The Use of Unit-level Accuracy Indicators

Chris Skinner

London School of Economics and Political Science

with Damião da Silva (University of Southampton)
and Jae-Kwang Kim (Iowa State University)

with thanks to ESRC for Professorial Fellowship

5th ESRC Research Methods Festival, Oxford, 2-5 July 2012

Outline

- accuracy indicators
- models
- identification
- estimation
- application to earnings
- simulation study

Example: English Longitudinal Study of Ageing

How accurate do you think the answers given by the respondent to questions about pay were?

1. Very accurate
2. Fairly accurate
3. Not very accurate
4. Not at all accurate

Examples of Accuracy Indicators in the Literature

- Mathiowetz (1998, *Public Opinion Quarterly*) considers $a_i =$ respondent expression of uncertainty, continuum from no uncertainty to item nonresponse,
  uses a_i in definition of imputation classes.

- Battistin, Miniaci and Weber (2003, *J. Human Resources*) consider heaping of household expenditure where $a_i =$ interviewer's assessment of respondent's understanding of question (fair, good, excellent), interview length.

- Kreider and Pepper (2007, *J. Amer. Statist. Ass.*) consider $y_i =$ disability status, $a_i =$ latent binary accuracy variable.

General Trends

'new survey data quality evaluation techniques have provided more information regarding the validity and reliability of survey results than was previously thought possible' (Biemer and Lyberg, 2003, *Introduction to Survey Quality*)

'unprecedented information about the data collection process' (Groves and Heeringa, 2006, *J. Roy. Statist. Soc. A*)

