The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

The Stata Journal (2007)
7, Number 2, pp. 167–182

Maximum likelihood and two-step estimation of an ordered-probit selection model

Richard Chiburis
Princeton University
Princeton, NJ
chiburis@princeton.edu

Michael Lokshin
The World Bank
Washington, DC
mlokshin@worldbank.org

Abstract. We discuss the estimation of a regression model with an ordered-probit selection rule. We have written a Stata command, `oheckman`, that computes two-step and full-information maximum-likelihood estimates of this model. Using Monte Carlo simulations, we compare the performances of these estimators under various conditions.

Keywords: st0123, oheckman, selection bias, ordered probit, maximum likelihood

1 Introduction

We implement full-information maximum likelihood (FIML) and two-step algorithms for the estimation of a linear regression model with an underlying ordered-probit selection rule. The selection rule may cause sample selection, regime switching, or a combination of both.

Several existing studies have used an ordered-probit selection model, but no estimation command has been available for Stata. In all articles discussed below, the two-step estimation procedure was used.

- Jimenez and Kugler (1987) analyze how the choice to attend a long vocational training program, a short program, or no program affects an earnings function for workers in Colombia. The instruments in the selection equation are data on primary education history and father's educational status.
- Idson and Feaster (1990) and Main and Reilly (1993) compute wage functions for workers in companies of different sizes, controlling for the worker's selection of company size. Idson and Feaster use marital and veteran status, and Main and Reilly use data on children as instruments in the selection equation.
- Ermisch and Wright (1993) and Paci et al. (1995) estimate wage equations for full-time and part-time workers by using an ordered probit to model the decision to work full-time, part-time, or not at all. Marital status and data on children are used as first-stage instruments for employment status.
- Carlsson (2004) computes regressions for airfares between pairs of cities separately depending on whether one, two, or more than two airlines operate between those cities. He models the selection of number of airlines by using an ordered probit.

However, he cannot reject the hypothesis that the equations are independent and therefore there is no selection bias.

- Bellemare and Barrett (2006) analyze livestock markets in Kenya and Ethiopia by using an ordered tobit model that consists of an ordered probit for classifying households into net buyers, autarkic, or net sellers, and then regressions to estimate quantity bought or sold.

Miranda and Rabe-Hesketh (2006) developed a wrapper command, `ssm`, for the Stata program `gllamm` (Rabe-Hesketh, Skrondal, and Pickles 2002) that fits a wide variety of selection models with a binary selection variable and discrete outcome variable. In contrast, the model we consider involves two or more selection categories and a continuous outcome variable. We also implement our FIML algorithm using the `d2` method for the `ml` command, which uses an analytically calculated gradient and Hessian matrix for the log likelihood to dramatically speed up the optimization process.

2 Methods

2.1 Model specification

Consider a model in which individuals i are sorted into $J+1$ categories $0, 1, \ldots, J$ on the basis of an ordered-probit selection rule:

$$\begin{aligned} z_i^* &= \alpha' \mathbf{w}_i + u_i; \\ z_i &= \begin{cases} 0 & \text{if } -\infty < z_i^* \le \mu_1, \\ 1 & \text{if } \mu_1 < z_i^* \le \mu_2, \\ 2 & \text{if } \mu_2 < z_i^* \le \mu_3, \\ \vdots & \\ J & \text{if } \mu_J < z_i^* < \infty \end{cases} \end{aligned} \tag{1}$$

where α is an unknown vector of parameters, u_i is a standard normal shock, and the unknown cutoffs $\mu_1, \mu_2, \ldots, \mu_J$ satisfy $\mu_1 < \mu_2 < \cdots < \mu_J$. We also define $\mu_0 \equiv -\infty$ and $\mu_{J+1} \equiv \infty$ to avoid having to handle the boundary cases separately. We assume that the independent variables $\mathbf{w}_i$ and the categorical variable z_i are observed, but the latent selection variable z_i^* is unobserved.

There is also an observed dependent variable y_i that is a linear function of some observed independent variables $\mathbf{x}_i$, but the coefficients of $\mathbf{x}_i$ depend on the category z_i:

$$y_i = \begin{cases} \beta_0' \mathbf{x}_i + \varepsilon_{i0} & \text{if } z_i = 0, \\ \beta_1' \mathbf{x}_i + \varepsilon_{i1} & \text{if } z_i = 1, \\ \vdots & \\ \beta_J' \mathbf{x}_i + \varepsilon_{iJ} & \text{if } z_i = J \end{cases} \tag{2}$$

where for each $j \in \{0, \ldots, J\}$, ε_{ij} has mean 0, has variance σ_j^2, and is bivariate normal with u_i with correlation ρ_j. We assume that the shocks ε_{ij} and u_i are independently and identically distributed across observations. Our goal is to estimate the parameter vectors $\beta_0, \ldots, \beta_J$. y_i could also be missing for certain categories j, in which case β_j, ρ_j, and σ_j do not exist.

