ITSEW 2010 – Stowe
June 13-16, 2010

To misreport or not to report? The case of the Italian Survey on Household Income and Wealth

Andrea Neri
Economic and Financial Statistics Department, Bank of Italy, Rome, Italy
M. Giovanna Ranalli
Department of Economics, Finance and Statistics, University of Perugia, Italy

SHIW - Survey on Household Income and Wealth

- Surveys households income, financial assets, e.g. government bonds, private bonds, shares, managed savings, and non-financial wealth
- Run every 2 years by the Bank of Italy (latest edition: 2008)
- Two stage clustered design: municipalities (Stratified πPS) are PSUs and households HHs (SRSWOR) are SSUs
- Panel component
- Face to face CAPI
- Post-stratified estimator on demographic characteristics
- Nonsampling errors:
  - Item nonresponse (negligible)
  - Measurement error and Unit nonresponse (major issues)

National Financial Accounts - NFAs

	Gov. Bonds	Private Bonds	Shares	Mutual Funds
Number of households (Millions)				
SHIW estimate	2.180	1.710	1.450	1.495
NFA estimate	1.845	4.471	2.865	2.716
SHIW/NFA	*1.182*	*0.383*	*0.506*	*0.551*
Amount held (Billions of Euros)				
SHIW estimate	65.610	61.190	31.862	50.992
NFA estimate	160.840	303.450	63.026	108.219
SHIW/NFA	*0.408*	*0.202*	*0.506*	*0.471*

Outline

- Unit nonresponse
  - Panel HHs
  - Non-panel HHs
- Measurement error for
  - 1/0 variables (possession of a financial asset)
  - continuous variables (amount held)
- Final total estimates obtained
  - via imputation
  - or via calibration
- Conclusions and points for discussion

Two major issues - 1 - Unit Nonresponse

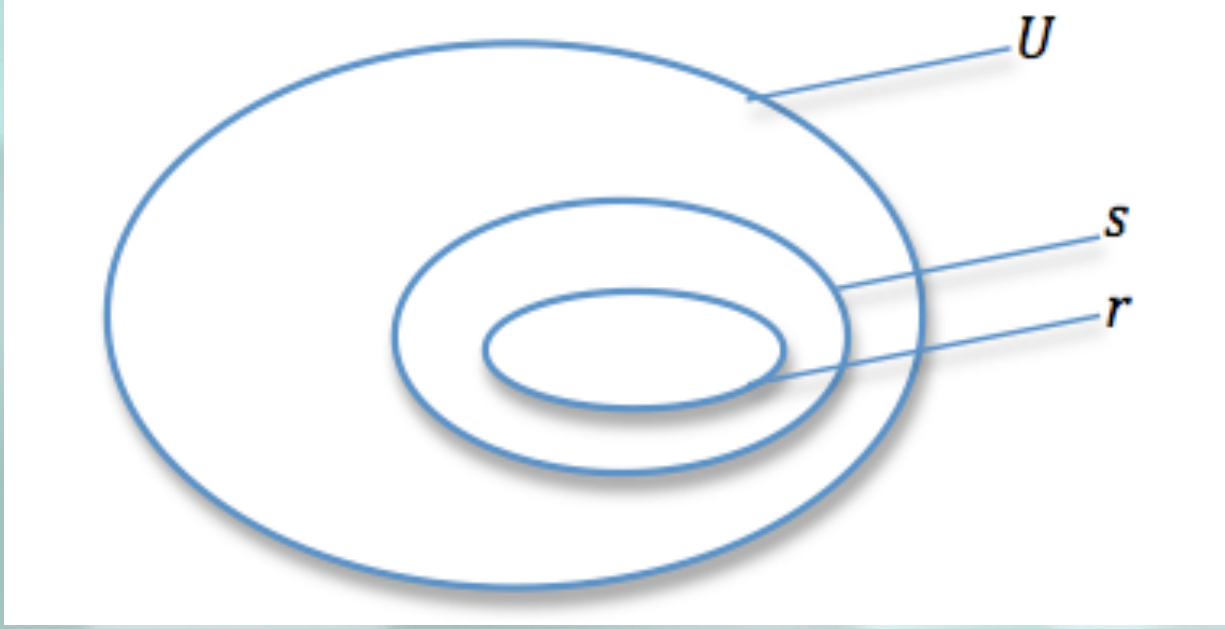
Households contacted and reasons for non-participation

	Panel		Non-panel		Total	
	number	%	number	%	number	%
Respondents	4,345	*79.3*	3,632	*41.6*	7,977	*56.1*
Refusals	1,012	*18.5*	3,589	*41.1*	4,601	*32.4*
Not at home	120	*2.2*	1,511	*17.3*	1,631	*11.5*
Total	5,477	*100.0*	8,732	*100.0*	14,209	*100.0*
Ineligible*	150	*2.7*	629	*6.7*	779	*5.2*

*Households not found at their address (wrong address, death, change of address).

