Imputing a continuous income variable from a bracketed income variable with special attention to missing observations

Steven Stern *

University of Virginia, Charlottesville, VA 22901, USA

Received 8 April 1991
Accepted 14 August 1991

This paper proposes a method, based on polychotomous discrete choice methods, to impute a continuous measure of income when only a bracketed measure of income is available and for only a subset of the observations. The method is shown to perform well with CPS data.

1. Introduction

Many cross section data sets report a respondent’s income in terms of a bracketed measure. Furthermore, income questions, whether bracketed or not, are notorious for high non-response rates. In fact, sometimes it is argued that income questions should be asked in bracketed form to achieve higher response rates. On the other hand, a researcher may need a continuous measure of income for his analysis. This may be because he wants to conserve on degrees of freedom or because it is the continuous measure of income that appears in the model. Also he may find it painful to exclude all observations with missing income variables.

This paper proposes a method to construct a continuous measure of income for all observations in a data set in which only bracketed income is observed and only for a subset of the observations. It uses ordered logit and piece-wise log-linear splining. For those observations where bracketed income is observed, the error is relatively small. For those observations where bracketed income is not observed, the size of the error depends upon how it is measured and upon other characteristics of the data set.

In the next section, I describe the method. Section 3 applies it to the January 1984 CPS where a continuous measure of income is observed. Section 4 concludes.

* I would like to thank Derek Neal for assistance with data. Peggy Pasternak provided excellent typing. The research was supported by the National Institute of Aging under Grant No. 5 R29 AG06387-05 and the AARP Andrus Foundation. All errors are mine.

1 For example, the 1979 National Health Interview Survey reports family income as falling into under $1000, $1000–$1999, $2000–$2999, $3000–$3999, $4000–$4999, $5000–$5999, $6000–$6999, $7000–$7999, $10000–$14999, $15000–$24999, and above $24999.
2. Methodology

I assume that the reader has a data set with \( N \) observations where bracketed income is observed for the first \( N_1 \) observations and is not observed for the last \( N_2 \) observations; \( N_1 + N_2 = N \). More precisely, for the first \( N_1 \) observations a variable \( y_i, i = 1, 2, \ldots, N_1 \) is observed where \( y_i = j \) if \( c_j \leq y_i^{*} \leq c_{j+1} \), \( y_i^{*} \) is true income, and \( c_j, j = 1, 2, \ldots, M + 1 \) are known constants. The latent variable, \( y_i^{*} \), is not observed. There is a set of exogenous variables, \( X_i, i = 1, 2, \ldots, N \), observed for everyone that affects \( y_i^{*} \) according to

\[
g(y_i^{*}) = X_i \beta + u_i,
\]

where \( u_i \sim \text{iid } N(0, \sigma^2) \). The goal is to construct an estimate of \( y_i^{*} \) for \( i = 1, 2, \ldots, N \).

Let \( g(c_j) = a_j \) for \( j = 1, 2, \ldots, M + 1 \). Then

\[
\Pr(y_i = j) = \Pr[a_j \leq g(y_i^{*}) \leq a_{j+1}] = \phi\left(\frac{a_{j+1} - X_i \beta}{\sigma}\right) - \phi\left(\frac{a_j - X_i \beta}{\sigma}\right),
\]

where \( \phi \) is the standard normal distribution function. This is a standard ordered probit problem. Set \( a_1 = -\infty, a_2 = 0, a_{M+1} = \infty \), and \( \sigma = 1 \). Then \( a_j, j = 3, 4, \ldots, M \), and \( \beta \) can be estimated with maximum likelihood estimation.

Once \( \beta \) is estimated, I can construct an estimate of \( \mathbb{E}(g(y_i^{*})) \). If \( y_i \) is observed then,

\[
\mathbb{E}\left[ g(y_i^{*}) \mid y_i = j \right] = X_i \beta + \mathbb{E}\left[ u_i \mid a_j - X_i \beta \leq u_i \leq a_{j+1} - X_i \beta \right],
\]

\[
= X_i \beta + \frac{\phi(a_j - X_i \beta) - \phi(a_{j+1} - X_i \beta)}{\phi(a_{j+1} - X_i \beta) - \phi(a_j - X_i \beta)},
\]

where \( \phi \) is the standard normal density function. If \( y_i \) is not observed then

\[
\mathbb{E}(g(y_i^{*})) = X_i \beta.
\]

Equation (2.3) guarantees that \( a_j \leq g(y_i^{*}) \leq a_{j+1} \) while eq. (2.4) does not.

