of nonresponse

- **Item-nonresponse**
  - Imputation methods using matching
    - Lillard et al. (1986); Little and Rubin (1987)
- The idea:

Observations with…	$\boldsymbol{X}$	$\boldsymbol{Y}$
complete data	Yes	Yes
missing data	Yes	No

  - For sub-sample with complete data: $Y = M(X)$
  - Then impute missing data using: $\hat{Y} = \hat{M}(X)$

Types of nonresponse

- **Unit-nonresponse (“non-compliance”)**
- (non-participation to the survey altogether)

Unit-nonresponse: possible solutions

Ex-ante:

– Replace non respondents with similar households

– Increase the sample size to compensate for it

– Using call-backs, monetary incentives:

  - Van Praag et al. (1983), Alho (1990), Nijman and Verbeek (1992)

Ex-post: Corrections by re-weighting the data

– Use imputation techniques (hot-deck, cold-deck, warm-deck, etc.) to simulate the answers of nonrespondents

Unit-nonresponse: possible solutions

Ex-ante:

- Replace nonrespondents with similar households
- Increase the sample size to compensate for it
- Using call-backs, monetary incentives:
  - Van Praag et al. (1983), Alho (1990), Nijman and Verbeek (1992)

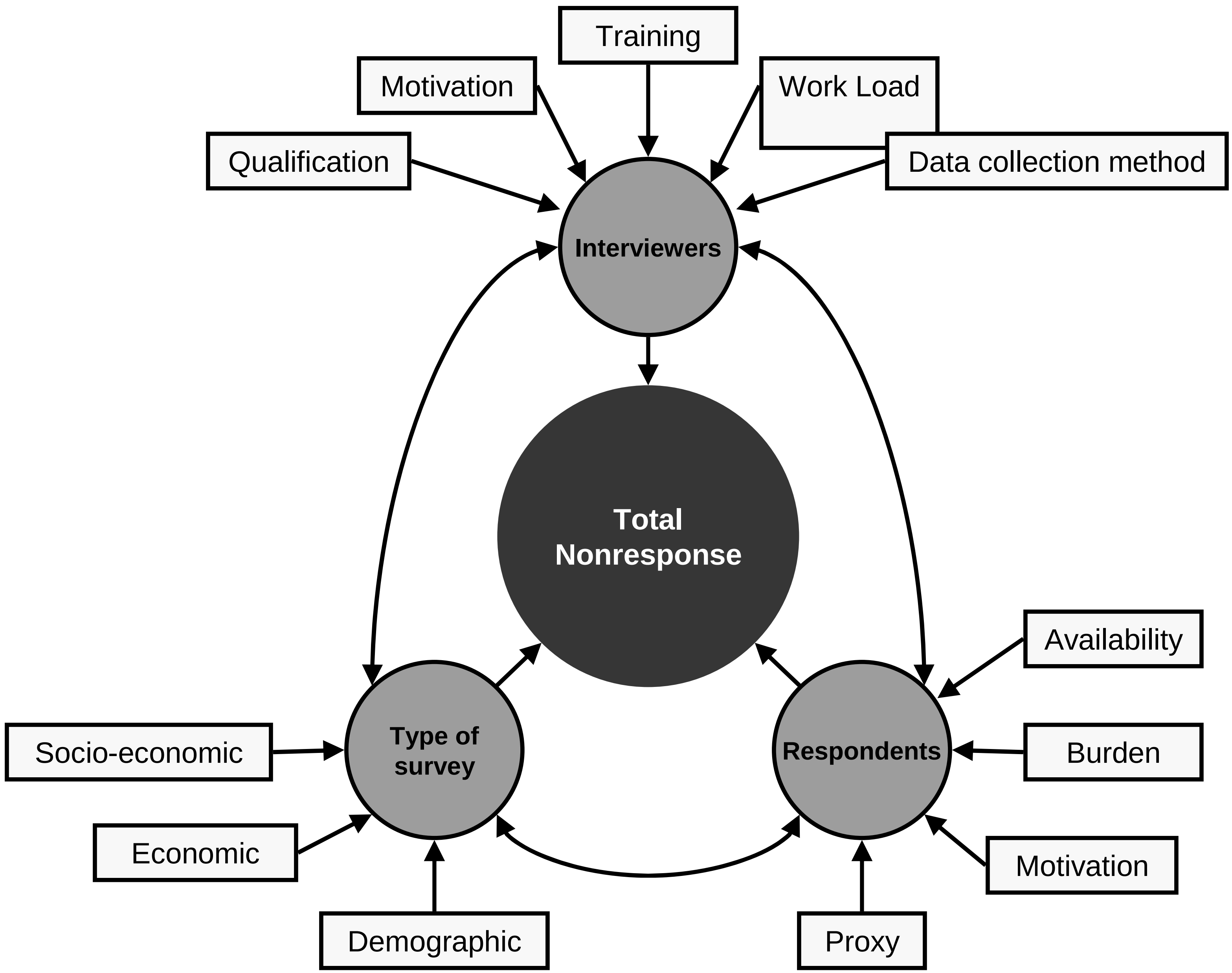
Ex-post: Corrections by re-weighting the data

- Use imputation techniques (hot-deck, cold-deck, warm-deck, etc.) to simulate the answers of nonrespondents

- None of the above...

The best way to deal with unit-nonresponse is to <u>prevent it</u>

Lohr, Sharon L. *Sampling: Design & Analysis* (1999)



Source: “Some factors affecting Non-Response.” by R. Platek. 1977. *Survey Methodology*. *3*. 191-214

Rising concern about unit-nonresponse

- **High nonresponse rates of 10-30% are now common**
  - LSMS: 0-26% nonresponse (Scott and Steele, 2002)
  - UK surveys: 15-30%
  - US: 10-20%
- **Concerns that the problem might be increasing**

Nonresponse is a <u>choice</u>, so we need to understand behavior

- Survey participation is a matter of choice
  - nobody is obliged to comply with the statistician's randomized assignment
- There is a perceived utility gain from compliance
  - the satisfaction of doing one's civic duty
- But there is a cost too
- An income effect can be expected

Nonresponse bias in measuring poverty and inequality

Compliance is unlikely to be random:

- Rich people have:
  - higher opportunity cost of time
  - more to hide (tax reasons)
  - more likely to be away from home?
  - multiple earners
- Poorest might also not comply:
  - alienated from society?
  - homeless

2: Implications for poverty and inequality measures

Implications for poverty

- $F(y)$ is the true income distribution, density $f(y)$
- $\hat{F}(y)$ is the observed distribution, density $\hat{f}(y)$
- Note: $F(y_p) = \hat{F}(y_p) = 0$ and $F(y_r) = \hat{F}(y_r) = 0$

Implications for poverty

- $F(y)$ is the true income distribution, density $f(y)$
- $\hat{F}(y)$ is the observed distribution, density $\hat{f}(y)$
- Note: $F(y_p) = \hat{F}(y_p) = 0$ and $F(y_r) = \hat{F}(y_r) = 0$

Definition: <u>correction factor</u> $w(y)$ such that:

$$f(y) = w(y)\hat{f}(y)$$

$$F(y) = \int_{y_p}^{y} w(x)\hat{f}(x)dx$$

Implications for poverty cont.,

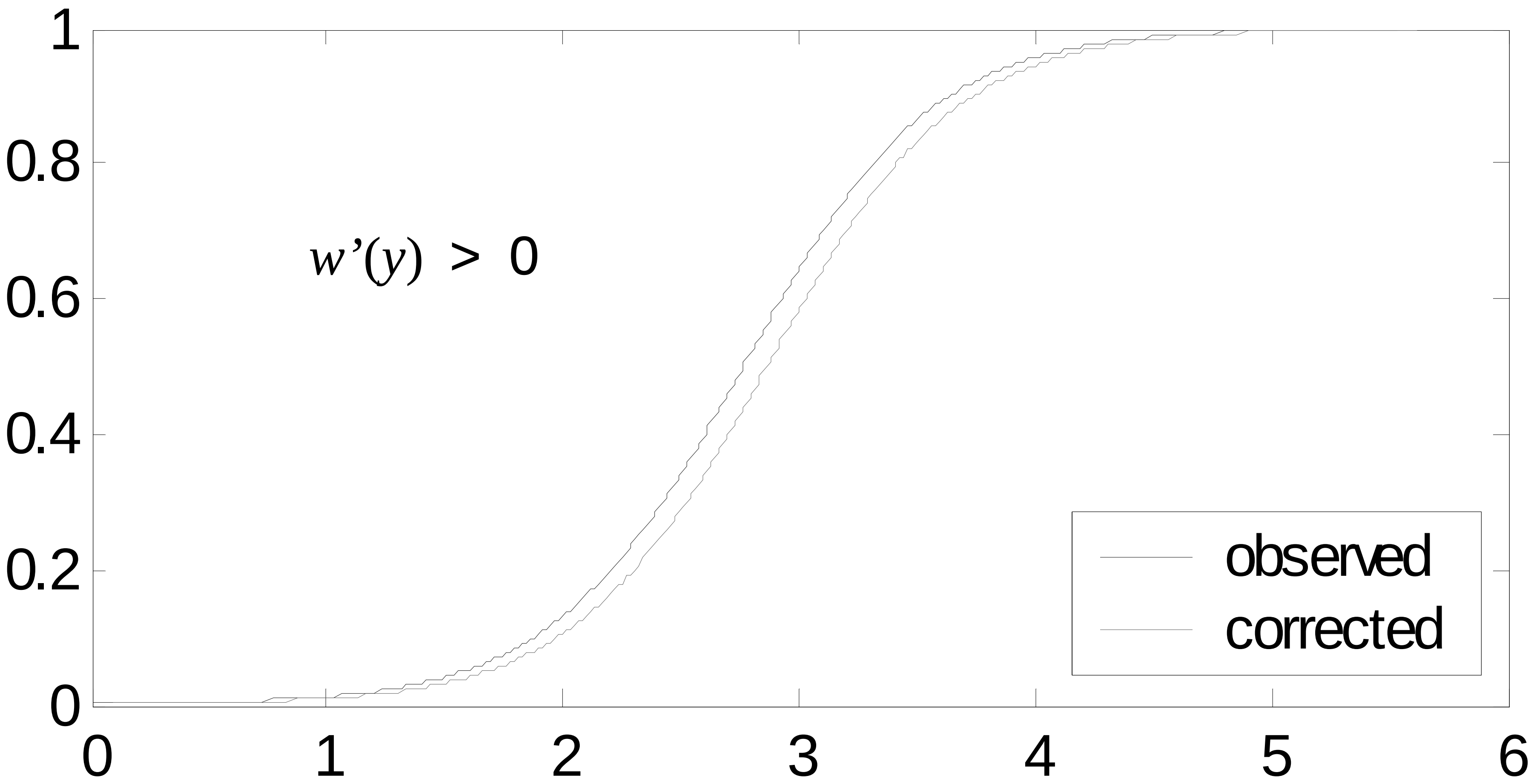
If compliance falls with income then poverty is overestimated for all measures and poverty lines.

i.e., <u>first-order dominance</u>:

if $w'(y) > 0$ for all $y \in (y_P, y_R)$,

then $F(y) < \hat{F}(y)$ for all $y \in (y_P, y_R)$

First-order dominance



Example

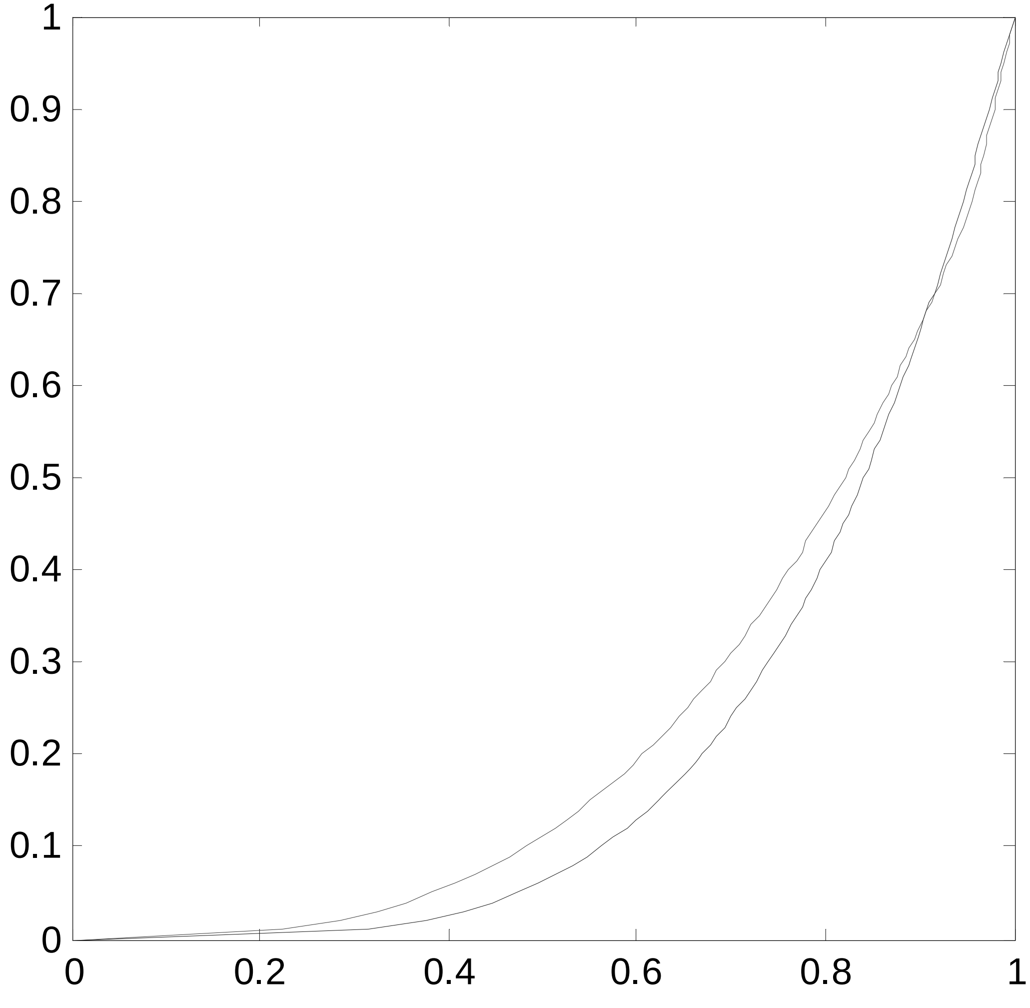
	"Poor"	"Non-poor"
Estimated distribution (%)	81	19
However,…		
Response rate (%)	90	50
True distribution of population (%)	70	30
Correction factors	0.87	1.56

Implications for inequality

If compliance falls with income ($w'(y) > 0$) then the implications for inequality are ambiguous

Lorenz curves intersect so some inequality measures will show higher inequality, some lower

Example of crossing Lorenz Curves



3: Evidence for the U.S.

Current Population Survey

Source: CPS March supplement, 1998 – 2002, Census Bureau

3 types of "non-interviews:"

- **type A**: individual refused to respond or could not be reached → what we define as "non-response"
- **type B**: housing unit vacant; type C: housing unit demolished → we ignore type B/C in our analysis

Year	total number of households	Type A households	rate of non-response (%)
1998	54,574	4,221	7.73%
1999	55,103	4,318	7.84%
2000	54,763	3,747	6.84%
2001	53,932	4,299	7.97%
2002	84,831	6,566	7.74%
All years	303,203	23,151	7.64%