Since only one category j is observed for each individual and the observations are independent, the correlations between ε_{ij} and ε_{ik} for $j \neq k$ cannot be identified, so we do not model or estimate them.[1]

As Heckman (1979) observed for the binary case, estimating any of the equations in (2) via ordinary least squares (OLS) generally leads to biased results. To see this, we define

$$\begin{aligned}
\lambda_i \equiv E[u_i \mid z_i, \mathbf{w}_i] &= \frac{\int_{\mu_j}^{\mu_{j+1}} (z_i^* - \alpha'\mathbf{w}_i)\phi(z_i^* - \alpha'\mathbf{w}_i)\, dz_i^*}{\Phi(\mu_{j+1} - \alpha'\mathbf{w}_i) - \Phi(\mu_j - \alpha'\mathbf{w}_i)} \\
&= \frac{-\int_{\mu_j}^{\mu_{j+1}} \phi'(z_i^* - \alpha'\mathbf{w}_i)\, dz_i^*}{\Phi(\mu_{j+1} - \alpha'\mathbf{w}_i) - \Phi(\mu_j - \alpha'\mathbf{w}_i)} \\
&= \frac{\phi(\mu_j - \alpha'\mathbf{w}_i) - \phi(\mu_{j+1} - \alpha'\mathbf{w}_i)}{\Phi(\mu_{j+1} - \alpha'\mathbf{w}_i) - \Phi(\mu_j - \alpha'\mathbf{w}_i)} \qquad (3)
\end{aligned}$$

where $j = z_i$. Then

$$\begin{aligned}
E[y_i \mid z_i, \mathbf{w}_i, \mathbf{x}_i] &= \beta_j'\mathbf{x}_i + E[\varepsilon_{ij} \mid z_i = j, \mathbf{w}_i] \\
&= \beta_j'\mathbf{x}_i + \rho_j\sigma_j\lambda_i \qquad (4)
\end{aligned}$$

Now consider an OLS regression of y on $\mathbf{x}$ over the subsample $\{i\colon z_i = j\}$. If we had added λ as an extra regressor, then the estimate $\widehat{\beta}_j$ would have been consistent, but without λ, the regression suffers from omitted-variable bias if $\rho_j \neq 0$ and will generally be inconsistent.

We next describe two methods for consistent estimation of the model: a two-step procedure and an FIML procedure.

2.2 Two-step estimation

The two-step estimation procedure has previously been described by Greene (2002) and is a generalization of Heckman's (1979) estimator for the binary case.

1. The correlation between ε_{ij} and ε_{ik} does matter when we want to counterfactually predict y_i in category k for an individual who actually chose category j. Our `yif()` postestimation statistic, which is described in section 4.4, implements such predictions under the assumption that ε_{ij} and ε_{ik} are conditionally independent given u_i.

In the first step, we estimate (1) by an ordered probit of z on $\mathbf{w}$, yielding the consistent estimates $\widehat{\alpha}, \widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_J$. Define $\widehat{z}_i^* \equiv \widehat{\alpha}'\mathbf{w}_i$. Then by using (3), a consistent estimator of λ_i is[2]

$$\widehat{\lambda}_i \equiv \frac{\phi(\widehat{\mu}_j - \widehat{z}_i^*) - \phi(\widehat{\mu}_{j+1} - \widehat{z}_i^*)}{\Phi(\widehat{\mu}_{j+1} - \widehat{z}_i^*) - \Phi(\widehat{\mu}_j - \widehat{z}_i^*)} \tag{5}$$

where $j = z_i$.

With (4), we can consistently estimate β_j with an OLS regression of y on $\mathbf{x}$ and $\widehat{\lambda}$ by using only the observations i for which $z_i = j$.

Let $\widehat{C}_j$ be the coefficient on $\widehat{\lambda}$ in this regression, and let RSS_j be the residual sum of squares for the regression. Let n_j be the number of observations in which equation j is observed. Then σ_j can be estimated as

$$\begin{aligned}
\widehat{\sigma_j} &\equiv \frac{1}{n_j}\left(RSS_j - \widehat{C}_j^2 \sum_{i:j=j} \frac{\partial\widehat{\lambda}_i}{\partial\widehat{z}_i^*}\right) \\
&= \frac{RSS_j}{n_j} - \frac{\widehat{C}_j^2}{n_j}\sum_{i:j=j}\left\{\frac{(\widehat{\mu}_j - \widehat{z}_i^*)\phi(\widehat{\mu}_j - \widehat{z}_i^*) - (\widehat{\mu}_{j+1} - \widehat{z}_i^*)\phi(\widehat{\mu}_{j+1} - \widehat{z}_i^*)}{\Phi(\widehat{\mu}_{j+1} - \widehat{z}_i^*) - \Phi(\widehat{\mu}_j - \widehat{z}_i^*)} - \widehat{\lambda}_i^2\right\}
\end{aligned}$$

Finally, since $\widehat{C_j}$ is a consistent estimator for $\rho_j\sigma_j$,

$$\widehat{\rho_j} \equiv \frac{\widehat{C_j}}{\widehat{\sigma_j}}$$

is a consistent estimator for ρ_j.

