Some methodological issues in the implementation of subjective poverty definitions
Kapteyn, A.J.; Kooreman, P.; Willemse, R.

Publication date:
1990

Citation for published version (APA):
Some Methodological Issues in the Implementation of Subjective Poverty Definitions

by

Arie Kapteyn, Peter Kooreman and Rob Willemse

Reprinted from The Journal of Human Resources Vol. 23, No. 2, 1988
Scientific Council

Eduard Bomhoff
Willem Buiter
Jacques Drèze
Theo van de Klundert
Simon Kuipers
Jean-Jacques Laffont
Merton Miller
Stephen Nickell
Pieter Ruys
Jacques Sijben

Erasmus University Rotterdam
Yale University
Université Catholique de Louvain
Tilburg University
Groningen University
Université des Sciences Sociales de Toulouse
University of Chicago
Tilburg University
Tilburg University

Board

Anton P. Barten, director
Eric van Damme
John Driffill
Arie Kepteijn
Frederick van der Ploeg

Research Staff and Resident Fellows

Anton P. Barten
Eric van Damme
John Driffill
Arie Kapteijn
Hugo Keuzenkamp
Pieter Kop Jansen
Jan Magnus
Frederick van der Ploeg

Address: Hogeschoollaan 225, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
Phone: +3113663050
Telex: 52426 kub nl
Telefax: +3113663066
E-mail: "CENTER@TICKUB5.BITNET"
Some Methodological Issues in the Implementation of Subjective Poverty Definitions

by

Arie Kapteyn, Peter Kooreman and Rob Willemse

Reprinted from The Journal of Human Resources Vol. 23, No. 2, 1988

Reprint Series no. 19
Some Methodological Issues in the Implementation of Subjective Poverty Definitions

Arie Kapteyn
Peter Kooreman
Rob Willemse

ABSTRACT
This paper discusses an investigation of the effects of systematic underreporting of income and of sample selectivity on the estimated levels of two subjective definitions of poverty: the so-called subjective poverty line and the Leyden poverty line. Both turn out to have substantially biasing effects. We present methods to remedy the biases. The resulting adjusted poverty lines prove to be quite accurate. Furthermore, we make suggestions for the design of questionnaires that are used in the surveys on which these poverty definitions are based.

I. Introduction

Whatever poverty definition one adheres to, a proper implementation is prerequisite for its possible use in social policy. In this paper we pay attention to a number of methodological issues that arise in the empirical implementation of certain subjective definitions of poverty. In particular, we are concerned with the so-called subjective poverty line (SPL) and Leyden poverty line (LPL), which have been introduced and discussed in a string of papers by, mainly, Dutch and American authors (see references). Both approaches are subjective in that they are based on responses to survey questions which try to elicit either a respondent’s evaluation of income levels or his judgment about minimum needs. Furthermore, both approaches are model based, in the sense that the re-

Kapteyn is a professor of economics at Tilburg University; Kooreman is a professor of economics at Tilburg as well. Willemse is a researcher for the Federal Reserve Bank of Richmond, Virginia.

[Submitted January 1987; accepted July 1987]
responses themselves do not generate poverty lines immediately. One needs to estimate a model that explains interhousehold variation in the responses to the survey questions. These two aspects identify two crucial methodological issues in the implementation of the SPL and LPL: the responses should measure what they are supposed to measure and the model should be correctly specified and estimated.

In this paper we present some empirical evidence on the sensitivity of the poverty line definitions to systematic biases in the responses and to an incorrect estimation method for the model. The systematic response bias is a result of the rather general tendency of respondents to underestimate their own after tax household income. The incorrect estimation method stems from a disregard of sample selectivity due to partial nonresponse.

In Section II we briefly explain the SPL and LPL and we discuss and estimate the model that is assumed to generate the responses. We point out some implausible outcomes.

In Section III, we present evidence that respondents severely underestimate their after tax household income, which results in a downward bias of the poverty lines. We also present a method to correct this bias. In Section IV we take up the problem of selectivity bias. It turns out that selectivity is a statistically significant problem. In Section V we compare the poverty lines that result from the various models. Both the income correction and the correction for sample selectivity bias lead to sizable changes in the estimated level of the poverty line.

In Section VI we draw conclusions from our findings for the design of questionnaires used in the empirical work that underlies the subjective poverty line definitions considered here. All in all, the problems considered can be remedied rather easily.

II. Two Subjective Definitions of Poverty and Their Implementation

The so-called subjective poverty line (SPL) is based on the following survey question, posed to the head of the household:

Which after tax monthly income do you, in your circumstances, consider to be absolutely minimal? That is to say that with less you could not make ends meet.