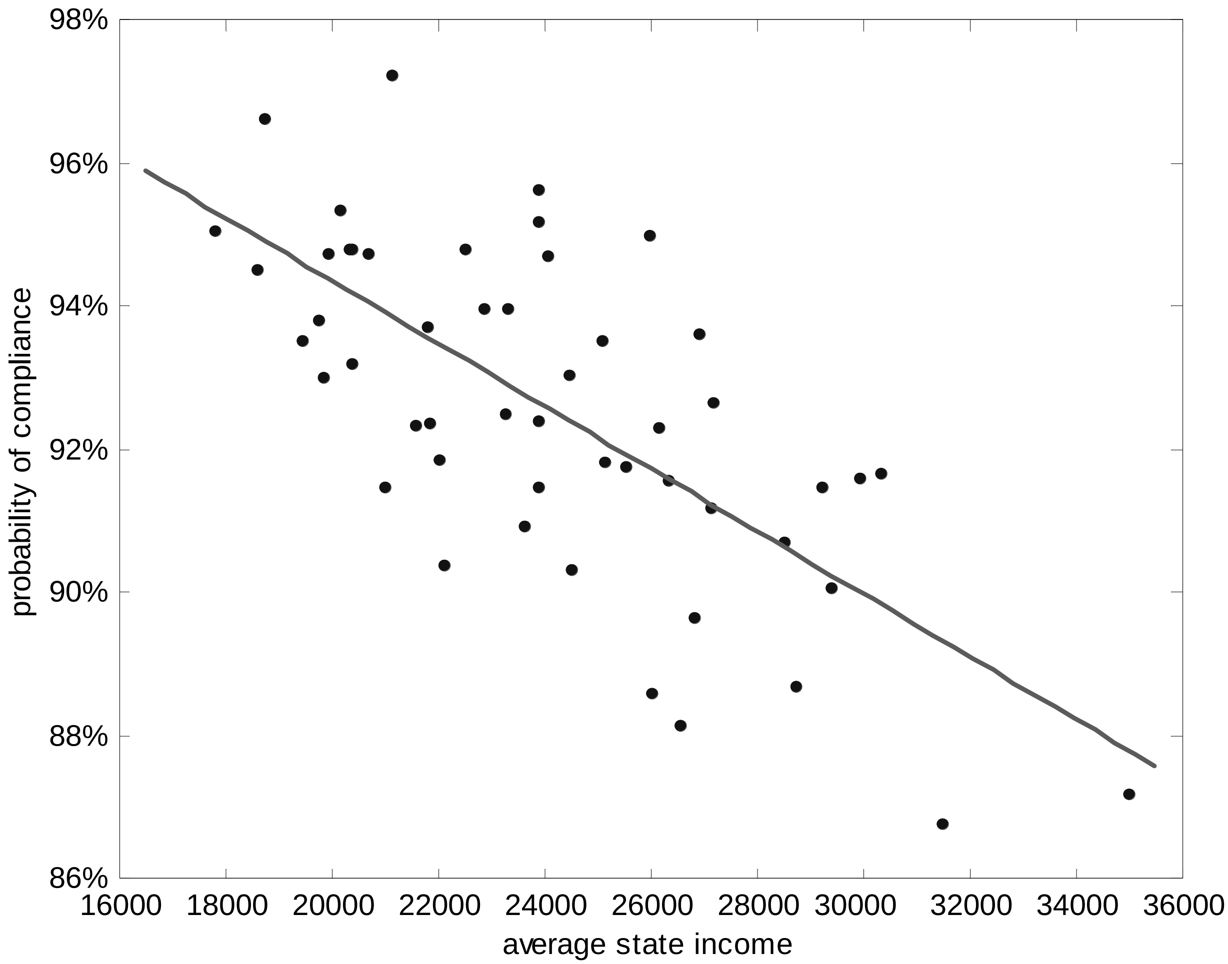
Dependence of response rate on income

Response rate and average per-capita income for 51 US states,
CPS March supplement 2002

State	Response Rate	Average Income
Maryland	86.77%	$31,500
District Of Columbia	87.21%	$34,999
Alaska	88.16%	$26,564
New York	88.61%	$26,013
New Jersey	88.71%	$28,746
California	89.66%	$26,822
...	...	...
Mississippi	95.08%	$17,821
Indiana	95.21%	$23,909
North Dakota	95.36%	$20,154
Georgia	95.66%	$23,893
West Virginia	96.65%	$18,742
Alabama	97.24%	$21,155

Dependence of response rate on income

Response rate and average per-capita income for 51 US states, CPS March supplement 2002



Estimation method

- In survey data, the income of non-responding households is by definition unobservable.
- However, we can observe the survey compliance rates by geographical areas.
- The observed characteristics of responding households, in conjunction with the observed compliance rates of the areas in which they live, allow one to estimate the household-specific probability of survey response.
- Thus we can correct for selective compliance by re-weighting the survey data.

Estimation method cont.,

- $\{(X_{ij}, m_{ij})\}$ ... set of households in state j s.t. m_{ij} households each carry characteristics X_{ij},
  where X_{ij} includes e.g. ln(y_{ij}), a constant, etc.
- total number of households in state j: M_j
- representative sample S_j in state j with sampled households $m_j = \Sigma\, m_{ij}$
- for each sampled household ε there's a probability of response $D_{\varepsilon ij}$ {0,1}

$$P(D_{\varepsilon ij} = 1 \mid X_{ij}, \theta) = P_i = \operatorname{log} istic(X_i \theta)$$

Estimation method cont.,

- The observed mass of <u>respondents</u> of group *i* in state/area *j* is

$$E(m_{ij}^{obs}) = m_{ij} P_i$$

$$E[\frac{m_{ij}^{obs}}{P_i}] = m_{ij}$$

- Then summing up for a given j yields:

$$\left[\sum_i \frac{m_{ij}^{obs}}{P_i} \right] = \sum_{i=1} w_{ij} = m_j$$

- Now let's define:

$$\psi_j(\theta) \equiv \sum_i \{ \frac{m_{ij}^{obs}}{P_i} - E[\frac{m_{ij}^{obs}}{P_i}] \} = \sum_i \{ \frac{m_{ij}^{obs}}{P_i} - m_j \}$$

These are the individual weights

This is known!

Estimation method cont.,

$$\psi_j(\theta) \equiv \sum_i \{\frac{m_{ij}^{obs}}{P_i} - E[\frac{m_{ij}^{obs}}{P_i}]\} = \sum_i \{\frac{m_{ij}^{obs}}{P_i} - m_j\}$$

where obviously $E\lfloor \psi_j(\theta) \rfloor = 0$

Then we can estimate

$$\hat{\theta} = \arg\min_{(\theta)} \Psi(\theta) \equiv \psi(\theta)' W^{-1} \psi(\theta)$$

Estimation method cont.,

Optimal weighting matrix W = Var($\psi(\theta)$) ... Hansen (1982)

Assume for single state j: $$Var[\psi_j(\theta)] = m_j \sigma^2$$

This can be estimated as $$\hat{\sigma}^2 = \frac{\sum \psi_j(\theta)^2}{\sum w_j}$$

Finally, $\hat{Var}(\hat{\theta}) = \hat{\sigma}^2 [G' NG]^{-1}$ where $G = \dfrac{\partial \psi(\theta)}{\partial \theta}$

Alternative Specifications

Specification	$\Psi(\theta)_{\min}$	θ_1	θ_2	θ_3
1: $P = logit(\theta_1 + \theta_2 \, ln(y) + \theta_3 \, ln(y)^2)$	27.866	32.55 (85.95)	-4.151 (-15.90)	0.1193 (0.7320)
2: $P = logit(\theta_1 + \theta_2 \, ln(y))$	27.940	17.81 (3.51)	-1.489 (-0.329)	
3: $P = logit(\theta_1 + \theta_3 \, ln(y)^2)$	28.068	9.551 (1.646)		-0.0666 (-0.0145)
4: $P = logit(\theta_2 \, ln(y) + \theta_3 \, ln(y)^2)$	28.324		1.725 (0.289)	-0.1438 (-0.0272)
5: $P = logit(\theta_1 + \theta_2 \, y)$	34.303	2.995 (0.202)	$-13.11 \cdot 10^{-6}$ $(-4.76 \cdot 10^{-6})$	
6: $P = logit(\theta_1 + \theta_2 \, y + \theta_3 \, y^2)$	28.639	3.792 (0.463)	$-37.45 \cdot 10^{-6}$ $(-14.92 \cdot 10^{-6})$	$67.21 \cdot 10^{-12}$ $(58.33 \cdot 10^{-12})$
7: $P = logit(\theta_1 + \theta_2 \, y + \theta_3 \, ln(y))$	27.891	19.64 (12.13)	$1.889 \cdot 10^{-6}$ $(17.35 \cdot 10^{-6})$	-1.671 (-1.229)