Some notation

- More formally, we wish to estimate the total of a p-dimensional vector of variables of interest $\boldsymbol{y}$ for $U = \{1, \ldots, k, \ldots, N\}$, $\boldsymbol{t}_y = \sum_U \boldsymbol{y}_k$.



- sample s has dimension n; $p(s)$ has first order inclusion probabilities $\pi_k = P(k \in s)$.
- respondents set r has dimension m with $r \subseteq s$ and $m \leqslant n$.

A two phases approach for unit nonresponse

1st phase sampling design $p(s)$;

2nd phase response mechanism $q(r|s)$ with $P(k \in r|s) = \theta_k$.

$$\hat{\boldsymbol{t}}_{y,\text{2p}} = \sum_r \frac{\boldsymbol{y}_k}{\pi_k \theta_k}$$

is unbiased under $p(s)q(r|s)$ but the θ_k's are unknown.

A two phases approach for unit nonresponse

1st phase sampling design $p(s)$;

2nd phase response mechanism $q(r|s)$ with $P(k \in r|s) = \theta_k$.

$$\hat{\boldsymbol{t}}_{y,\text{2p}} = \sum_r \frac{\boldsymbol{y}_k}{\pi_k \theta_k}$$

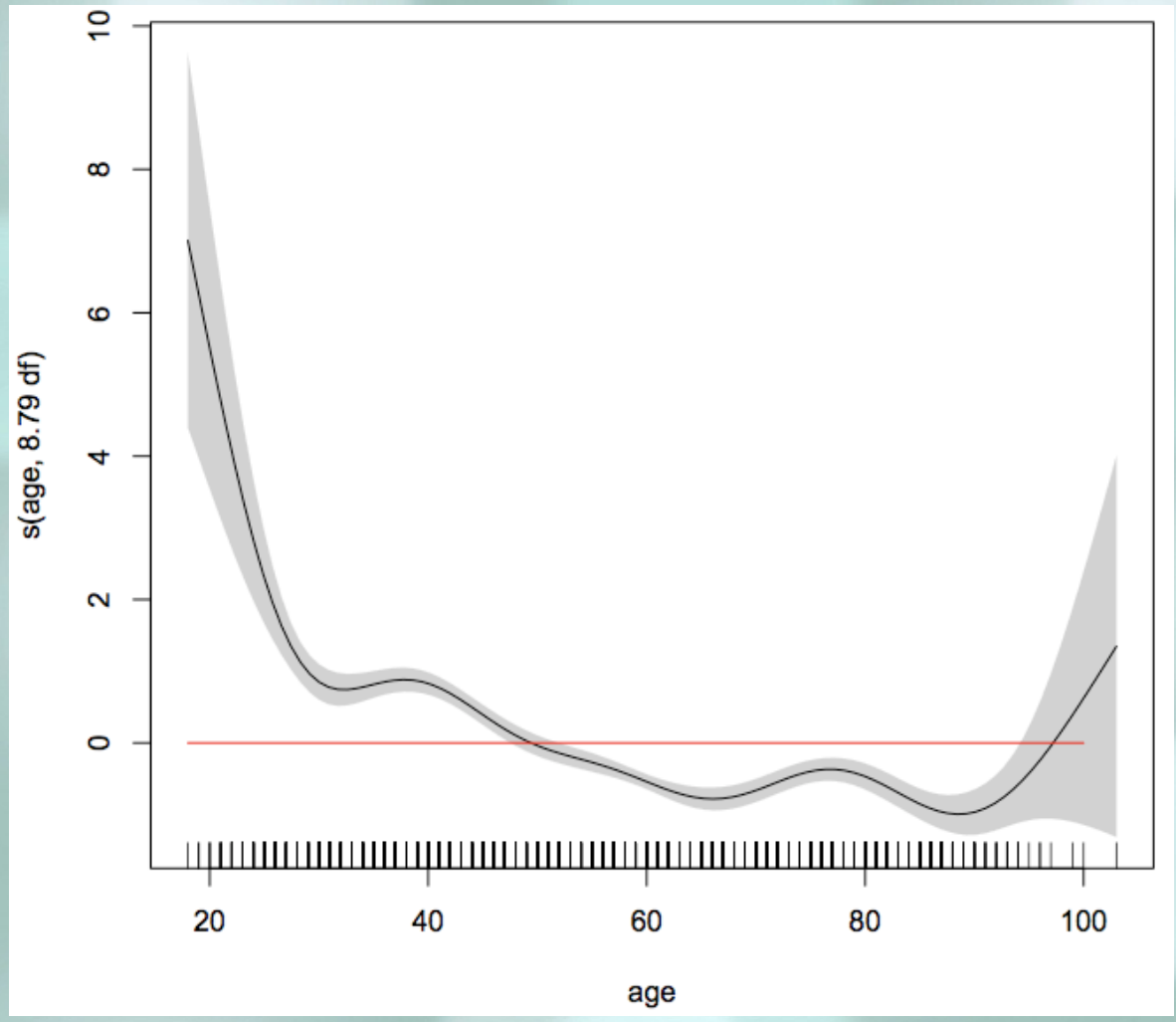
is unbiased under $p(s)q(r|s)$ but the θ_k's are unknown.