2.3 FIML estimation

FIML estimation consists of finding the parameter values that maximize the likelihood of the data. The parameters to be estimated are

$$\alpha; \beta_0, \beta_1, \ldots, \beta_{J-1}; \quad \mu_1, \mu_2, \ldots, \mu_J; \quad \rho_0, \rho_1, \ldots, \rho_{J-1}; \quad \sigma_0, \sigma_1, \ldots, \sigma_{J-1}$$

but β_j, ρ_j, and σ_j do not exist for categories j in which y is missing.

Given the parameters, the likelihood of an observation i in which the category is j and y_i is observed is

$$\begin{aligned}
L_{ij}^y &\equiv \mathrm{L}[y_i, j \mid \mathbf{x}_i, \beta_j, \sigma_j, \rho_j, \alpha, \mathbf{w}_i, \mu_j, \mu_{j+1}] \\
&= \mathrm{L}[y_i \mid \mathbf{x}_i, \beta_j, \sigma_j]\,\Pr[j \mid y_i, \mathbf{x}_i, \beta_j, \sigma_j, \rho_j, \alpha, \mathbf{w}_i, \mu_j, \mu_{j+1}] \\
&= \frac{1}{\sigma_j}\phi\left(t_i\right)\left[\Phi\left(\frac{\alpha'\mathbf{w}_i + \rho_j t_i - \mu_j}{\sqrt{1-\rho_j^2}}\right) - \Phi\left(\frac{\alpha'\mathbf{w}_i + \rho_j t_i - \mu_{j+1}}{\sqrt{1-\rho_j^2}}\right)\right] \qquad (6)
\end{aligned}$$

2. In the special case $\widehat{\mu}_j = 0$ and $\widehat{\mu}_{j+1} = \infty$, this simplifies to $\phi(\widehat{z}_i^*)/\Phi(\widehat{z}_i^*)$, which Heckman (1979) called the "inverse Mills' ratio."

where $t_i \equiv (y_i - \beta_j' \mathbf{x}_i)/\sigma_j$, ϕ is the standard normal density function, and Φ is the standard normal cumulative distribution function. The derivation uses the fact that if ε, u are standard bivariate normal with correlation ρ, then the conditional distribution of u given ε is normal with mean $\rho\varepsilon$ and variance $1 - \rho^2$.

If j is a category for which y is unspecified, then the likelihood is simply

$$L_{ij}^{\cdot} \equiv \Phi(\alpha' \mathbf{w}_i - \mu_j) - \Phi(\alpha' \mathbf{w}_i - \mu_{j+1}) \tag{7}$$

We take the logarithm of (6) or (7) to get the log likelihood for observation i, and since observations are independent we can add the log likelihood across observations to get the log likelihood for the entire sample:

$$\mathcal{L} \equiv \sum_{i=1}^{n} \begin{cases} \log L_{iz_i}^{y}, & \text{if } y_i \text{ is observed;} \\ \log L_{iz_i}^{\cdot}, & \text{if } y_i \text{ is missing} \end{cases} \tag{8}$$

2.4 Identification problems

If all variables in $\mathbf{w}$ are also in $\mathbf{x}$, then the identification of β_j is weak because $\widehat{\lambda}_i$ in (5) is a function of $\widehat{z}_i^* = \widehat{\alpha}' \mathbf{w}_i$. Since $\mathbf{x}$ and $\widehat{z}_i^*$ are collinear, $\mathbf{x}$ and $\widehat{\lambda}_i$ would be collinear except for the nonlinearity of the function $\widehat{\lambda}_i(\widehat{z}_i^*)$. Therefore, the identification of β_j relies on the specific form of the nonlinearity of $\widehat{\lambda}_i(\widehat{z}_i^*)$, in particular the normality of u_i, which is often a dubious assumption in practice. As noted by Puhani (2000), this is a well-known problem for Heckman's original estimator for the probit selection model.

In the ordered-probit selection model, this identification problem is especially bad for the selection categories $1 \leq j \leq J - 1$ in the interior of the range of z, for which both the lower and upper cutoffs $\widehat{\mu}_j$ and $\widehat{\mu}_{j+1}$ are finite. As shown in figure 1, $\widehat{\lambda}(\widehat{z}^*)$ is nearly linear when both cutoffs are finite, even when the cutoffs are relatively far apart. For smaller categories for which the cutoffs are closer together, the linearity is even stronger, and in the limit, $\lim_{\widehat{\mu}_{j+1} \to \widehat{\mu}_j} \widehat{\lambda}_i = \widehat{\mu}_j - \widehat{z}_i^*$, which is perfectly linear. The near-linearity of $\widehat{\lambda}(\widehat{z}^*)$ means that β_j is barely identified at all for interior categories j when $\mathbf{w}$ is a subset of $\mathbf{x}$.

Therefore, for the ordered-probit selection model $\mathbf{w}$ must contain a variable that is not in $\mathbf{x}$. That is, the researcher must have at least one instrument for the selection variable z that has no effect on y except through its effect on z. Such a variable w must be a significant determinant of z yet satisfy the exclusion restriction $\mathrm{Cov}(w, \varepsilon_j) = 0$ for all j.