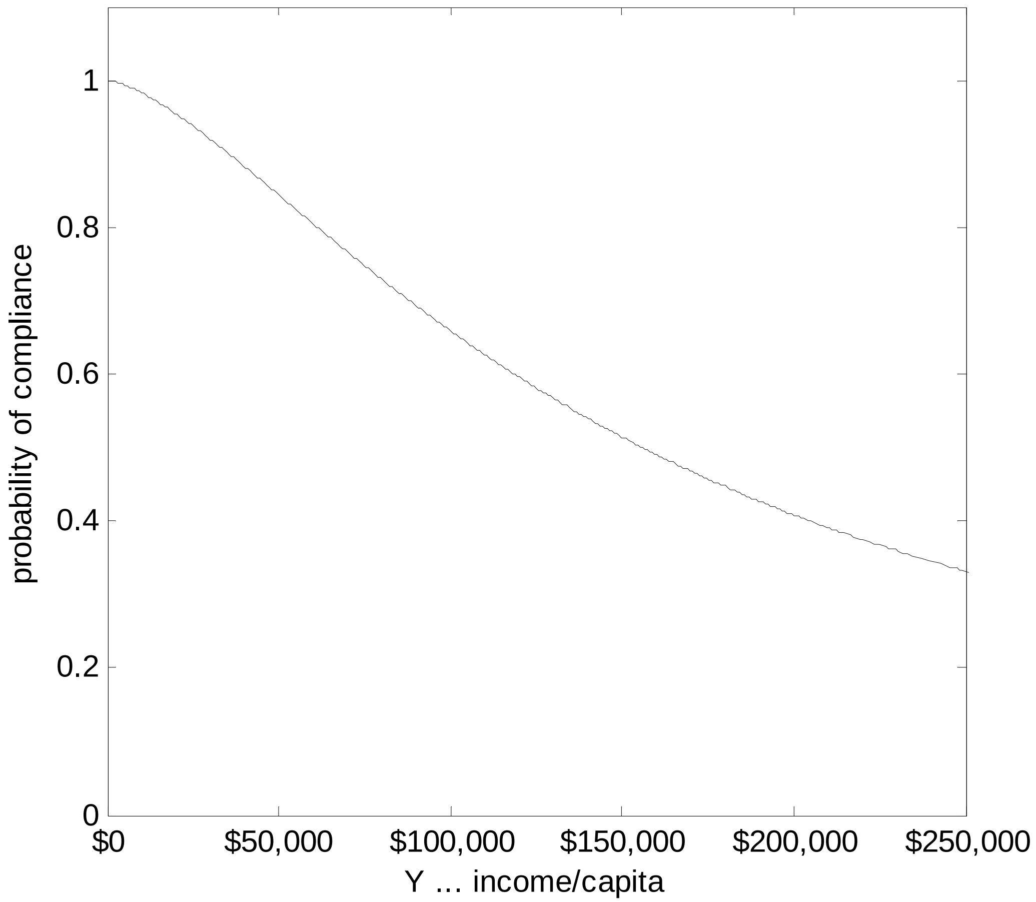
Results From Specification 2

$P = logit(\theta_1 + \theta_2\ ln(y))$

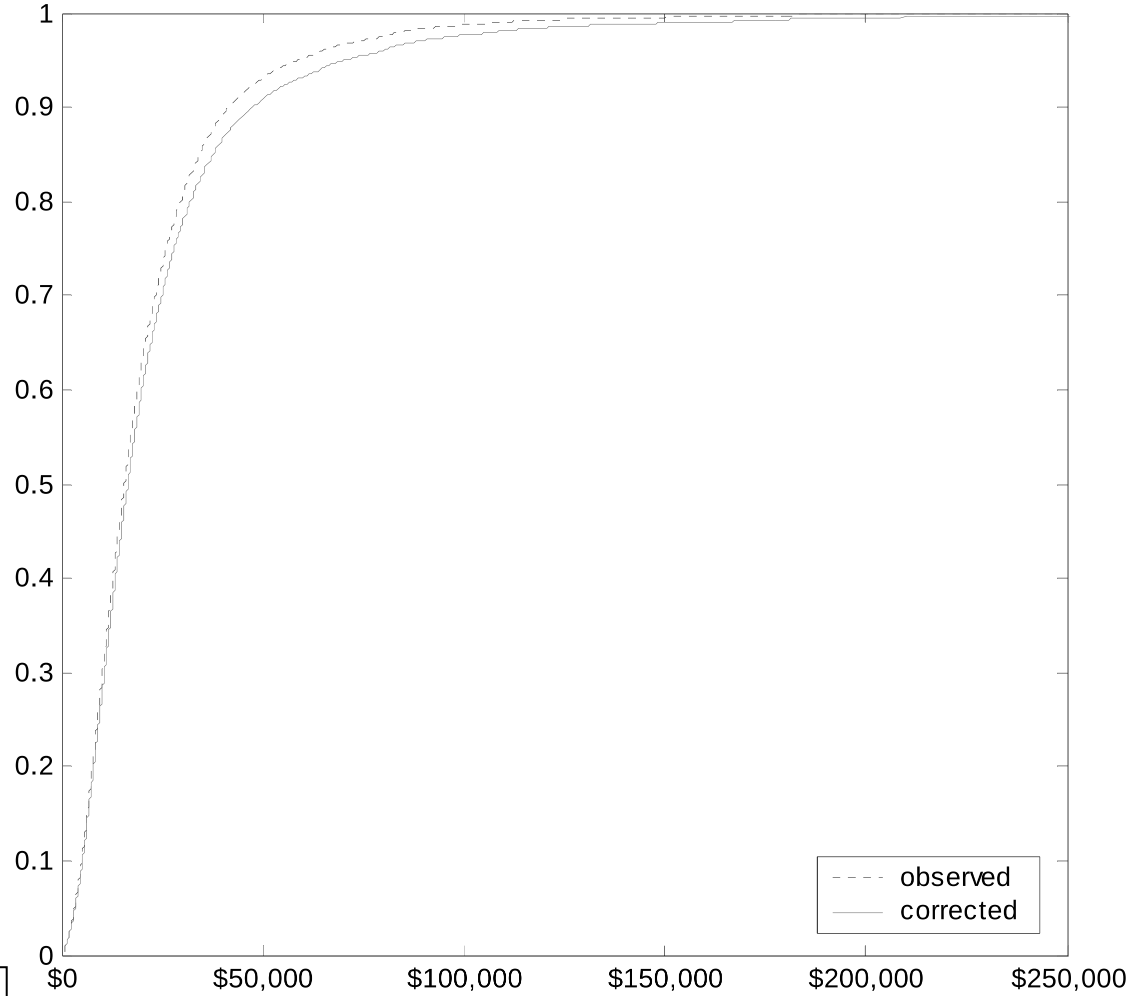
Year	$\Psi(\theta)_{min}$	θ_1	θ_2	Gini $_{uncorr}$	Gini $_{corr}$	ΔGini
1998	17.321	19.90 (4.58)	-1.697 (-0.43)	45.49%	50.92%	5.43%
1999	21.437	18.10 (4.42)	-1.528 (-0.418)	45.21%	49.03%	3.82%
2000	12.558	22.21 (4.46)	-1.890 (-0.413)	44.30%	47.67%	3.37%
2001	17.793	20.11 (3.82)	-1.702 (-0.355)	44.99%	49.47%	4.48%
2002	27.94	17.81 (3.51)	-1.489 (-0.329)	44.36%	48.02%	3.66%
All	102.16	19.47 (1.89)	-1.654 (-0.177)	44.83%	49.07%	4.24%