Absolutely minimal per month $\ldots$

A respondent's answer to this minimum income question (MIQ) will be referred to as his minimum income $y_{\text{min}}$. It turns out that $y_{\text{min}}$ depends on
the respondent's actual after tax income and a number of other factors, including family composition. See, for instance, Kapteyn, Van de Geer, Van de Stadt (1985) for details. In formula:

\[(2.1) \quad y_{\text{min}} = f(y;x),\]

where \(y\) is the respondent's actual income\(^1\) and \(x\) is a vector of other factors. The function \(f\) is monotonically increasing in \(y\) and there exists an income level \(y_{\text{min}}^*\) defined by

\[(2.2) \quad y_{\text{min}}^* = f(y_{\text{min}}^*;x),\]

such that for all incomes \(y\) less than \(y_{\text{min}}^*\) we have that \(y < y_{\text{min}}\) and for all incomes \(y\) greater than \(y_{\text{min}}^*\), \(y > y_{\text{min}}\). See Figure 2.1. The income level \(y_{\text{min}}^*\) is the SPL, as it is the point where families can just make ends meet; with less income they cannot make ends meet and with more income than \(y_{\text{min}}^*\) they can. Since the position of the function \(f(y;x)\) depends on \(x\), it is clear from (2.2), and from Figure 2.1 that the SPL depends on \(x\). That is, if families have different characteristics they will require different amounts of money to make ends meet.

\(^1\) In this paper "income" is always after tax household income.
The Leyden Poverty Line

The *Leyden poverty line* (LPL) is based on the so-called *income evaluation question* (IEQ):

Which after tax monthly income would you, in your circumstances, consider to be very bad? And bad? Insufficient? Sufficient? Good? Very good?

(we mean after tax *household* income)

<table>
<thead>
<tr>
<th>Label</th>
<th>$</th>
</tr>
</thead>
<tbody>
<tr>
<td>very bad</td>
<td>$100</td>
</tr>
<tr>
<td>bad</td>
<td>$100</td>
</tr>
<tr>
<td>insufficient</td>
<td>$100</td>
</tr>
<tr>
<td>sufficient</td>
<td>$100</td>
</tr>
<tr>
<td>good</td>
<td>$100</td>
</tr>
<tr>
<td>very good</td>
<td>$100</td>
</tr>
</tbody>
</table>

A respondent's answers to the IEQ are used to estimate his *welfare function of income* (WFI). The measurement method is illustrated in Figure 2.2.

The verbal labels "very good," "good," etc. have been identified with midpoints of the six equal intervals on a zero-one scale. In this way the verbal evaluations have been transformed into numerical evaluations.
The response to the IEQ can now be represented by a scatter of six points. According to Van Praag (1968) the relation between an income level $z$ and its numerical evaluation on a zero-one scale, $U(z)$, can be approximated quite well by a lognormal distribution function:

$$ U(z) = \Lambda(z; \mu, \sigma), $$

where $\Lambda(\cdot; \mu, \sigma)$ is the lognormal distribution function with median $e^\mu$ and log-variance $\sigma^2$. The parameters $\mu$ and $\sigma$ of a respondent are estimated by fitting a lognormal function through the scatter of points in $(z, U(z))$-space, as illustrated in Figure 2.2. Van Praag and Kapteyn (1973) and Van Herwaarden and Kapteyn (1981) provide further details.

The estimated lognormal distribution function represents the respondent's WFI. The quantity $e^\mu$ is a location parameter, it is the income level which is evaluated at 0.5 by the individual; $\sigma$ is a slope parameter. The evaluation of an income $z$, by an individual with welfare parameters $\mu$ and $\sigma$ is given by

$$ U(z) = \Lambda(z; \mu, \sigma) = N(lnz; \mu, \sigma) = N(\frac{lnz - \mu}{\sigma}; 0, 1). $$

where $N(\cdot; \mu, \sigma)$ is the normal distribution function with mean $\mu$ and variance $\sigma^2$.

The LPL is based on the notion that poverty is a state of low utility. If the WFI is taken as a cardinal utility function of income, someone is defined as poor, if his income $y$ is such that

$$ U(y) \leq \alpha, $$

where $\alpha$ is a "welfare level" (a number between zero and one), which has to be set by politicians. Let us define $u_\alpha$ by

$$ \Lambda(u_\alpha; 0, 1) = N(lnu_\alpha; 0, 1) = \alpha, $$

then according to the LPL someone is poor if his income $y$ satisfies

$$ \frac{lny - \mu}{\sigma} \leq u_\alpha, $$

or

$$ y \leq \exp(\mu + \sigma \cdot u_\alpha). $$

It turns out that the parameters $\mu$ and $\sigma$ both depend on a vector of family characteristics $x$. Thus, just like the SPL, the LPL also depends on $x$. 

(For details and justification see Van Praag 1971, Van Praag and Kapteyn 1973, Kapteyn 1977.)
A. The Model

To make the SPL and LPL operational we have to specify the relation between \( y_{\text{min}} \), \( \mu \) and \( \sigma \) on the one hand and the vector of household characteristics \( x \) on the other hand. The explanation of \( \mu \) and \( \sigma \) is derived from a "theory of preference formation" developed in Kapteyn (1977); see, for example, Van de Stadt, Kapteyn, Van de Geer (1985) (SKG hereafter) and Kapteyn, Van de Geer, Van de Stadt (1985) (KGS hereafter) for details. Estimation of the complete model requires panel data. Only cross section data are available, however, for our analysis. Therefore all lagged explanatory variables in the model are ignored. The resulting specification for \( \mu \) is

\[
(2.9) \quad \mu_n = \beta_0 + \beta_1 (1 - \beta_2) \ln f s_n + \beta_2 \ln y_n + \beta_3 (m_n - \beta_1 h s_n) + \epsilon_n,
\]

where \( \mu_n \) is the value of \( \mu \) for family \( n \), \( f s_n \) is the size of family \( n \), \( y_n \) is its after tax income, \( m_n \) is mean log-income in the reference group of household \( n \), \( h s_n \) is mean log-family size in the reference group of household \( n \), \( \epsilon_n \) is an error term capturing all omitted factors.\(^2\) Since \( \mu_n \) is an indicator of the level of financial wants of a family, Equation (2.9) says that a family's financial wants are determined by its income, family size and by the geometric mean of incomes in the reference group, adjusted for the geometric mean of family sizes in the reference group.