- $\theta_{kh} = \theta_h$ for each $k \in h$, $h = 1, \ldots, L \rightarrow$ post-stratification;
- logistic models for θ_k (Little, 1986; Ekholm & Laaksonen, 1991; Kim & Kim, 2007). Different models for
  - Non-panel HHs (from a subsample of nonrespondents CATI for socio-demo info)
  - Panel HHs (from previous waves)

Logistic Additive model for Non-panel HHs

Variables	coef	exp(coef)	p-value
Number of components of the HH*	-0.38	0.68	$<.0001$
Graduated head of HH	-0.48	0.62	$<.0001$
North/Centre	-0.67	0.51	$<.0001$
Self-employed head of HH	-0.60	0.55	$<.0001$
Retired head of HH	-0.35	0.70	0.0010
Home owner head of HH	-0.73	0.48	$<.0001$
Originally selected HH (vs subst.)	-0.24	0.79	$<.0001$
Intercept	1.72	5.58	$<.0001$

* quantitative variables



The estimated effect of AGE of the head of HH (and 95% confidence bounds) on the linear predictor scale

Logistic model for Panel HHs

Variables	coeff	exp(coef)	p-value
Municipalities with more than 500,000 inhabitants	-0.14	0.74	0.002
Number of components of the HH*	0.07	1.09	0.001
Number of waves HH has participated*	0.18	1.21	$<.0001$
Climate of previous interview	0.17	1.20	$<.0001$
Interviewer's age*	0.01	1.01	0.011
Number of waves (interviewer)	0.05	1.20	$<.0001$
Workload of interviewer** 21 – 100	-0.13	0.73	0.020
Workload of interviewer** 101 – 300	-0.34	0.59	$<.0001$
Workload of interviewer** $>$ 300	0.29	1.12	$<.0001$
Intercept	-1.33	0.27	$<.0001$

*quantitative variables; **baseline $<= 20$

Two major issues - 2 - Measurement Error

Two major issues - 2 - Measurement Error

$$\hat{t}_{y,\text{2p}} = \sum_r \frac{y_k}{\pi_k \hat{\theta}_k}$$

Kim & Kim (27) prove design consistency for this estimator if the logistic model is well specified.

Two major issues - 2 - Measurement Error

$$\hat{\boldsymbol{t}}_{y,\text{2p}} = \sum_r \frac{\boldsymbol{y}_k}{\pi_k \hat{\theta}_k}$$

Kim & Kim (27) prove design consistency for this estimator if the logistic model is well specified.

However, respondents likely report $\tilde{\boldsymbol{y}}_k \neq \boldsymbol{y}_k$. Two different types of variable of interest (see Table 2)

- Possession (1/0 variable)
- Amount held (continuous variable)

× Subsample with more accurate measures

✓ Independent sample with more accurate measures

The Extended Supplementary Sample (ESS)

- Sample of 1,681 HHs from customers of a major Bank in Italy (stratified according to geographical area of residence, municipality size, and to the financial wealth held within the Bank)
- Same questionnaire and interviewers as for SHIW
- Survey data *matched* with the Bank database containing the amounts of the assets *actually* held by the individuals selected in the sample