This identification problem under the lack of an exclusion restriction affects both the two-step and the FIML estimation procedures, as we demonstrate with Monte Carlo simulations in section 6.6.

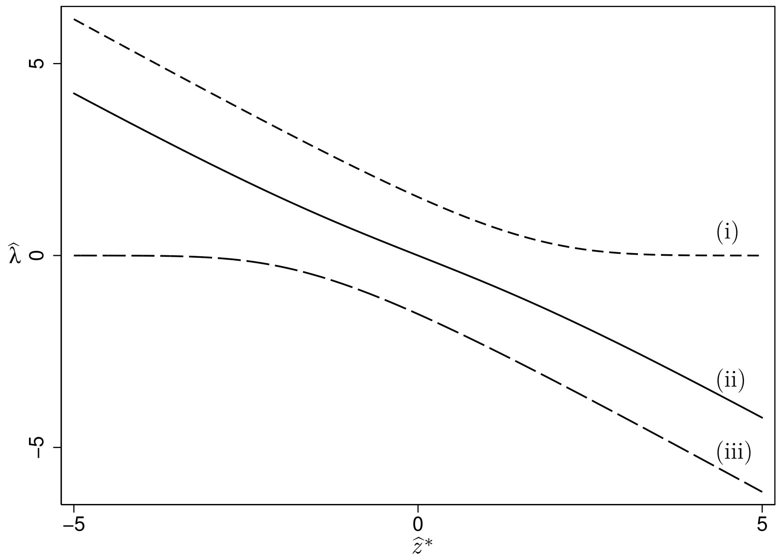

Figure 1: $\widehat{\lambda}$ as a function of $\widehat{z}^*$, from (5). From top to bottom: (i) $\widehat{\mu}_j = 1$, $\widehat{\mu}_{j+1} = \infty$; (ii) $\widehat{\mu}_j = -1$, $\widehat{\mu}_{j+1} = 1$; (iii) $\widehat{\mu}_j = -\infty$, $\widehat{\mu}_{j+1} = -1$. Observe that $\widehat{\lambda}(\widehat{z})$ is clearly nonlinear in cases (i) and (iii), in which the selection category has one infinite cutoff, but $\widehat{\lambda}(\widehat{z})$ is nearly linear for an interior selection category (ii), with both cutoffs finite.

3 Implementation details

FIML estimation of the model is implemented using the fast `d2` method for the `ml` command. In the `d2` method, the log likelihood is computed for each observation, along with its analytical gradient and Hessian matrix. The log-likelihood function is maximized using the modified Newton–Raphson algorithm.

For numerical reasons it is easiest for `ml` to estimate parameters that have domain $(-\infty, \infty)$, so we rescale some of the parameters before passing them to `ml`. We pass $\text{arctanh}(\rho_j)$ in place of ρ_j, $\ln(\sigma_j)$ in place of σ_j, and $\ln(\delta_j)$, where $\delta_j \equiv \mu_j - \mu_{j-1}$, in place of μ_j. Results are displayed both for the transformed parameters and for the original parameters.

The initial values passed to `ml` are obtained using the two-step estimation procedure described in section 2.2. To avoid passing `ml` infeasible or extreme feasible initial values, we follow the `heckman` implementation in censoring the initial $\widehat{\rho}_j$ into the range $[-0.85, 0.85]$ for all j.

We estimate a constant term for the second-step equation (2) but not for the first step (1) since the cutoffs $\mu_1, \mu_2, \ldots, \mu_J$ make a constant redundant. As a result, for binary selection our `oheckman` command produces results identical to those of the `heckman` command, but the output format differs because `heckman` reports a constant term for the first step, whereas `oheckman` reports a cutoff. This difference in output formats parallels the difference between the output formats of the `probit` and `oprobit` commands.

4 The oheckman command

4.1 Syntax

`oheckman` *depvar* [`=`] [*indepvars*] [*if*] [*in*] [*weight*] `,`
`select(`*categoryvar* [`=`] *indepvars_sel*`)` [`twostep robust cluster(`*varname*`)`
`level(`#`)` *maximize_options*]

`fweight`s, `pweight`s, and `iweight`s are allowed, but only if `twostep` is not specified.

4.2 Options

`select(`*categoryvar* [`=`] *indepvars_sel*`)` is required and specifies the categorical variable z and the independent variables $\mathbf{w}$ that determine z through an ordered probit as in (1).

`twostep` implements the two-step estimation procedure. The default is to use FIML.

`robust` computes robust estimates of variance. `robust` may not be used with `twostep`.

`cluster(`*varname*`)` adjusts standard errors for intragroup correlation. It implies `robust` and may not be used with `twostep`.

`level(`#`)` sets the level for confidence intervals, as a percentage. The default is `level(95)` or as set by `set level`.

maximize_options are passed directly to `ml` and control the maximization process. They are used only by the FIML algorithm and are rarely needed; see [R] **maximize**.