The theory of preference formation mentioned above implies that \( \sigma_n \) is determined by the dispersion of incomes and family sizes in family \( n \)'s reference group, both present and past, and by the variability of past incomes of family \( n \). Although this dependency raises various interesting policy issues (see KGS) we will ignore it here for simplicity's sake. Empirically, this amounts to taking the dispersion of incomes and family sizes as given and dealing with the observed variation of \( \sigma \) across families as being determined exogenously. Furthermore, as \( \sigma \) appears to be uncorrelated with the explanatory variables on the right hand side of (2.9), we take \( \sigma \) as exogenous and for the purpose of constructing a poverty line according to (2.8) we set \( \sigma \) equal to its sample mean \( \bar{\sigma} \).

The explanation of the logarithm of \( y_{\text{min}} \) is based on a similar specification as the explanation of \( \mu \):

\[
(2.10) \quad \ln y_{\text{min},n} = \alpha_0 + \alpha_1 (1 - \alpha_2) \ln f s_n + \alpha_2 \ln y_n + \alpha_3 (m_n - \alpha_1 h s_n) + u_n,
\]

where \( u_n \) is an error term, possibly correlated with \( \epsilon_n \) in (2.9).

The specification of the influence of family size so far is primitive for

---

2. Equation (2.9) can be obtained from (22) in KGS by omitting all lagged variables on the right side.
two reasons. First, it is restrictive to simply count the number of family members without regard for their ages. Therefore, \( \ln fs_n \) is redefined as follows:

\[
(2.11) \quad \ln fs_n := \sum_{j=1}^{I_n} \omega_j f(a_j),
\]

where \( I_n \) is the number of persons in family \( n \); \( \omega_1 = 1 \) and

\[
(2.12) \quad \omega_j := \ln(j/(j - 1)), \quad j = 2, \ldots, I_n;
\]

\[
(2.13) \begin{cases} 
  f(a_j) := 1, & a_j > 18 \\
  f(a_j) := 1 + \gamma_2(18 - a_j)^2 + \gamma_3(18 - a_j)^2(36 + a_j) & 0 \leq a_j \leq 18,
\end{cases}
\]

so that the age function is a third degree polynomial for \( a_j \leq 18 \), with \( f(18) = 1, f'(18) = 0 \); \( \gamma_2 \) and \( \gamma_3 \) are parameters, which have to be estimated. Thus, the logarithmic weighting of family members has been retained but in addition children under 18 are also weighted on the basis of their age. Second, both (2.9) and (2.10) specify the cost of an increase in family size as a fixed percentage of income irrespective of the level of this income (see below). For example, if a baby costs 15 percent of household income at a very low income level, then (2.9) and (2.10) imply that it will also cost 15 percent of household income at a much higher income level. A simple way to relax this restriction is to replace \( \beta_1 \) by \( \beta_1 + \delta \ln y_n \) and \( \alpha_1 \) by \( \alpha_1 + \psi \ln y_n \), as will become clear shortly. If the same adjustment is carried through for the family size in the reference group, this yields

\[
(2.14) \quad \mu_n = \beta_0 + \beta_1 (1 - \beta_2) \ln fs_n + \delta(1 - \beta_2) \ln fs_n \ln y_n + \beta_2 \ln y_n + \beta_3 m_n - \beta_3 \beta_1 h s_n - \beta_3 \delta m_n h s_n + \epsilon_n,
\]

with \( \ln fs_n \) defined by (2.11). The expression for \( \ln y_{min,n} \) is analogous:

\[
(2.15) \quad \ln y_{min,n} = \alpha_0 + \alpha_1 (1 - \alpha_2) \ln fs_n + \psi(1 - \alpha_2) \ln fs_n \ln y_n + \alpha_2 \ln y_n + \alpha_3 m_n - \alpha_3 \alpha_1 h s_n - \alpha_3 \psi m_n h s_n + u_n
\]

Once the parameters in (2.14) and (2.15) are known, we have our operational version of the LPL and the SPL. With respect to the LPL we note that the poverty line corresponding to a welfare level \( \alpha \) is given by (substitute (2.14) in (2.7)):

\[
(2.16) \quad \text{LPL}(\alpha) = \exp \frac{\beta_0 + \beta_1 (1 - \beta_2) \ln fs_n + \beta_3 m_n - \beta_3 \delta m_n h s_n - \beta_3 \beta_1 h s_n + \sigma u_\alpha}{(1 - \beta_2)(1 - \delta \ln fs_n)}.
\]
The SPL is given by (substitute (2.15) in (2.2)):

\[
(2.17) \quad \text{SPL} = \exp \frac{\alpha_0 + \alpha_1(1 - \alpha_2)\ln f s_n + \alpha_3 m_n - \alpha_3 \psi m_n h s_n - \alpha_3 \alpha_1 h s_n}{(1 - \alpha_2)(1 - \psi \ln f s_n)}
\]

The error terms \( \varepsilon_n \) and \( u_n \) have been ignored in the derivation of (2.16) and (2.17).