The last step is to transpose the estimate of \( \mathbb{E}g(y_i^{*}) \) into an estimate of \( y_i^{*} \). One can use any appropriate method to estimate \( g(\cdot) \). However, the only relevant information is the estimates of \( g(c_j) = a_j \) and the known values of \( c_j, j = 1, 2, \ldots, M + 1 \). Thus, I use a piece wise log-linear spline, mainly because it is easy to use. In particular, assume

\[
y_i^{*} = \exp(a_j + \delta_j g(y_i^{*})),
\]

if \( a_j \leq g(y_i^{*}) \leq a_{j+1} \). One can solve for the \( \alpha \)'s and \( \delta \)'s by using

\[
c_j = \exp(a_j + \delta_j a_j), \quad j = 2, 3, \ldots, M.
\]

One must assume that \( \delta_1 = \delta_2 \) and \( \delta_{M+1} = \delta_M \). Given the \( \alpha \)'s and \( \delta \)'s, one can apply equation (2.5) to each observation to construct an estimate of \( y_i^{*} \).

---

2 See, for example, Maddala (1985).
3 Also \( \alpha_1 = \alpha_2 \) and \( \alpha_{M+1} = \alpha_M \), but this is no restriction.
While the estimate of $g(y^*)$ is unbiased, the estimate of $y^*_i$ (or $\ln y^*_i$) is not. However, the bias may be small. Further, there is usually no other alternative.

3. Results

The method described in section 2 is applied to the January 1984 Current Population Survey (CPS). The CPS has 6000 observations for whom the first 3000 are assumed to have observed bracketed income. The income measure is weekly labor income, and the bracket levels are multiples of $100 up to $500. A regression on log income is

$$\log y_i = 1.14 + 0.074 \, EDUC_i - 0.053 \, BLACK_i + 0.010 \, AGE_i + 0.196 \, URBAN_i + 0.867 \, \log HOURS_i + e_i,$$

$$(0.11) \quad (0.003) \quad (13.96) \quad (0.026) \quad (0.12) \quad (0.001) \quad (37.13)$$

where numbers in parentheses under coefficient estimates are standard errors, numbers in parentheses under variables are variable means, and $R^2 = 0.480$. One can see that CPS estimates are precise and of the right sign and magnitude.

The ordered probit equation is

$$y^*_i = 0.815 + 0.157 \, EDUC_i - 0.103 \, BLACK_i + 0.021 \, AGE_i - 0.824 \, FEMALE_i + u_i,$$

$$(0.131) \quad (0.007) \quad (0.063) \quad (0.002) \quad (0.040)$$

where $c_1 = 0.909 \, (0.034)\), $c_2 = 1.603 \, (0.038)\), $c_3 = 2.179 \, (0.041)\), $c_4 = 2.710 \, (0.044)\)$, and numbers in parentheses are standard errors. Using the estimates in eq. (3.2) and eq. (2.5), I constructed the $\alpha$'s and $\delta$'s. These are reported in table 1. One can see that income increases a little slower than log linearly in terms of $X\beta$.