4.3 Syntax for predict

`predict` [*type*] *newvar* [*if*] [*in*] [`, xbsel xbif(`*j*`) mills psel(`*j*`)`
`millsif(`*j*`) yif(`*j*`)`]

4.4 Options for predict

`xbsel` computes $\widehat{z}_i^* = \widehat{\alpha}'\mathbf{w}_i$.

`xbif(`*j*`)` calculates $\widehat{\beta}_j'\mathbf{x}_i$.

`mills` returns $\widehat{\lambda}_i$ in (5), the estimate of the expected value of u_i given z_i and $\mathbf{w}_i$. This approach is the generalization of the Mills' ratio computed by `heckman`. However, unlike `heckman`, `oheckman` computes $\widehat{\lambda}_i$ for the actual value of the categorical variable z_i, not as if z_i were equal to 1. To get behavior similar to that of the `mills` statistic for `heckman`, use `millsif()`.

`psel(`j`)` estimates the probability that the categorical variable z_i would take on the value j, using the independent variables $\mathbf{w}_i$ in the selection equation.

`millsif(`j`)` estimates the expected value of u_i for each observation by using $\mathbf{w}_i$, under the assumption that the categorical variable z_i is equal to j for every observation.

`yif(`j`)` estimates the counterfactual $\widetilde{y}_j$ for the given equation j, if all observations were to switch to category j, *but taking into account the category that was actually chosen*. That is,

$$\widetilde{y}_j = \widehat{\beta}_j'\mathbf{x}_i + \widehat{\rho}_j\widehat{\sigma}_j\widehat{\lambda}_i$$

with $\widehat{\lambda}_i$ calculated as in (5), using the z_i actually chosen. This calculation differs from the `ycond` postestimation statistic for `heckman`, which computes $\widehat{\lambda}_i$ as if $z_i = j$ for all observations.

4.5 Saved results

In addition to the results returned by `ml`, the following program-specific results are saved:

Scalars			
`e(numeq)`	# of second-step equations	`e(ll_0)`	log likelihood if $\rho_j = 0$ for all j (FIML only)
`e(chi2_c)`	χ^2 for test $\rho_j = 0$ for all j (FIML only)	`e(p_c)`	p-value for test $\rho_j = 0$ for all j (FIML only)
Macros			
`e(x_sel)`	selection regressors $\mathbf{w}$	`e(x_reg)`	second-step regressors $\mathbf{x}$
`e(y_sel)`	categorical variable z	`e(y_reg)`	dependent variable y
`e(method)`	`ml` or `two-step`		
Matrices			
`e(cat)`	unique values of `e(y_sel)`	`e(cutoffs)`	estimates $\widehat{\mu}_1, \ldots, \widehat{\mu}_J$
`e(rho)`	estimates $\{\widehat{\rho}_j\}$	`e(sigma)`	estimates $\{\widehat{\sigma}_j\}$