Both poverty lines depend on the family composition and on the distribution of incomes and family compositions in a family's reference group. If \( \delta \) and \( \psi \) are equal to zero, it is clear from (2.16) and (2.17) that both poverty lines increase proportionally with (redefined) family size. For the LPL this is restrictive for the following reason.

The dependence of the LPL on family size defines equivalence scales (see, e.g., Kapteyn and Van Praag 1976, 1980), which tell us how family income has to vary with family composition in order to keep the family at welfare level \( \alpha \). Proportionality implies that in order to compensate for the birth of a baby, for example, it takes the same percentage of family income for both a very high welfare level \( \alpha \) (and hence a high income prior to the birth of the baby) and a very low welfare level \( \alpha \) (with a corresponding low income prior to the birth of the baby). It is more likely, however, that the cost of a baby for a richer family is a higher amount but a smaller proportion of income, than for a poorer family. For \( \delta < 0 \), this is generally what the model implies (see Section V).

**B. Preliminary Empirical Results**

Models (2.14) and (2.15) have been estimated for a sample of 773 households taken from the Dutch population in 1982. This so-called labor mobility survey only samples families whose head is under 65. In other respects the sample is random. Table 2.1 presents some of the estimation results that have been obtained with the LISREL-V maximum likelihood computer program. Equations (2.14) and (2.15) have been estimated jointly, allowing for correlation between \( \varepsilon_n \) and \( u_n \). A likelihood ratio test of the restriction that the age function parameters \( \gamma_2 \) and \( \gamma_3 \) are identical in both the \( \mu \)- and \( y_{\text{min}} \)-equation did not lead to rejection. This equality was therefore imposed. Equality of the remaining parameters across the two equations was rejected at the 5 percent level by a likelihood ratio test.

The large standard errors of \( \hat{\beta}_1 \), \( \hat{\delta} \), \( \hat{\alpha}_1 \), \( \hat{\psi} \) are not necessarily disturbing because one should consider the effects of \( \hat{\beta}_1 \) and \( \hat{\delta} \) (and of \( \hat{\alpha}_1 \) and \( \hat{\psi} \)) jointly. A closer inspection of the estimated values of \( \beta_1 \) and \( \delta \) reveals an anomaly, however. For large values of the welfare level \( \alpha \) (\( \geq 0.7 \)) and
Table 2.1
First Estimation Results (Standard Errors in Parentheses)\textsuperscript{a}

<table>
<thead>
<tr>
<th>Equation (2.14)</th>
<th>Equation (2.15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1 = 31.07 (22.05)$</td>
<td>$\hat{\alpha}_1 = 18.64 (11.33)$</td>
</tr>
<tr>
<td>$\hat{\beta}_2 = 0.89 (0.06)$</td>
<td>$\hat{\alpha}_2 = 0.85 (0.07)$</td>
</tr>
<tr>
<td>$\delta = -2.98 (2.11)$</td>
<td>$\hat{\psi} = -1.79 (1.03)$</td>
</tr>
<tr>
<td>$\hat{\gamma}_2 = 10.11 \times 10^{-2} (2.17 \times 10^{-2})$</td>
<td>$\hat{\gamma}_3 = -2.77 \times 10^{-3} (0.55 \times 10^{-3})$</td>
</tr>
</tbody>
</table>

Number of observations: 773

\textsuperscript{a} Only those results are reported that are relevant for the analysis in subsequent sections.

with all variables in (2.16) set at their sample mean we find that an increase in family size leads to a reduction in the poverty line. More generally, the effects of family size on the cost of living of a family turn out to be quite small. This result has been obtained more often and it has been criticized as being implausible (e.g., by Watts 1985). This problem will be addressed in the next section.

III. Measurement of Income

A central notion emerging from the previous sections is that an individual's evaluation of income levels strongly depends on the level of his own income. A complicating factor in this context is that people in general only know approximately the level of their actual income. When answering the income evaluation question and the minimum income question, the respondent will take his estimate of his actual income as a frame of reference. Due to the fact that in our survey household income is measured twice, however, we are able to determine the biasing effects of the respondents' systematic errors in "estimating" their own income.

The income evaluation question in the questionnaire is preceded by a question which asks the respondent to indicate in which one of seven income brackets net household income falls. It is likely that the income level the respondent has in mind when answering this question is also the income level he refers to when answering the subsequent income evaluation questions. The last section of the questionnaire asks the respondent
to provide detailed information on a large number of different components of the household's net income, such as earned income, fringe benefits, family allowance, spouse’s income, etc. The aggregate of these components is likely to provide a much more reliable measure of the household’s income. A comparison of the two income measures suggests that respondents tend to underestimate household income when answering the income class question; see Table 3.1.

In order to analyze the systematic difference between the two income measures, we postulate the following relation between the income $y_n^*$ underlying the answer to the income question in brackets and the income components $y_{ni} (i = 1, \ldots, m)$ recorded at the end of the questionnaire:

$$y_n^* = \left( \sum_{i=1}^{m} \alpha_i y_{ni} \right) \cdot e^{\eta_n}$$

where the $\alpha_i$'s are unknown parameters and $\eta_n$ is a normally distributed error term with zero mean and variance $\sigma_{\eta}^2$. The $n$th respondent’s answer falls into the $i$th bracket if $y^*$ is between the upper and lower bound of this bracket. The values of $\alpha_i$ are expected to lie in the unit interval. The smaller a parameter $\alpha_i$, the more the respondent “forgets” the $i$th income component when answering the income question in brackets.