Graph of specification 2:

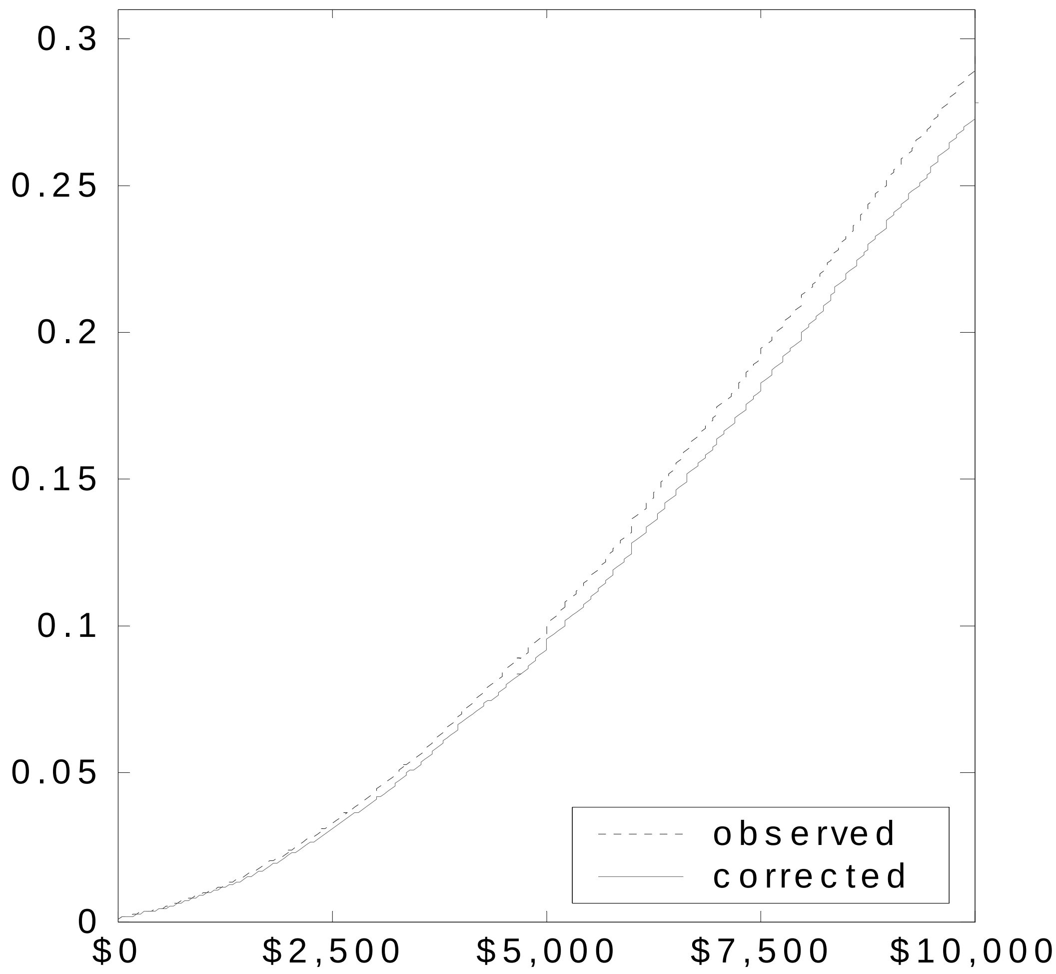
Probability of compliance as a function of income