5 Example

We illustrate the `oheckman` command by estimating wage equations in the public, private, and informal sectors for male workers in India. The categorical variable is `inf_prv_pub`, which takes on the value 1 for a worker in the informal sector, 2 for a worker in the private sector, and 3 for a worker in the public sector. Log wage is regressed against age, years of education, and a non–Hindu religion dummy. Household size and marital status dummies are used as extra regressors in the selection equation. The data come from the 55th round of India's National Sample Survey.

```
. use http://siteresources.worldbank.org/INTPOVRES/Resources/55th_short
. local selvar hhsize married widowed divorced
. local indvar age educ nonhindu
. oheckman log_realwage `indvar' [pw=weight] if sex==1,
> select(inf_prv_pub `selvar' `indvar')
Warning: Two-step initial estimate of rho2 = -.96588101 truncated to +/-.85.
Iteration 0:   log pseudolikelihood = -1.787e+08  (not concave)
Iteration 1:   log pseudolikelihood = -1.750e+08  (not concave)
Iteration 2:   log pseudolikelihood = -1.748e+08
  (output omitted)
Iteration 7:   log pseudolikelihood = -1.745e+08
Ordered probit selection model                  Number of obs   =     58349
                                                Wald chi2(7)    =  10630.71
Log pseudolikelihood = -1.745e+08               Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
inf_prv_pub  |
      hhsize |  -.0142523   .0024162    -5.90   0.000    -.0189879   -.0095166
     married |   .1307387   .0166229     7.86   0.000     .0981584     .163319
     widowed |  -.2637936   .0486937    -5.42   0.000    -.3592316   -.1683556
    divorced |  -.4980634   .1146349    -4.34   0.000    -.7227437   -.2733831
         age |   .0275979   .0007404    37.27   0.000     .0261466    .0290491
        educ |   .2499107    .002549    98.04   0.000     .2449148    .2549066
    nonhindu |   .1049969   .0189047     5.55   0.000     .0679443    .1420495
-------------+----------------------------------------------------------------
log_realwa~1 |
         age |   .0020197   .0003532     5.72   0.000     .0013275    .0027118
        educ |   .0226429   .0016513    13.71   0.000     .0194064    .0258794
    nonhindu |   .1488653   .0106898    13.93   0.000     .1279137     .169817
       _cons |   5.200081   .0144964   358.72   0.000     5.171668    5.228493
-------------+----------------------------------------------------------------
log_realwa~2 |
         age |   -.002199   .0021025    -1.05   0.296    -.0063198    .0019217
        educ |  -.1520386   .0119005   -12.78   0.000    -.1753631    -.128714
    nonhindu |  -.0391257   .0248368    -1.58   0.115     -.087805    .0095536
       _cons |   7.718006   .1660724    46.47   0.000      7.39251    8.043502
-------------+----------------------------------------------------------------
log_realwa~3 |
         age |   .0191914   .0014842    12.93   0.000     .0162823    .0221004
        educ |    .082742   .0059923    13.81   0.000     .0709973    .0944867
    nonhindu |  -.0166981   .0233705    -0.71   0.475    -.0625034    .0291073
       _cons |    5.48088    .136613    40.12   0.000     5.213124    5.748637
-------------+----------------------------------------------------------------
    /cutoff1 |   2.767841   .0339927    81.42   0.000     2.701217    2.834465
    /lndelta2 |  -.0922004    .011462    -8.04   0.000    -.1146656   -.0697352
    /athrho1 |  -.1011615   .0159755    -6.33   0.000    -.1324729   -.0698501
    /athrho2 |  -1.280433   .0576339   -22.22   0.000    -1.393393   -1.167473
    /athrho3 |  -.0403424   .0666513    -0.61   0.545    -.1709766    .0902918
   /lnsigma1 |  -.5630402   .0061925   -90.92   0.000    -.5751772   -.5509032
   /lnsigma2 |   .1307919   .0339859     3.85   0.000     .0641807    .1974031
   /lnsigma3 |  -.5955589   .0194662   -30.59   0.000     -.633712   -.5574057
```

```
     cutoff1 |   2.767841   .0339927                      2.701217    2.834465
     cutoff2 |   3.679763   .0345111                      3.612123    3.747404
        rho1 |  -.1008178   .0158131                     -.1317034   -.0697367
        rho2 |  -.8566002   .0153442                     -.8839152   -.8234603
        rho3 |  -.0403205    .066543                     -.1693298    .0900473
      sigma1 |   .5694751   .0035265                      .5626052    .5764289
      sigma2 |   1.139731   .0387348                      1.066285    1.218235
      sigma3 |   .5512544   .0107309                      .5306185    .5726929

Wald test of indep. eqns. (rho = 0): chi2(3) =   521.08   Prob > chi2 = 0.0000
```

The Wald test at the end of the output is a test of the null hypothesis $\rho_1 = \rho_2 = \rho_3 = 0$. If this hypothesis is true, then OLS is unbiased and there is no need to use a selection-bias correction model. Here the null hypothesis is strongly rejected.

After fitting the model, we can predict what wages would be in the public and private sectors. Then we estimate how much each public-sector employee gained by working in the public sector rather than in the private sector.

```
. predict public_log_wage if sex==1
(Option xbsel assumed; estimation of latent selection variable)
. predict private_log_wage if sex==1
(Option xbsel assumed; estimation of latent selection variable)
. predict informal_log_wage if sex==1
(Option xbsel assumed; estimation of latent selection variable)
. gen diff = public_log_wage - private_log_wage if sex==1 & inf_prv_pub==3
(62606 missing values generated)
```

6 Monte Carlo simulations

Many Monte Carlo studies have been done comparing the performance of the two-step estimator and FIML in the binary selection case considered by Heckman (1979). Surveying these studies, Puhani (2000) finds that FIML is usually more efficient than the two-step estimator. In this section, we describe the results of Monte Carlo simulations for the more general ordered-probit selection model.

6.1 Data-generating process

We generate x_{1i} and x_{2i} as independent standard normal random variables. Shocks u_i and ε_i are generated as standard bivariate normal with correlation ρ. The selection process is

$$\begin{aligned} z_i^* &= \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i; \\ z_i &= \begin{cases} 0 & \text{if } -\infty < z_i^* \leq -1, \\ 1 & \text{if } -1 < z_i^* \leq 1, \\ 2 & \text{if } 1 < z_i^* < \infty \end{cases} \end{aligned} \tag{9}$$

The dependent variable y_i is defined by

$$y_i = \begin{cases} \beta x_{1i} + \varepsilon_i & \text{if } z_i = j^*, \\ .\ (\text{missing}) & \text{otherwise} \end{cases} \tag{10}$$

where $j^* \in \{0, 1, 2\}$ is a parameter that we vary.

6.2 Baseline specification

Our baseline parameters in (9) and (10) are $\rho = 0.5$, and $\alpha_1 = \alpha_2 = \beta = 1$. Our regressors are $\mathbf{w}_i = [x_{1i}\ x_{2i}]'$ for the selection equation and $\mathbf{x}_i = [x_{1i}]$ for the main equation.

We report results for both $j^* = 0$ and $j^* = 1$. We test the performance of FIML and the two-step algorithm, as well as simple OLS (`regress`). We run 1,000 trials of every experiment, and we use the same datasets across the different methods (table 1).