The parameters $\alpha_i$ and $\sigma_{\eta}^2$ can be estimated by means of maximum likelihood. The likelihood function is given by

<table>
<thead>
<tr>
<th>Income Bracket</th>
<th>Average Income$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 17,500</td>
<td>8,769</td>
</tr>
<tr>
<td>17,500–20,000</td>
<td>23,918</td>
</tr>
<tr>
<td>20,000–24,000</td>
<td>28,552</td>
</tr>
<tr>
<td>24,000–28,000</td>
<td>30,743</td>
</tr>
<tr>
<td>28,000–34,000</td>
<td>36,086</td>
</tr>
<tr>
<td>34,000–43,000</td>
<td>41,656</td>
</tr>
<tr>
<td>&gt; 43,000</td>
<td>59,189</td>
</tr>
</tbody>
</table>

a. Dutch Guilders per year.
b. Average income of all households in the income bracket according to the second income measure.
\[(3.2) \quad L(\alpha_1, \ldots, \alpha_n, \sigma_n^2) = \prod_{n \in \theta_1} \Phi\left( \frac{\log a_1 - \log(\Sigma \alpha_i y_{ni})}{\sigma_n} \right) \prod_{n \in \theta_2} \Phi\left( \frac{\log a_k - \log(\Sigma \alpha_i y_{ni})}{\sigma_n} \right) \prod_{n \in \theta_3} 1 - \Phi\left( \frac{\log a_6 - \log(\Sigma \alpha_i y_{ni})}{\sigma_n} \right) \]

where

- \( n \in \theta_1 \): if the \( n \)th household’s income falls into the first income bracket
- \( n \in \theta_2 \): if the \( n \)th household’s income falls into the \( k \)th income bracket \((k = 2, \ldots, 6)\)
- \( n \in \theta_3 \): if the \( n \)th household’s income falls into the seventh income bracket
- \( a_k \): upper bound of the \( k \)th income bracket
- \( \Phi(\cdot) \): standard normal distribution function

and \( y_{ni} \) denotes the \( i \)th income component of the \( n \)th household. The results of the maximum likelihood estimation of the parameters in model (3.1) are presented in Table 3.2. Apparently, only the respondent’s own

<table>
<thead>
<tr>
<th>Table 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimation Results Equation (3.1)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income Component</th>
<th>( \alpha_i )</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earned income of respondent from main occupation</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>Holiday allowance plus fringes of respondent</td>
<td>0.60</td>
<td>0.11</td>
</tr>
<tr>
<td>Rent subsidy plus income from subletting rooms</td>
<td>0.06</td>
<td>0.14</td>
</tr>
<tr>
<td>Property income plus income from secondary jobs</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>Family allowance</td>
<td>0.17</td>
<td>0.09</td>
</tr>
<tr>
<td>Income of the spouse</td>
<td>0.77</td>
<td>0.04</td>
</tr>
<tr>
<td>Income of oldest child of respondent</td>
<td>0.21</td>
<td>0.05</td>
</tr>
<tr>
<td>Income of other household members</td>
<td>0.29</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Number of observations: 811
earned income is fully taken into account when answering the income question in brackets. All other components are to some extent "forgotten," especially rent subsidies and family allowances.

We assume that if the respondent would have been aware of the actual value of his household's income as measured by the sum of the income components, the resulting values for $e^\mu$ and $y_{min}$ would have been higher by the same percentage as by which $\sum y_i$ exceeds $\sum \alpha_i y_i$. Thus, we adjust the measured values of $e^\mu$ and $y_{min}$ for each respondent as follows:

\[
(3.3) \quad e^{\mu_n} = e^\mu \left( \frac{\sum_{i=1}^{m} y_{ni}}{\sum_{i=1}^{m} \alpha_i y_{ni}} \right)
\]

\[
(3.4) \quad \hat{y}_{min,n} = y_{min,n} \left( \frac{\sum_{i=1}^{m} y_{ni}}{\sum_{i=1}^{m} \alpha_i y_{ni}} \right)
\]

where $e^{\mu_n}$ and $\hat{y}_{min,n}$ are the adjusted values for household $n$.

Subsequently, models (2.14)-(2.15) has been reestimated with $\mu_n$ and $\ln y_{min,n}$ replaced by $\hat{\mu}$ and $\ln \hat{y}_{min}$. Table 3.3 gives some of the results. The earlier observed anomaly has disappeared. For a very wide range of incomes, the estimates of $\delta$ and $\beta_1$ imply an increase of the cost of living when family size increases. These effects are presented in greater detail in Section V. Furthermore, it appears that $\psi$ is not significantly different from zero. Imposition of the constraint $\psi = 0$, yields the estimation results reported in Table 3.4. (Since the results of this model will be used for the construction of poverty lines, in Section V, they are presented in more detail than the previous ones.) The estimates in Tables 3.3 and 3.4 are almost identical. A discussion of the economic interpretation of the estimates is deferred to Section V.