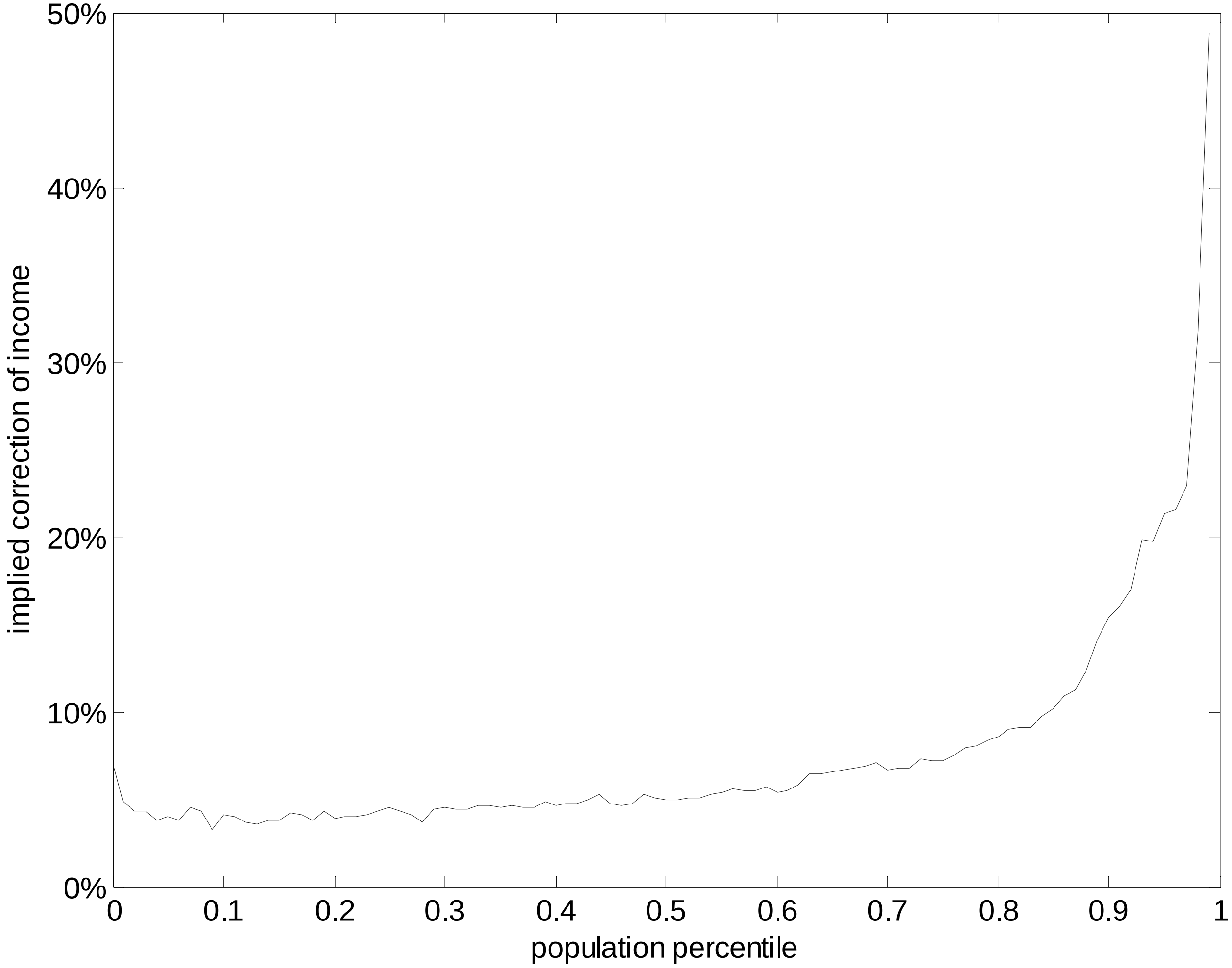
Empirical and Corrected Cumulative Income Distribution