Table 1: Results from first run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.9988 (0.0777)	94.8	1.0005 (0.0656)	95.1
Two-step	1,000	0.9989 (0.0797)	94.6	1.0003 (0.0656)	95.2
OLS	1,000	0.8360 (0.0669)	31.8	0.7871 (0.0525)	2.2
FIML	500	1.0000 (0.1097)	94.1	1.0029 (0.0959)	94.3
Two-step	500	1.0014 (0.1133)	93.9	1.0028 (0.0959)	94.2
FIML	300	0.9958 (0.1459)	93.1	0.9996 (0.1201)	94.9
Two-step	300	1.0006 (0.1482)	93.8	0.9989 (0.1197)	95.2
FIML	200	1.0008 (0.1836)	92.9	1.0055 (0.1505)	95.7
Two-step	200	1.0043 (0.1884)	93.7	1.0049 (0.1503)	95.5
FIML	100	0.9923 (0.2836)	89.9	1.0029 (0.2261)	93.0
Two-step	100	0.9957 (0.2738)	92.4	0.9981 (0.2202)	93.9
FIML	50	0.9965 (0.4243)	85.2	1.0436 (0.3442)	91.1
Two-step	50	1.0101 (0.4441)	91.1	1.0174 (0.3217)	93.7

The OLS estimator is clearly biased. FIML is slightly more efficient than the two-step estimator when $j^* = 0$, but FIML is not noticeably better than the two-step estimator when $j^* = 1$. As the sample size gets small, the coverage rate of the confidence intervals deteriorates, particularly for FIML. Robust confidence intervals for FIML have even worse coverage (results not shown).

6.3 High correlation of errors

For this experiment, we set $\rho = 0.99$ (table 2).

Table 2: Results from second run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.9992 (0.0500)	95.1	1.0021 (0.0465)	95.6
Two-step	1,000	0.9975 (0.0643)	94.9	1.0023 (0.0535)	95.1
OLS	1,000	0.6734 (0.0573)	0.0	0.5807 (0.0432)	0.0

The relative efficiency of FIML over the two-step method increases with $|\rho|$. OLS also becomes more biased as $|\rho|$ increases.

6.4 Multiple equations

An advantage of FIML over the two-step estimator appears to be that FIML uses the value of y_i across all equations when estimating β_j for one particular equation j. However, the feedback across equations is only indirect through the selection-equation parameters α and $\{\mu_j\}$.

To test whether the simultaneous estimation of all equations improves the FIML estimator, we replace (10) with

$$y_i = \beta x_{1i} + \varepsilon_i \text{ always, regardless of } z_i$$

Doing so results in the estimation of three separate equations, for $j = 0$, 1, and 2.[3] We report the estimate $\widehat{\beta}$ for equation j^* (table 3), and we report results only for FIML since the two-step results are the same as in the baseline case.

3. In real applications, β_j and/or ρ_j would vary with j, since otherwise there is no need to use a selection model. We chose β and ρ to be independent of j for simplicity of presentation, and this choice has little effect on the simulation results.

Table 3: Results from third run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.9988 (0.0778)	94.7	1.0004 (0.0653)	95.3
FIML	500	0.9999 (0.1097)	94.0	1.0026 (0.0960)	94.1
FIML	300	0.9960 (0.1459)	93.1	0.9994 (0.1200)	95.1
FIML	200	1.0013 (0.1837)	93.1	1.0065 (0.1511)	95.2
FIML	100	0.9931 (0.2836)	90.3	1.0042 (0.2301)	92.7
FIML	50	0.9969 (0.4255)	85.2	1.0403 (0.3459)	91.4

Comparing these results to the baseline case shows that including all the equations results in no noticeable improvement for FIML.

6.5 Nonnormal shocks

We modify the shocks u_i, ε_i to be nonnormal by squaring them (table 4).

Table 4: Results from fourth run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.9568 (0.1799)	84.2	1.0002 (0.0850)	95.1
Two-step	1,000	0.9956 (0.1437)	94.8	1.0021 (0.0861)	95.0
OLS	1,000	0.9722 (0.1316)	93.8	0.9422 (0.0784)	88.6

For $j^* = 0$, the two-step estimator handles the nonnormality well, but FIML is biased and has poor coverage. FIML makes full use of the assumption of joint normality of the shocks, so it makes more mistakes when the shocks are not normal. The two-step estimator is actually consistent even if the second-step shocks ε_{ij} are not normally distributed.

For $j^* = 1$, both FIML and the two-step estimator handle the nonnormality well, since the near-linearity of $\widehat{\lambda}_i$ in figure 1 (ii) makes FIML nearly equivalent to the two-step estimator.

6.6 No exclusion restriction

For this experiment, we let $\mathbf{w}_i = [x_{1i}]$ (and hence $\alpha_2 = 0$), so that there is no exclusion restriction because $\mathbf{w}_i = \mathbf{x}_i$. As discussed in section 2.4, the identification for β here depends entirely on the weak nonlinearity of $\widehat{\lambda}_i$, so we expect our estimators to have some trouble.