Table 3.3

<table>
<thead>
<tr>
<th>Equation (2.14)</th>
<th>Equation (2.15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1 = 1.94 (1.10)$</td>
<td>$\hat{\alpha}_1 = 0.40 (0.75)$</td>
</tr>
<tr>
<td>$\hat{\beta}_2 = 0.57 (0.06)$</td>
<td>$\hat{\alpha}_2 = 0.44 (0.07)$</td>
</tr>
<tr>
<td>$\delta = -0.15 (0.10)$</td>
<td>$\hat{\psi} = 0.002 (0.07)$</td>
</tr>
<tr>
<td>$\hat{\gamma}_2 = 0.27 \times 10^{-2} (1.17 \times 10^{-2})$</td>
<td>$\hat{\gamma}_3 = -0.12 \times 10^{-3} (0.30 \times 10^{-3})$</td>
</tr>
</tbody>
</table>

Number of observations: 773
Table 3.4

*Estimation Results with $\psi = 0$ (Standard Errors in Parentheses)*

<table>
<thead>
<tr>
<th>Equation (2.14)</th>
<th>Equation (2.15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_0 = 2.72$ (0.41)</td>
<td>$\hat{\alpha}_0 = 3.16$ (0.39)</td>
</tr>
<tr>
<td>$\hat{\beta}_1 = 1.92$ (0.78)</td>
<td>$\hat{\alpha}_1 = 0.43$ (0.04)</td>
</tr>
<tr>
<td>$\hat{\beta}_2 = 0.58$ (0.05)</td>
<td>$\hat{\alpha}_2 = 0.45$ (0.03)</td>
</tr>
<tr>
<td>$\delta = -0.15$ (0.07)</td>
<td>$\hat{\alpha}_3 = 0.19$ (0.05)</td>
</tr>
<tr>
<td>$\hat{\beta}_3 = 0.13$ (0.04)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_2 = 0.24 \times 10^{-2}$ (1.18 $\times$ 10$^{-2}$)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_3 = 0.12 \times 10^{-3}$ (0.3 $\times$ 10$^{-3}$)</td>
<td></td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon = 0.22$ (0.0056)</td>
<td>$\hat{\sigma}_\mu = 0.25$ (0.0065)</td>
</tr>
<tr>
<td>$\hat{\sigma}_{\epsilon\mu} = 0.04$ (0.002)</td>
<td></td>
</tr>
<tr>
<td>$R^2 = 0.59$</td>
<td>$R^2 = 0.56$</td>
</tr>
</tbody>
</table>

Number of observations: 773

IV. Sample Selectivity

Almost any empirical work based on micro-data is confronted with the problem that a number of observations in the available sample cannot be used in the analysis, because of missing information on one or more variables. Usually, these observations are simply left out in the hope that the omitted observations are more or less "random" with respect to the analysis. If the number of deleted observations is relatively large, however, and the dropout is connected with the endogenous variables of the model, the results might be subject to selection bias.

To analyze this problem in more detail, we assume that the process by which observations are removed from the sample can be described by a selection equation of the form:

\[ y_{3n}^* = X_{3n}' \eta + \nu_n \]  

The household is removed from the sample if $y_{3n}^* \geq 0$ and retained otherwise. The variables in $X_{3n}$ are thought to affect the selection probability of the $i$th household.\(^3\) We assume that $\epsilon_n$, $u_n$ and $v_n$ follow a multivariate normal distribution with zero mean and covariance matrix

\[ \Omega = \begin{pmatrix} \sigma^2_\epsilon & * & * \\ \sigma_{\epsilon u} & \sigma^2_v & * \\ \sigma_{\epsilon v} & \sigma_{\epsilon u} & 1 \end{pmatrix} \]

\(3\). We assume that $X_{3n}'$ is observed for all households.
If the researcher ignores the possibility of selection bias, he simply estimates (2.14) and (2.15) under the assumption that $Ee_n = Eu_n = 0$. The expectation of the error terms in (2.14) and (2.15), however, has to be taken conditionally on the household being retained (selected) in the sample, that is

\[(4.3) \ E(e_n|y_n^* < 0) = E(e_n|v_n < -X_{3n}^\prime \eta)\]

\[(4.4) \ E(u_n|y_n^* < 0) = E(u_n|v_n < -X_{3n}^\prime \eta)\]

These expectations are generally unequal to zero. Consequently, $\beta_1$ and $\beta_2$ will be estimated inconsistently.

We will estimate the selection Equation (4.1) jointly with (2.14) and (2.15) and test whether selection bias is present, which is equivalent to testing the null hypothesis $\sigma_{e_v} = \sigma_{u_v} = 0$. We will first give a verbal description of the selection process.

In the analysis of Sections II and III a fairly large proportion of the observations could not be used due to deficient information with respect to one or more variables. In 122 cases the respondent did not answer the MIQ or provided insufficient information to estimate $\mu$, whereas in 145 cases net household income could not be calculated due to insufficient information on one or more income components. We assume that the selection of these 267 households can be described by a single selection equation. In addition, 18 observations had to be left out because of various other deficiencies. Since this latter selection pertains to a small number of observations only, we ignore the possible selection bias that may result.