× Experiment run in 2003

- We will use this data to model mis-reporting for profiles of HHs and to extrapolate it to SHIW in two steps
  i. estimation of mis-reporting on the possesion;
  ii. estimation of mis-reporting on the amount held.

i. Mis-reporting models for the possession

For each of 4 financial assets (gov. bonds, private bonds, shares, mutual funds) let

$$\tau_{kj} = \begin{cases} 1 & \text{if unit } k \text{ possesses financial asset } j = 1, \ldots, 4 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\delta_{kj} = \begin{cases} 1 & \text{if unit } k \text{ declares to possess financial asset } j = 1, \ldots, 4 \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\text{logit}\{P(\tau_{kj} = 1)\} = f(\delta_{kj}, \text{HH characteristics})$$

is modeled for each $j = 1, \ldots, 4$.

A brief summary of the models

Probability of possession (most significant variables)

Gov. Bonds ↑ Graduated head of HH; North-Center of Italy; Number of components with income

↓ above 3rd quartile of income

Private Bonds ↑ Graduated head of HH; Age of the head of HH; North-Center of Italy

↓ above 3rd quartile of income; Large municipality

Shares ↑ Graduated head of HH; ↓ Truthful response

Mutual Funds ↑ Age of the head of HH; North-Center of Italy

ii. Mis-reporting models for the amount held

Let

$$r_{kj} = \frac{y_{kj}}{\tilde{y}_{kj}} \quad \text{for } j = 1, ..., 4$$

be the ratio between true and declared amount held for each financial asset. Then $\log r_{kj}$ is modeled as a linear function of the amount declared and HH characteristics.

ii. Mis-reporting models for the amount held

Let

$$r_{kj} = \frac{y_{kj}}{\tilde{y}_{kj}} \quad \text{for } j = 1, ..., 4$$

be the ratio between true and declared amount held for each financial asset. Then $\log r_{kj}$ is modeled as a linear function of the amount declared and HH characteristics.

Gov. Bonds ↑ Age of the head of HH

↓ Amount declared; below 1st quartile of income

Private Bonds ↑ above 3rd quartile of income;

↓ Graduated head of HH; below 1st quartile of income; Truthful response

Shares ↓ Amount declared;

Mutual Funds ↑ Graduated head of HH; Age of the head of HH;

↓ Truthful response

Alternative estimators

SHIW estimate Original design weights $1/\pi_k$ on declared variables $\tilde{\boldsymbol{y}}_k$

NR adjustment Nonresponse adjusted weights $1/\pi_k\hat{\theta}_k$ on declared variables $\tilde{\boldsymbol{y}}_k$

ME adjustment Original design weights $1/\pi_k$ on imputed variables $\hat{\boldsymbol{y}}_k$

NR+ME adjustment Nonresponse adjusted weights $1/\pi_k\hat{\theta}_k$ on imputed variables $\hat{\boldsymbol{y}}_k$

All estimators are calibrated w.r.t. demographic benchmarks (sex, age classes, geographical area, education)

Number of HHs (Millions) Possessing Financial Assets

	Gov. Bonds	*% of NFA*	Priv. Bonds	*% of NFA*	Shares	*% of NFA*	Managed savings	*% of NFA*
SHIW estimate	2.180	*118.2*	1.710	*38.2*	1.450	*50.6*	1.495	*55.0*
NR adj	2.299	*124.6*	1.769	*39.6*	1.502	*52.4*	1.601	*58.9*
ME adj	1.205	*65.3*	3.649	*81.6*	2.364	*82.5*	2.293	*84.4*
NR + ME adj	1.196	*64.8*	3.724	*83.3*	2.461	*85.9*	2.297	*84.6*
NFA estimate	1.845	*100.0*	4.471	*100.0*	2.865	*100.0*	2.716	*100.0*

Amount held (Billions of Euros)

	Gov. Bonds	*% of NFA*	Priv. Bonds	*% of NFA*	Shares	*% of NFA*	Managed savings	*% of NFA*
SHIW estim	65.610	*40.8*	61.190	*20.2*	31.862	*50.6*	50.992	*47.1*
NR adj	64.826	*40.3*	71.589	*23.6*	35.626	*56.5*	56.211	*51.9*
ME adj	69.311	*43.1*	456.435	*150.4*	53.447	*84.8*	109.373	*101.1*
NR+ME adj	66.655	*41.4*	472.852	*155.8*	56.318	*89.4*	115.219	*106.5*
NFA estim	160.840	*100.0*	303.450	*100.0*	63.026	*100.0*	108.219	*100.0*