paradata associated with survey data collection process

Indicators of Measurement Accuracy

$y_i^* =$ measured variable

$y_i =$ true variable

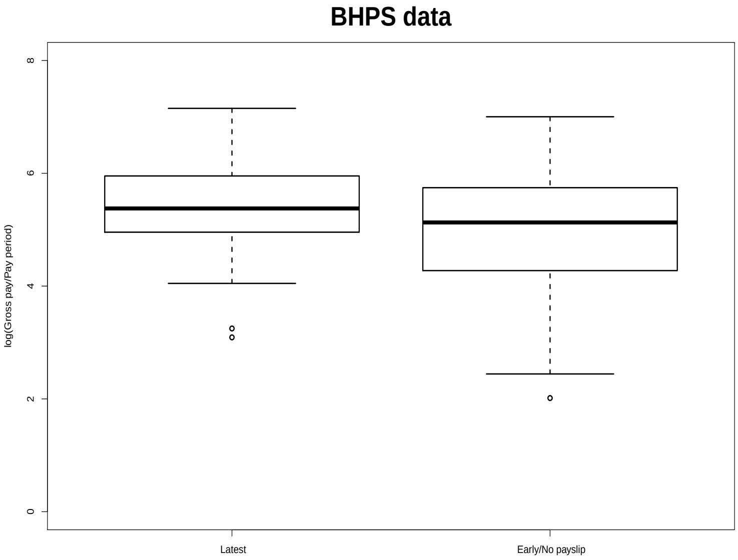
$y_i^* - y_i =$ **measurement error**

$a_i =$ accuracy indicator, associated with magnitude of measurement error

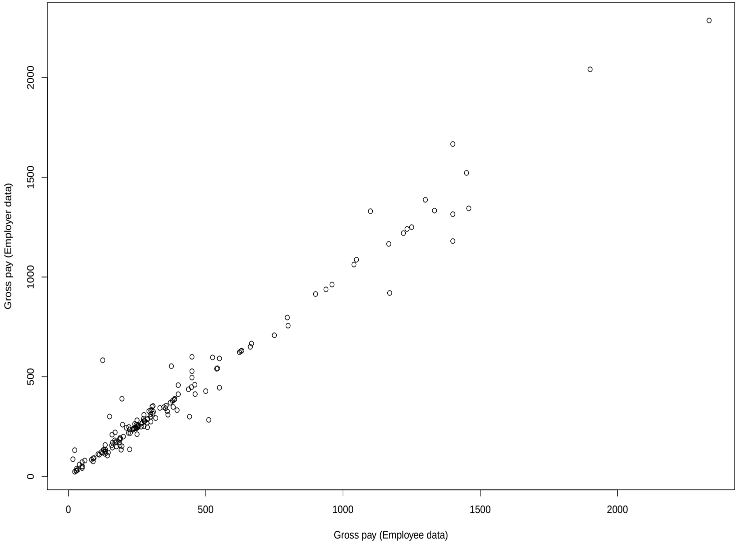
Example: British Household Panel Survey

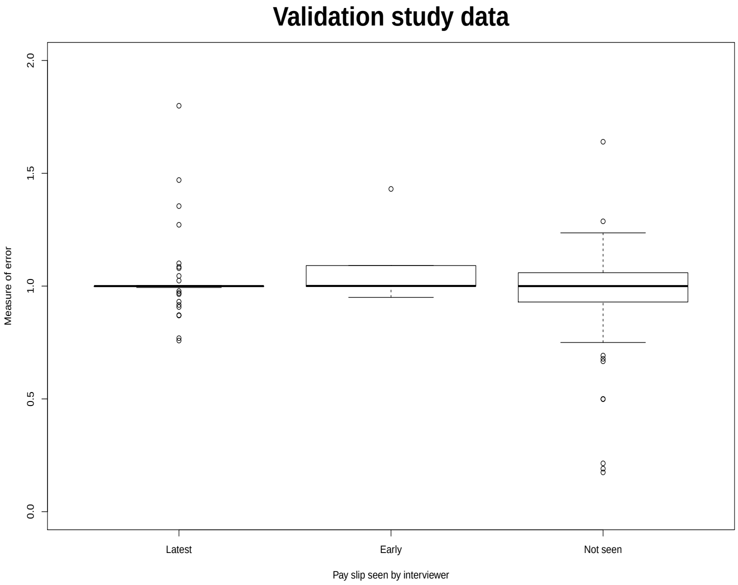
Pay slip seen by interviewer:

- Latest payslip seen
- Early payslip seen
- No payslip seen


BHPS data
log(Gross pay/Pay period)
0
2
4
6
8
Latest
Early/No payslip

Validation study data




Validation study data
Measure of error
0.0
0.5
1.0
1.5
2.0
Latest
Early
Not seen
Pay slip seen by interviewer