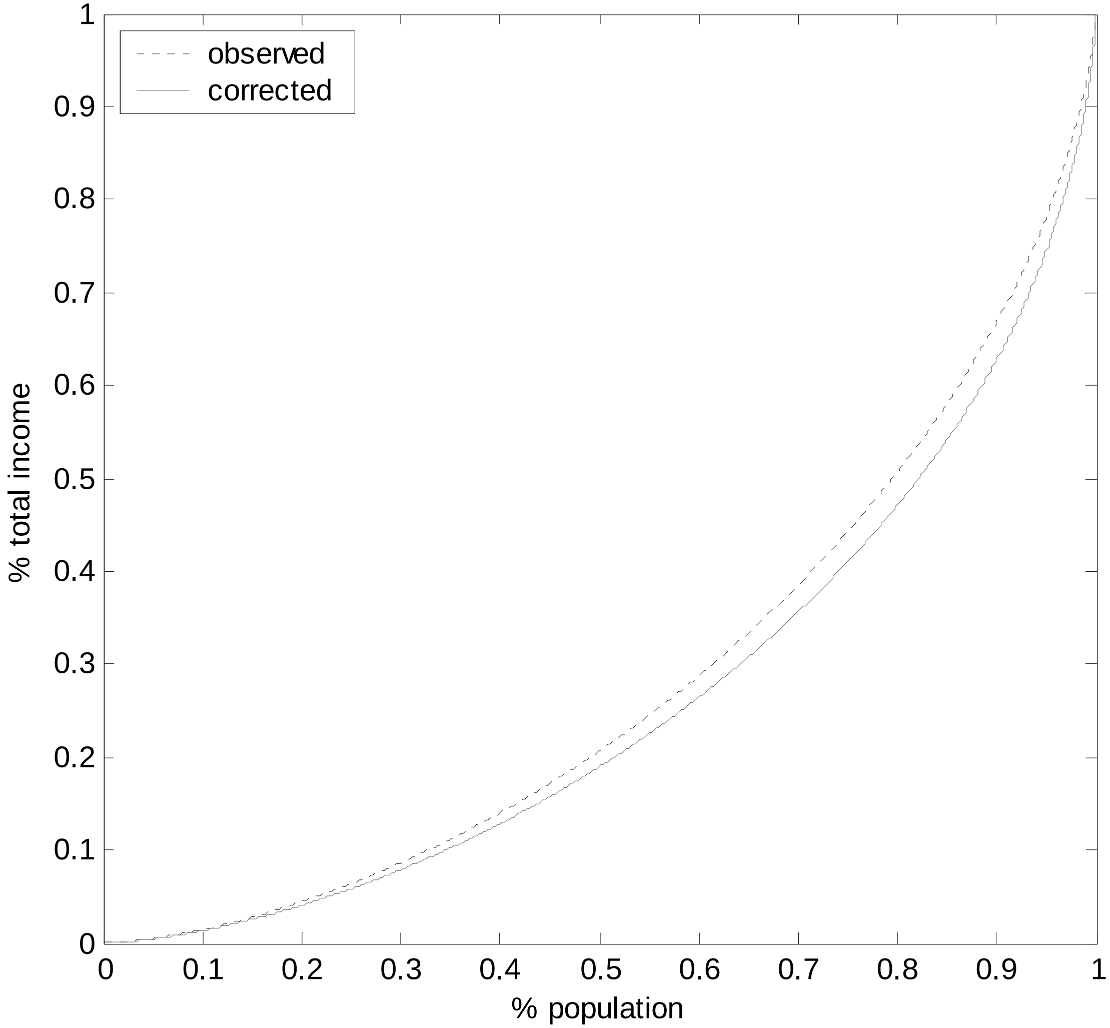
Income Distribution: Magnification



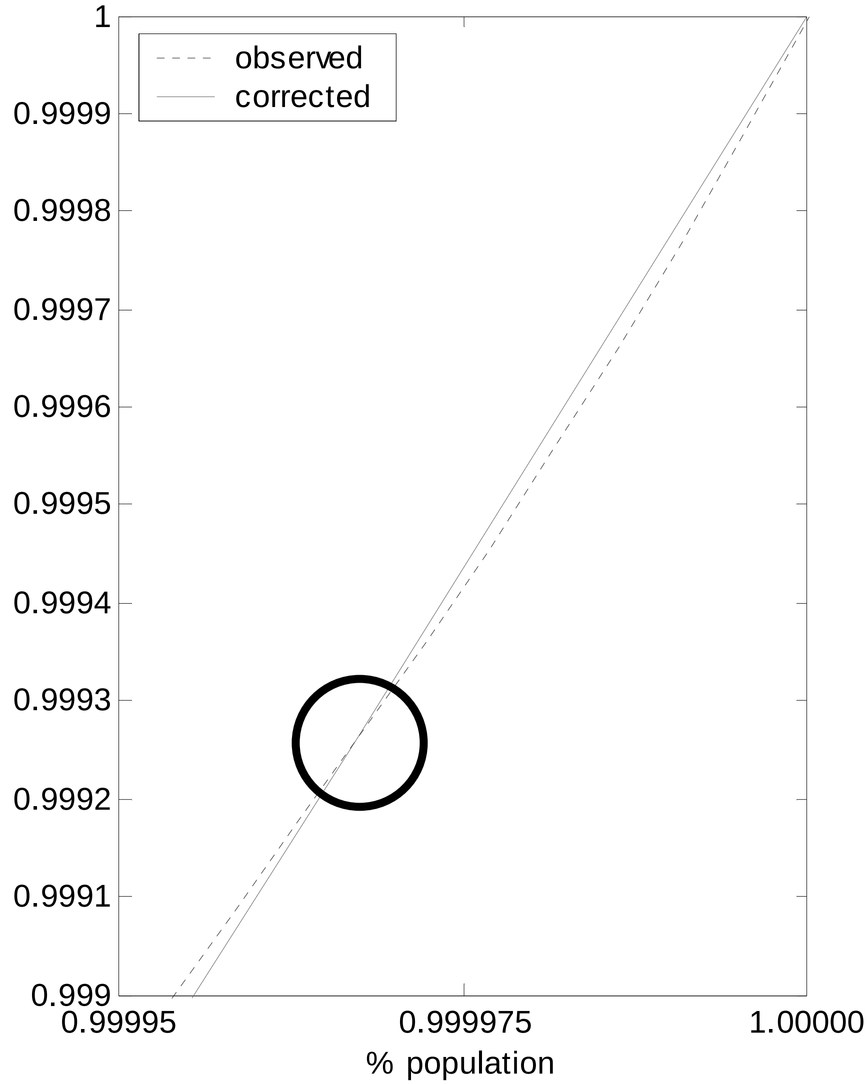
Correction by Percentile of Income



Empirical and Corrected Lorenz Curve



Lorenz Curves: Magnification



Specifications with Other Variables

Specifications 10 – 18, P = logit($\theta_1 + \theta_2 \ln(y) + \theta_3 X_1 + \theta_4 X_2$):

Specification	$\Psi(\theta)_{min}$	θ_1	θ_2	θ_3	θ_4	Gini $_{corrected}$	ΔGini
2: (baseline)	102.16	19.47 (1.89)	-1.654 (-0.177)			49.07%	4.24%
10: X_1 = *age* X_2 = *age*^{2}	100.19	17.78 (3.52)	-1.695 (-0.188)	0.09321 (0.10569)	-0.00092 (-0.00095)	49.09%	4.26%
11: X_1 = *age*	101.84	20.17 (2.61)	-1.679 (-0.201)	-0.00888 (-0.01648)		49.23%	4.40%
12: X_2 = *age*^{2}	101.57	20.11 (2.31)	-1.688 (0.198)		-0.00010 (-0.00014)	49.26%	4.43%
13: X_1 = *(age>64)*	100.52	20.04 (2.06)	-1.696 (-0.188)	-0.6123 (-0.5753)		49.23%	4.40%
14: X_1 = *edu* X_2 = *edu*^{2}	99.795	25.90 (7.59)	-1.469 (-0.334)	-1.481 (-1.447)	0.06235 (0.06456)	48.52%	3.69%
15: X_1 = *edu*	101.15	18.71 (2.48)	-1.502 (-0.333)	-0.08292 (-0.12667)		48.68%	3.85%
16: X_1 = *(edu>39)*	98.725	18.44 (1.93)	-1.456 (-0.233)	-1.352 (-1.187)		48.53%	3.70%
17: X_1 = *sex*	101.00	19.37 (1.92)	-1.627 (-0.187)	-0.4785 (-0.5315)		48.84%	4.01%
18: X_1 = *race*	93.353	17.51 (1.96)	-1.516 (-0.183)	0.5877 (0.1592)		48.26%	3.43%
19: X_1 = *size*	100.11	21.51 (2.12)	-1.777 (-0.189)		-0.3102 (-0.1316)	49.15%	4.32%
20: X_1 = *race* X_2 = *size*	91.709	19.15 (2.16)	-1.618 (-0.189)	0.5672 (0.1574)	-0.229 (-0.1289)	48.38%	3.55%

4: China

Example for China

- Urban Household Survey of NBS
- Two stages in sampling
  - Stage 1: Large national random sample with very short questionnairre and high repsonse rate
  - Stage 2: Random sample drawn from Stage 1 sample, given very detailed survey, including daily diary, regular visits etc
- Use Stage 1 data to model determinants of compliance
- Then re-weight the data

Further reading

- Korinek, Anton, Johan Mistiaen and Martin Ravallion, “An Econometric Method of Correcting for unit Nonresponse Bias in Surveys,” *Journal of Econometrics,* (2007), 136: 213-235

- Korinek, Anton, Johan Mistiaen and Martin Ravallion, “Survey Nonresponse and the Distribution of Income.” *Journal of Economic Inequality*, (2006), 4:33-55