Alternatively via (model) calibration

A more general framework provided by calibration (Deville & Särndal, 1992), model calibration (Wu & Sitter, 2001; Wu & Luan; 2003) and a combination of the two (Montanari & Ranalli, 2009)

$$\hat{\boldsymbol{t}}_{yc} = \sum_r w_k \tilde{\boldsymbol{y}}_k$$

Weights w_k are found to be as close as possible – in terms of a distance measure $\Phi(\cdot, \cdot)$ – to the NR adjusted weights $d_k = 1/\pi_k \hat{\theta}_k$ and, at the same time, to satisfy benchmark (coherence and/or efficiency) constraints:

$$\min_{w_k} \sum_r \Phi(w_k, d_k) \qquad \text{s.t.}$$

1. Vector of demographic variables: $\sum_r w_k \boldsymbol{x}_k = \boldsymbol{t}_x$
2. Vector of measurement error imputed variables $\sum_r w_k \tilde{\boldsymbol{y}}_k = \sum_r d_k \hat{\boldsymbol{y}}_k$

Number of HHs (Millions) Possessing Financial Assets again

	Gov. Bonds	*% of NFA*	Priv. Bonds	*% of NFA*	Shares	*% of NFA*	Managed savings	*% of NFA*
SHIW estimate	2.180	*118.2*	1.710	*38.2*	1.450	*50.6*	1.495	*55.0*
NR adj	2.299	*124.6*	1.769	*39.6*	1.502	*52.4*	1.601	*58.9*
ME adj	1.205	*65.3*	3.649	*81.6*	2.364	*82.5*	2.293	*84.4*
NR + ME adj	1.196	*64.8*	3.724	*83.3*	2.461	*85.9*	2.297	*84.6*
CAL adj	1.247	*67.6*	3.888	*87.0*	2.554	*89.1*	2.364	*87.0*
NFA estimate	1.845	*100.0*	4.471	*100.0*	2.865	*100.0*	2.716	*100.0*

Amount held (Billions of Euros) again

	Gov. Bonds	*% of NFA*	Priv. Bonds	*% of NFA*	Shares	*% of NFA*	Managed savings	*% of NFA*
SHIW estim	65.610	*40.8*	61.190	*20.2*	31.862	*50.6*	50.992	*47.1*
NR adj	64.826	*40.3*	71.589	*23.6*	35.626	*56.5*	56.211	*51.9*
ME adj	69.311	*43.1*	456.435	*150.4*	53.447	*84.8*	109.373	*101.1*
NR+ME adj	66.655	*41.4*	472.852	*155.8*	56.318	*89.4*	115.219	*106.5*
CAL adj	71.218	*44.3*	501.661	*165.3*	59.284	*94.1*	118.627	*109.6*
NFA estim	160.840	*100.0*	303.450	*100.0*	63.026	*100.0*	108.219	*100.0*

Conclusions and points for discussion

- ME seems to have a larger impact than NR
  - × Self selection issues in the subsample of nonrespondents
  - × Non contacts/refusal confounding
  - × Portfolio issues from time mis-alignement (2003 ESS; 2008 SHIW)
  - × Correlation between response probabilities and misclassification probabilities

Conclusions and points for discussion

- ME seems to have a larger impact than NR
  - × Self selection issues in the subsample of nonrespondents
  - × Non contacts/refusal confounding
  - × Portfolio issues from time mis-alignement (2003 ESS; 2008 SHIW)
  - × Correlation between response probabilities and misclassification probabilities
- (Model) Calibration seems a natural environment to handle NR and ME adjustments

	Design weights $1/\pi_k$	NR adjusted weights $1/\pi_k\hat{\theta}_k$	CAL weights w_k
% CV	96.6	104.6	187.0
Q3 - Q1	2548.9	2703.7	2765.1

  - × Extreme weights due to the unrepresentativeness of the wealthiest HHs

THANK YOU!

QUESTIONS, COMMENTS, SUGGESTIONS... ??

Essential Bibliography

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Ekholm, A. and S. Laaksonen (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics 7*, 325–337.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review 54*, 139–157.

Kim, J. K. and J. J. Kim (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 35*(4), 501–514.

Montanari G.E. and M.G. Ranalli (2009). Multiple and ridge model calibration, Proceedings of WCES, Workshop on Calibration and Estimation in Surveys, Statistics Canada, Ottawa.

Wu, C. and Y. Luan, (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, *19*, 119–131.

Wu, C. and R. R. Sitter (2001). A model-calibration to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.