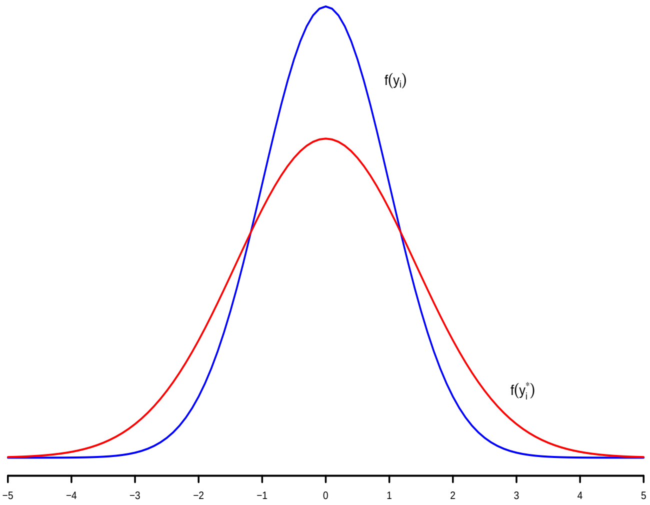
Bias Impact of Measurement Error

y_i^* = measured variable, unit i

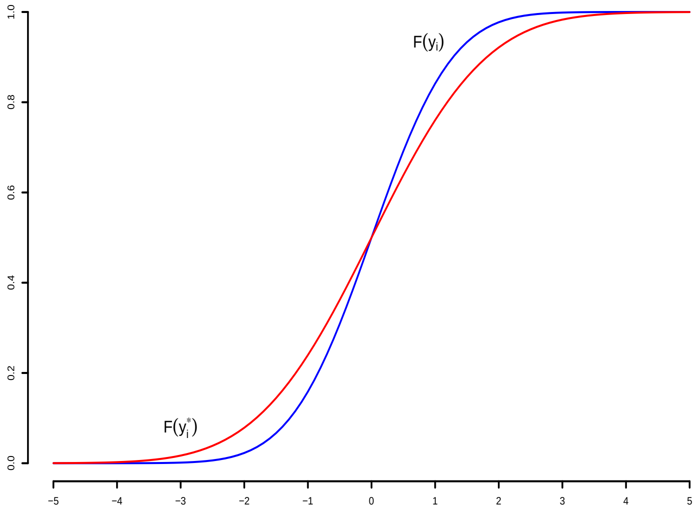
y_i = true variable

Classical measurement error model

$y_i^* = y_i + \epsilon_i$, $E(\epsilon_i) = 0$, $var(\epsilon_i) = \sigma^2$



$$y_i^* = y_i + \epsilon_i, \qquad y_i \sim \mathrm{N}(0,1), \qquad \epsilon_i \sim \mathrm{N}(0,1)$$



$$y_i^* = y_i + \epsilon_i, \qquad y_i \sim N(0,1), \qquad \epsilon_i \sim N(0,1)$$

Problem

Can we use accuracy indicators to correct for bias due to measurement error?

Existing Methods for Measurement Error Bias Adjustment

- methods which employ error characteristics of measurement instrument obtained from validation study
- latent variable modelling employing multiple indicators
- instrumental variable estimation

Binary Accuracy Indicator - Basic Model

$$y_i^* = \begin{cases} y_i + \epsilon_i & \text{if } a_i = 1 \\ \\ y_i & \text{if } a_i = 0, \end{cases}$$

Extended Model

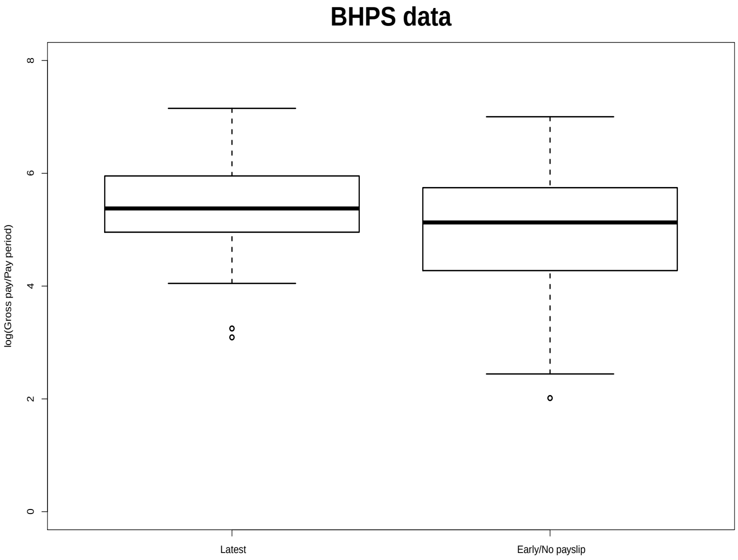
$$a_i^* \quad = \quad \begin{cases} 1 \quad \Rightarrow \quad a_i = 1 \quad \Rightarrow \quad y_i^* = y_i + \epsilon_i \\ \\ 0 \quad \Rightarrow \quad a_i = \begin{cases} 1 \text{ (with probability } p) \quad \Rightarrow \quad y_i^* = y_i + \epsilon_i \\ \\ 0 \text{ (with probability } 1 - p) \quad \Rightarrow \quad y_i^* = y_i \end{cases} \end{cases}$$

Identification Challenge

Observe dependence of y_i^* on a_i.