In order to estimate Equations (4.1), (2.14) and (2.15), we have to choose a set of variables which are thought to affect the selection probability:

- Constant term
- Log age of respondent
- Squared log age of respondent
- Dummy variable = 1 if the respondent is an employed wage earner
  = 0 otherwise
- Dummy variable = 1 if the respondent is a higher executive of a firm
  = 0 otherwise
The likelihood function of the observations for model (2.14), (2.15), (4.1) is given by:

\[(4.5) \quad L = \prod_{n \in \theta} \left( \int_{-\infty}^{0} f_n(\hat{\mu}, \ln y_{\min}, y^*_3) dy^*_3 \right) \prod_{n \in \Theta} \int_{0}^{\infty} f_{3n}(y^*_3) dy^*_3 \]

where

- \(n \in \theta_0\) if the household is retained in the sample,
- \(n \in \Theta\) if the household is removed from the sample,
- \(f_n(\cdot)\) is the joint normal distribution function of \((\hat{\mu}, \ln y_{\min}, y^*_3)\)
- \(f_{3n}(\cdot)\) is the marginal distribution function of \(y^*_3\), for the \(n\)th household.

The estimation results are presented in Table 4.1. For comparison we also show the estimates from Section III.

Although the differences between the first two and the last two columns in Table 4.1 are generally small, a likelihood ratio test strongly rejects the null hypothesis \(\sigma_{uv} = \sigma_{\nu\nu} = 0\).

Furthermore, it appears that the probability of being removed from the sample increases with age. For example, the probability for a 25-year-old respondent is 0.11 on average, whereas it rises to 0.48 for a 55-year-old respondent. The probability of being removed from the sample also increases if the respondent is a higher executive.

V. Poverty Lines

To obtain a better feeling for the importance of the issues dealt with in the previous sections we compare the poverty lines implied by the three models of Sections II, III, and IV. First the age functions \(f(a_j)\) for each of the three models are considered. These have been drawn in Figure 5.1.

The age function for the model of Section II, illustrates the implausible results obtained on the basis of the incorrect income measures. The age functions for both other models do not differ widely, and both look plausible. The preferred model, with correction for sample selectivity, shows somewhat less steep age effects than the model of Section III.

Next we illustrate how the cost of an extra child varies with income and family size. Figure 5.2 presents a few examples, based on the estimates of the third model. Figure 5.2 shows that the cost of the extra child, that is the youngest child in the family, increases with income but less than proportionally. Furthermore, there are substantial economies of scale. For instance, the two-year-old child costs considerably less in the six person family than in the four person family (see Cases III and IV). Older
Table 4.1
Results of Joint Estimation of Equations (2.14), (2.15), (4.1)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Estimate from Table 3.4</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>2.86</td>
<td>0.41</td>
<td>2.72</td>
<td>0.41</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1.68</td>
<td>0.69</td>
<td>1.92</td>
<td>0.78</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.55</td>
<td>0.05</td>
<td>0.58</td>
<td>0.05</td>
</tr>
<tr>
<td>$\delta$</td>
<td>-0.13</td>
<td>0.06</td>
<td>-0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.14</td>
<td>0.04</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.185*10^{-2}</td>
<td>0.012</td>
<td>0.235*10^{-2}</td>
<td>0.012</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>-0.80*10^{-4}</td>
<td>0.0003</td>
<td>-0.117*10^{-3}</td>
<td>0.0003</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>3.23</td>
<td>0.39</td>
<td>3.16</td>
<td>0.39</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.39</td>
<td>0.04</td>
<td>0.43</td>
<td>0.04</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.43</td>
<td>0.03</td>
<td>0.45</td>
<td>0.03</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>0.20</td>
<td>0.05</td>
<td>0.19</td>
<td>0.05</td>
</tr>
<tr>
<td>$\eta_0$</td>
<td>-8.91</td>
<td>5.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>4.18</td>
<td>2.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>-0.52</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_3$</td>
<td>-0.018</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_4$</td>
<td>0.88</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_{\epsilon}$</td>
<td>0.27</td>
<td>0.009</td>
<td>0.22</td>
<td>0.006</td>
</tr>
<tr>
<td>$\sigma_{\mu}$</td>
<td>0.28</td>
<td>0.010</td>
<td>0.25</td>
<td>0.007</td>
</tr>
<tr>
<td>$\sigma_{\epsilon \mu}$</td>
<td>0.06</td>
<td>0.005</td>
<td>0.04</td>
<td>0.002</td>
</tr>
<tr>
<td>$\sigma_{\epsilon \nu}$</td>
<td>0.24</td>
<td>0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_{\mu \nu}$</td>
<td>0.18</td>
<td>0.025</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

children cost more than young children. Despite the economies of scale due to a large family, the eleven-year-old child in Case II costs more than the two-year-old child in Case III.

Finally, we have computed poverty lines for some selected family compositions based on the three models for the SPL and two versions of the LPL ($\alpha = 0.4$ and $\alpha = 0.5$) (see Table 5.1). In the calculation of the poverty lines both $m_n$ and $hs_n$ have been set equal to the sample means of log-income and log-family size, respectively. The standard errors in parentheses are based on an asymptotic approximation, which follows straightforwardly from the fact that the poverty lines are differentiable functions of the parameters. For comparison, the last column of Table 5.1
The official poverty line exceeds the amounts based on the model of Section II. This model also suggests substantial economies of scale for large families, that is the estimated poverty lines increase very slowly with an increase in family size and sometimes even decrease (compare, for instance, two adults and two adults with a six-year-old child). These findings are typical for the results that have been obtained in earlier studies. See, for example, Goedhart et al. (1977), Colasanto, Kapteyn, and Van der Gaag (1984). (In these studies a negative weight for certain ages was not observed, simply because no weighing on the basis of age took place.)