The precision of the two-step estimator can be improved by throwing out trials for which the estimated $\widehat{\rho}$ is infeasible ($|\widehat{\rho}| > 1$). We report this version of the estimator as Two-step* (table 5).

Table 5: Results from fifth run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.8784 (0.2916)	79.8	0.7370 (0.5426)	64.8
Two-step	1,000	0.9946 (0.4394)	97.9	1.0768 (2.6304)	95.0
Two-step*	1,000	0.9314 (0.3705)	98.0	0.7016 (0.7161)	100.0
OLS	1,000	0.6893 (0.0765)	1.7	0.6342 (0.0519)	0.0

As can be seen from comparing curves (i) and (ii) in figure 1, the nonlinearity of $\widehat{\lambda}_i$ is much weaker for $j^* = 1$ than for $j^* = 0$, and this explains why the results are much worse for $j^* = 1$.

The FIML estimates are much tighter than the two-step estimates, but they are biased. Throwing out infeasible estimates greatly improves the precision of the two-step estimator but introduces bias.

6.7 Exclusion restriction not satisfied

We replace (10) with

$$y_i = \begin{cases} \beta x_{1i} + x_{2i} + \varepsilon_i & \text{if } z_i = j^*, \\ .\ (\text{missing}) & \text{otherwise} \end{cases}$$

so that the exclusion restriction on x_2 is not satisfied (table 6).

Table 6: Results from sixth run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.1350 (0.0805)	0.0	0.0030 (0.0706)	0.0
Two-step	1,000	0.0887 (0.0887)	0.0	−0.0014 (0.0720)	0.0
OLS	1,000	0.5008 (0.0786)	0.0	0.3621 (0.0593)	0.0

The estimates are all tight, yet far from the true $\beta = 1$. This finding highlights the importance of having a valid exclusion restriction.

6.8 Weak instrument

We change $\alpha_2 = 0$, so that the instrument x_2 used for identification is invalid. However, identification can still be achieved with x_1 alone because of nonlinearity, but this nonlinearity is weak, as discussed in section 2.4 (table 7).

Table 7: Results from seventh run

Method	No. of obs./trial	$\widehat{\beta}$ for $j^* = 0$, mean (SD)	95% Conf. coverage (%)	$\widehat{\beta}$ for $j^* = 1$, mean (SD)	95% Conf. coverage (%)
FIML	1,000	0.8794 (0.2925)	79.7	0.6902 (0.5411)	62.0
Two-step	1,000	0.9770 (0.4198)	97.4	1.7948 (1.6622)	98.9
Two-step*	1,000	0.9233 (0.3590)	97.4	0.6775 (0.6477)	99.5
OLS	1,000	0.6893 (0.0765)	1.7	0.6342 (0.0519)	0.0

These results are similar to those for no exclusion restriction, which makes sense since in both cases the identification comes only from x_1.

6.9 Summary

The FIML estimator is slightly more efficient than the two-step estimator when the data exactly meet the model specifications and especially when $|\rho|$ is high. However, the FIML confidence intervals have poor coverage rates for small sample sizes, and FIML performs poorly when the shocks are not normal. Therefore, the two-step estimator is more robust and appears to be the better choice for almost all practical applications.

Both estimators perform poorly when there is no exclusion restriction imposed or when the exclusion restriction is not satisfied.

7 References

Bellemare, M. F., and C. B. Barrett. 2006. An ordered tobit model of market participation: Evidence from Kenya and Ethiopia. *American Journal of Agricultural Economics* 88: 324–337.

Carlsson, F. 2004. Prices and departures in European domestic aviation markets. *Review of Industrial Organization* 24: 37–49.

Ermisch, J. F., and R. E. Wright. 1993. Wage offers and full-time and part-time employment by British women. *Journal of Human Resources* 28: 111–133.

Greene, W. H. 2002. *LIMDEP Version 8.0 Econometric Modeling Guide*, vol. 2. Plainview, NY: Econometric Software.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–162.

Idson, T. L., and D. J. Feaster. 1990. A selectivity model of employer-size wage differentials. *Journal of Labor Economics* 8: 99–122.

Jimenez, E., and B. Kugler. 1987. The earnings impact of training duration in a developing country. *Journal of Human Resources* 22: 228–247.

Main, B. G. M., and B. Reilly. 1993. The employer size–wage gap: Evidence for Britain. *Economica* 60: 125–142.

Miranda, A., and S. Rabe-Hesketh. 2006. Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata Journal* 6: 285–308.

Paci, P., H. Joshi, G. Makepeace, and P. Dolton. 1995. Is pay discrimination against young women a thing of the past? A tale of two cohorts. *International Journal of Manpower* 16: 60–65.

Puhani, P. A. 2000. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* 14: 53–68.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal* 2: 1–21.

About the authors

Richard Chiburis is a Ph.D. candidate in economics at Princeton University.

Michael Lokshin is a senior economist at the Development Economics Research Group of the World Bank.