How to distinguish betweeen

- $y_i^* \mid y_i$ (measurement error) depends on a_i
- y_i depends on a_i (with possibly no measurement error)


BHPS data
log(Gross pay/Pay period)
0
2
4
6
8
Latest
Early/No payslip

Identifying Assumption

Observe covariate vector $\mathbf{x_i}$

Assume: a_i and y_i conditionally independent given $\mathbf{x_i}$

Parametric Modelling Assumptions

- $y_i \mid \mathbf{x}_i \sim f(y_i \mid \mathbf{x}_i; \gamma)$

- $y_i^* \mid \mathbf{x}_i, y_i, a_i = 1 \sim g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1; \eta)$

- $\psi = (\gamma, \eta)$

- treat p as known

Estimation of Finite Population Distribution Function

target of inference: $\theta_c = N^{-1} \sum_{i \in U} I(y_i < c)$

direct estimator: $\widehat{\theta}_c = (\sum_{i \in s} w_i)^{-1} \sum_{i \in s} w_i I(y_i^* < c)$

adjusted estimator:
$\widehat{\theta}_c^* = (\sum_{i \in s} w_i)^{-1} \sum_{i \in s} w_i \widehat{E}_m[I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i]$

Estimation of $E_m[I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i]$

pseudo MLE: obtain $\widehat{\psi}$ by solving survey weighted score equations, if in closed form, and use $E_m[I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i; \widehat{\psi}]$.

fractional imputation: estimate ψ by cycling between imputation of y_i from $f[y_i \mid \mathbf{x}_i, y_i^*, a_i; \widehat{\psi}^{(t)}]$
and maximizing likelihood including imputed data to obtain $\widehat{\psi}^{(t+1)}$.
Estimate $E_m[I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i]$ using imputed data.

Pseudo MLE

Assume $$\begin{aligned} y_i \mid \mathbf{x}_i &\sim \mathrm{N}(\mathbf{x}_i^\top \beta, \sigma^2) \\ y_i^* \mid \mathbf{x}_i, y_i, a_i = 1 &\sim \mathrm{N}(y_i, \tau^2) \end{aligned}$$

Then

$$y_i \mid \mathbf{x}_i, y_i^*, a_i = 1, \sim \mathrm{N}\big((1-\rho)\mathbf{x}_i^\top \beta + \rho y_i^*,\ \sigma^2(1-\rho)\big),$$

where $\rho = \sigma^2/(\sigma^2 + \tau^2)$, etc.

Construct weighted score equations.

Use linearization for variance estimation.

Fractional Imputation

$$
\begin{aligned}
f(y_i \mid \mathbf{x}_i, y_i^*, a_i = 1) &= \frac{f(y_i \mid \mathbf{x}_i, a_i = 1) g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1)}{\int f(y_i \mid \mathbf{x}_i, a_i = 1) g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1) dy_i} \\
&= \frac{f(y_i \mid \mathbf{x}_i; \gamma) g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1; \eta)}{\int f(y_i \mid \mathbf{x}_i, a_i = 0; \gamma) f(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1; \eta) dy_i}
\end{aligned}
$$

Fractional Imputation + EM Algorithm

Step 1. Obtain initial estimate $(\widehat{\gamma}^{(0)}, \widehat{\eta}^{0)})$

Step 2. For $a_i = 1$, generate $y_{il}^{(1)}, \cdots, y_{il}^{(M)}$, from $f(y_i \mid \mathbf{x}_i; \widehat{\gamma}^{(t)})$.

Step 3. For $a_i = 1$, compute fractional weights

$$w^*_{ij(t)} = \frac{g(y_i^* \mid \mathbf{x}_i, y_{il}^{(j)}, a_i = 1; \widehat{\eta}^{(t)})}{\sum_{k=1}^{M} g(y_i^* \mid \mathbf{x}_i, y_{il}^{(k)}, a_i = 1; \widehat{\eta}^{(t)})}$$

Step 4. Update parameter estimates $(\widehat{\gamma}^{(t+1)}, \widehat{\eta}^{(t+1)})$ by solving the weighted complete sample score equations with imputed data.