The model of Section III yields quite a different picture. The level of the SPL corresponds roughly with the official poverty line and both versions
of the LPL exceed the official poverty line. The economies of scale for a large family also look more plausible than in the previous model.

The most striking outcome of the model of Section IV is that the correction for sample selectivity bias leads to a downward revision of approximately 10 percent of the level of the poverty line as compared with the results of the model of Section III. As a result we see that the official poverty line is generally high enough to allow households to make ends meet. Furthermore, economies of scale are plausible. They are a little larger than implied by the official poverty line.

VI. Conclusions

Since income is the central concept in most social security and welfare policies, the implementation of any poverty line definition
Table 5.1
Poverty Lines for Various Family Compositions, According to Three Models (Dutch Guilders $\times 100$)

<table>
<thead>
<tr>
<th>Family Composition</th>
<th>Model of Section II</th>
<th>Model of Section III (Correction for income measurement errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPL</td>
<td>LPL(0.4)</td>
</tr>
<tr>
<td>1 Adult</td>
<td>116(1)</td>
<td>137(9)</td>
</tr>
<tr>
<td>2 Adults</td>
<td>160(6)</td>
<td>185(5)</td>
</tr>
<tr>
<td>2 Adults + 6</td>
<td>130(9)</td>
<td>153(9)</td>
</tr>
<tr>
<td>2 Adults + 12</td>
<td>157(7)</td>
<td>182(6)</td>
</tr>
<tr>
<td>2 Adults + 12,6</td>
<td>136(11)</td>
<td>159(11)</td>
</tr>
<tr>
<td>2 Adults + 12,6,1</td>
<td>144(10)</td>
<td>168(1)</td>
</tr>
<tr>
<td>2 Adults + 18,12,6,1</td>
<td>169(8)</td>
<td>195(7)</td>
</tr>
<tr>
<td>2 Adults + 12,6,2,1</td>
<td>141(13)</td>
<td>165(14)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family Composition</th>
<th>Model of Section IV (Correction for sample selection bias)</th>
<th>Official Poverty Line</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPL</td>
<td>LPL(0.4)</td>
</tr>
<tr>
<td>1 Adult</td>
<td>117(6)</td>
<td>132(7)</td>
</tr>
<tr>
<td>2 Adults</td>
<td>154(6)</td>
<td>171(6)</td>
</tr>
<tr>
<td>2 Adults + 6</td>
<td>174(7)</td>
<td>190(7)</td>
</tr>
<tr>
<td>2 Adults + 12</td>
<td>178(7)</td>
<td>194(7)</td>
</tr>
<tr>
<td>2 Adults + 12,6</td>
<td>195(8)</td>
<td>208(8)</td>
</tr>
<tr>
<td>2 Adults + 12,6,1</td>
<td>207(9)</td>
<td>218(9)</td>
</tr>
<tr>
<td>2 Adults + 18,12,6,1</td>
<td>225(10)</td>
<td>232(10)</td>
</tr>
<tr>
<td>2 Adults + 12,6,2,1</td>
<td>217(11)</td>
<td>226(10)</td>
</tr>
</tbody>
</table>

should be based on an accurate measurement of household income. As the subjective poverty line definitions try to elicit directly which income level is necessary to make ends meet or to guarantee a certain welfare level, it is important that respondents have an accurate knowledge of their income. For the purpose of questionnaire design, this yields two alternatives. The first alternative is the procedure adopted in this paper, where the subjective questions are preceded by a question which measures the respondent's perception of his own after tax household income. On the basis of an accurate measurement of income later in the questionnaire,
one can then adjust the response to the subjective questions. A second alternative is to begin with a large number of factual questions about household income, total the components and only then pose the subjective questions.

The finding that sample selectivity creates a significant problem is not surprising. The poor tend to have characteristics that give them a lower probability of participation in surveys. Correction for sample selectivity is important to avoid policies that are aimed at the poor, but which are mainly based on observations from a middle class population.

The poverty lines discussed here, are based on an explicit model of determinants of subjective evaluations. It appears that any poverty line will be based on some implicit or explicit model of behavior or valuation. To construct an internally consistent or reliable poverty line, the model on which it is based should be correctly specified. The model (2.14)-(2.15) used in this paper is misspecified, because, due to the lack of adequate data, the influence of lagged variables had to be ignored. Hence, the empirical results should not be taken too seriously. The poverty lines that are discussed here require both panel data and a questionnaire design that allows for an accurate treatment of one of the crucial variables in social policy: income.

References


Kapteyn, A., S. van de Geer, and H. van de Stadt. 1985. "The Impact of Changes in Income and Family Composition on Subjective Measures of


Reprint Series, CentER, Tilburg University, The Netherlands:


No. 8  Th. van de Klundert and F. van der Ploeg, Wage rigidity and capital mobility in an optimising model of a small open economy, *De Economist* 137, nr. 1, 1989, pp. 47 - 75.