Table 2 provides information on the bias and standard deviation of any esitmate of $y^*_i$ for the sample. This is done for both $y^*_i$ and $\ln(y^*_i)$. These are compared to setting the estimate of $y^*_i$

<table>
<thead>
<tr>
<th>Income bracket</th>
<th>CPS</th>
<th>$\alpha$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1</td>
<td>4.605</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td>1–2</td>
<td>4.605</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td>2–3</td>
<td>4.767</td>
<td>0.584</td>
<td></td>
</tr>
<tr>
<td>3–4</td>
<td>4.904</td>
<td>0.499</td>
<td></td>
</tr>
<tr>
<td>4–5</td>
<td>5.076</td>
<td>0.420</td>
<td></td>
</tr>
<tr>
<td>5–</td>
<td>5.076</td>
<td>0.420</td>
<td></td>
</tr>
</tbody>
</table>

* CPS weekly income brackets are in hundreds of dollars.
Table 2
Goodness of fit. a

<table>
<thead>
<tr>
<th></th>
<th>In sample</th>
<th></th>
<th>Crude average</th>
<th></th>
<th>Out of sample</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ordered probit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>Bias(Log) b</td>
<td>Bias</td>
<td>Bias</td>
<td>Bias</td>
<td>Bias(Log)</td>
</tr>
<tr>
<td>0–1 c</td>
<td>0.03</td>
<td>0.024</td>
<td>0.23</td>
<td>(0.19)</td>
<td>(0.078)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>1–2</td>
<td>0.16</td>
<td>0.089</td>
<td>0.13</td>
<td>0.132</td>
<td>(0.30)</td>
<td>(0.196)</td>
</tr>
<tr>
<td>2–3</td>
<td>0.14</td>
<td>0.053</td>
<td>0.10</td>
<td>0.059</td>
<td>(0.28)</td>
<td>(0.110)</td>
</tr>
<tr>
<td>3–4</td>
<td>0.14</td>
<td>0.039</td>
<td>0.10</td>
<td>0.037</td>
<td>(0.31)</td>
<td>(0.087)</td>
</tr>
<tr>
<td>4–5</td>
<td>0.19</td>
<td>0.041</td>
<td>0.14</td>
<td>0.036</td>
<td>(0.36)</td>
<td>(0.067)</td>
</tr>
<tr>
<td>5–</td>
<td>0.05</td>
<td>-0.005</td>
<td>0.09</td>
<td>-0.079</td>
<td>(1.17)</td>
<td>(1.166)</td>
</tr>
<tr>
<td>1–5</td>
<td>0.16</td>
<td>0.056</td>
<td>0.14</td>
<td>0.067</td>
<td>(0.31)</td>
<td>(0.126)</td>
</tr>
<tr>
<td>Total</td>
<td>0.12</td>
<td>0.041</td>
<td>(0.60)</td>
<td>(0.153)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Numbers in parentheses are centered standard deviations.
b Bias(Log) is bias of log income.
c Weekly income brackets are in hundreds of dollars.

equal to the midpoint of the income bracket levels:

\[
\hat{y}_i^* = \frac{(c_j + c_{j+1})}{2}
\]

if \( y_i = j \),

\[
(3.3)
\]

and

\[
\ln \hat{y}_i^* = \frac{(\ln c_j + \ln c_{j+1})}{2}
\]

if \( y_i = j \).

\[
(3.4)
\]

Equation (3.2) can not be used for \( j = M \), and eq. (3.4) can not be used for \( j = 1 \) or \( j = M \).

One sees in table 2 that the ordered probit estimates of \( y_i^* \) have larger bias (0.16 versus 0.14) and the same standard deviations (0.31). On the other hand, the ordered probit bias of \( \ln y_i^* \) is smaller (0.056 versus 0.067), leading to a smaller mean squared error (0.019 versus 0.021).

The biggest disadvantage of the crude estimator is that it can not be applied if bracketed income is not observed. The ordered probit estimator can as long as \( X_i \) is observed. Goodness of fit measures for the observations for whom bracketed income is assumed missing is reported in the last columns of table 2. While errors are much larger than in the first four columns, errors are still reasonably sized, especially for \( \ln y_i^* \).

4. Conclusion

The method suggested here can be used for observations with or without bracketed income as long as there are enough observations to estimate an ordered probit equation. Log income
estimates are quite precise for samples with informative exogenous variables and bracketed income precisely measured. Income is not as precisely measured but, if the only alternative, may be usable depending upon the quality of income data. The procedure is easy to apply and can be performed with a software package like LIMDEP. The researcher should use care in applying this method. But I recommend it as an alternative to using bracketed income and discarding many observations with missing income variables.

Reference