Kim (2011, *Biometrika*)

Application: British Household Panel Survey

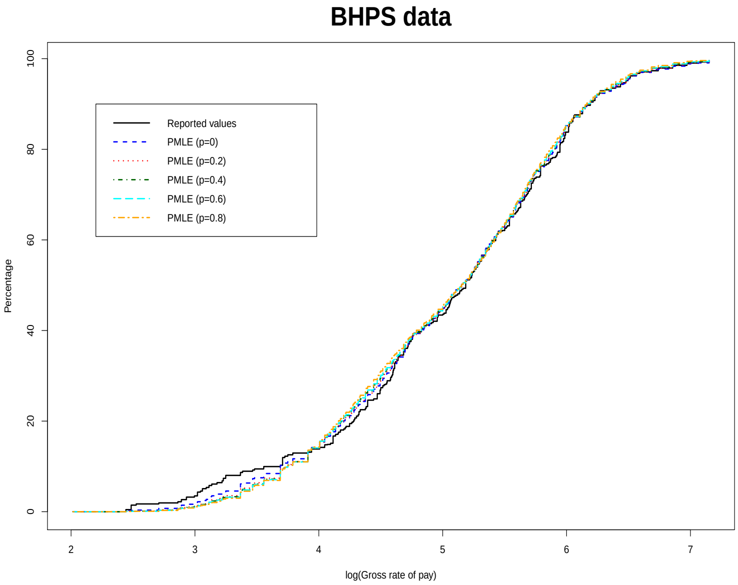
Wave 12 to correspond to ISMIE validation study

$y_i =$ gross weekly pay, aim to estimate distribution function

$$a_i = 0 \quad \text{if latest pay slip seen}$$
$$= 1 \quad \text{if not}$$

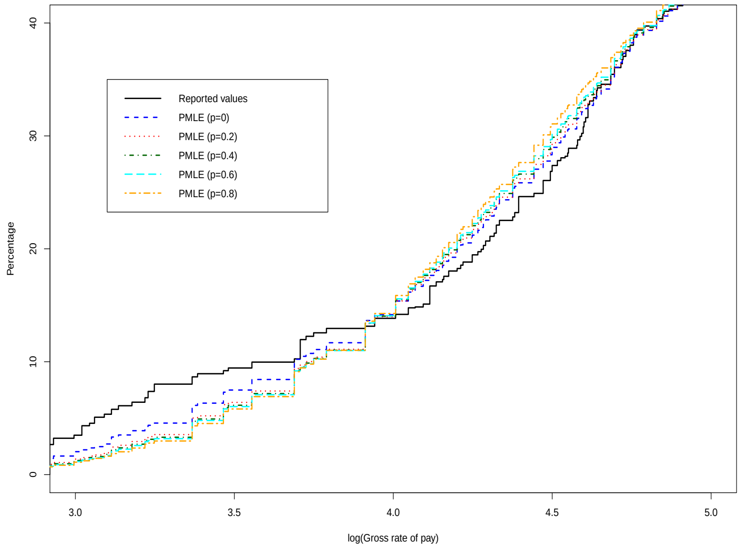
$\mathbf{x}_i$ includes hours worked, part-time status, qualifications, occupation, workplace size, region, sex, age, household position, household size, housing tenure, marital status

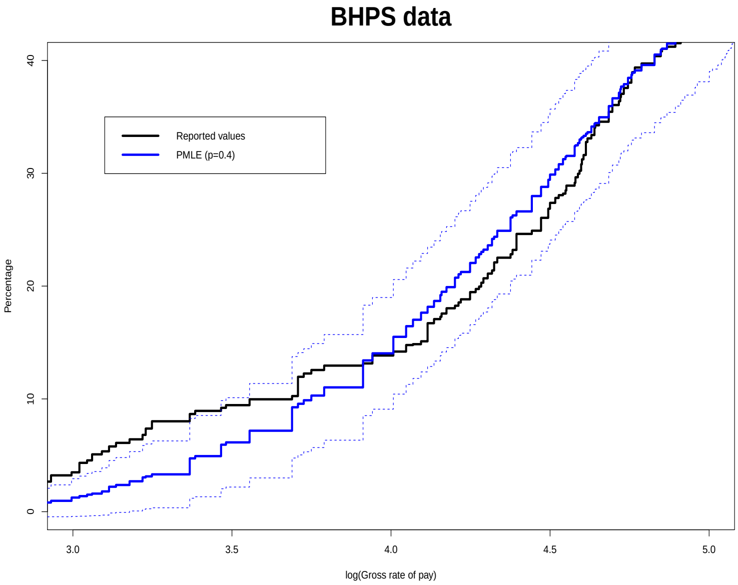
separate models for pay period = 1 week, 2-4 weeks, 1 month +



BHPS data
Reported values
PMLE (p=0)
PMLE (p=0.2)
PMLE (p=0.4)
PMLE (p=0.6)
PMLE (p=0.8)
Percentage
0
20
40
60
80
100
2
3
4
5
6
7
log(Gross rate of pay)

BHPS data


Reported values
PMLE (p=0)
PMLE (p=0.2)
PMLE (p=0.4)
PMLE (p=0.6)
PMLE (p=0.8)
Percentage
0
10
20
30
40
3.0
3.5
4.0
4.5
5.0
log(Gross rate of pay)


BHPS data
Reported values
PMLE (p=0.4)
Percentage
0
10
20
30
40
3.0
3.5
4.0
4.5
5.0
log(Gross rate of pay)

Simulation Comparison of Pseudo MLE and Fractional Imputation

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $y_i^* \sim N(y_i, \tau^2)$, $x_i \sim U(0, 1)$,
$a_i \sim Bin(1, \pi_i)$, $logit(\pi_i) = \delta_0 + \delta_1 x_i$

$n = 300$

$M = 20$ imputations for fractional imputation

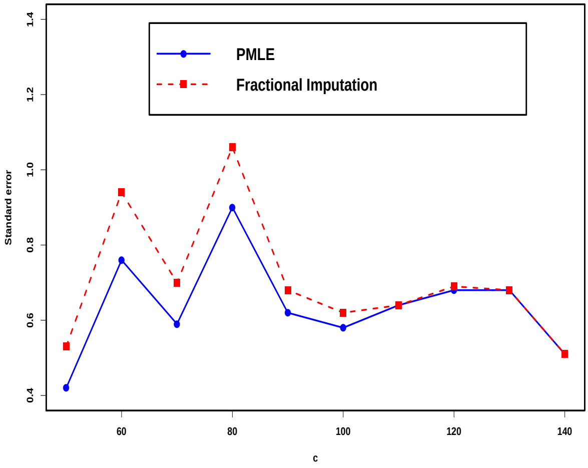
Relative Root MSE (%) of parameter estimators ($p = 0$)

Parameter	PMLE	Fractional Imputation
β_0	1.8	2.1
β_1	1.2	1.4
σ^2	11.8	11.9
τ^2	10.9	10.9

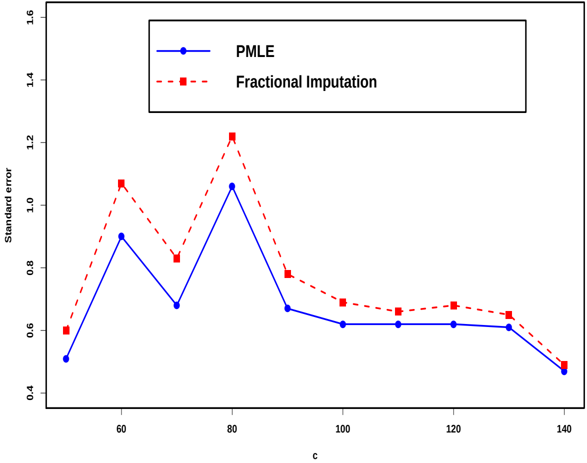
Relative Root MSE (%) of parameter estimators ($p = 0.2$)

Parameter	PMLE	Fractional Imputation
β_0	2.1	2.4
β_1	1.4	1.6
σ^2	18.5	35.2
τ^2	10.6	10.8

Standard errors of cdf estimators ($p = 0$)



Standard errors of cdf estimators ($p = 0.2$)



Further Research

Explore implementation of fractional imputation for alternative models