Title

Example 45g — Heckman selection model

Description Remarks and examples References Also see

Description

To demonstrate selection models, we will use the following data:

```
. use https://www.stata-press.com/data/r17/gsem_womenwk
(Fictional data on women and work)

. summarize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         age |      2,000      36.208     8.28656         20         59
        educ |      2,000      13.084    3.045912         10         20
     married |      2,000       .6705    .4701492          0          1
    children |      2,000      1.6445    1.398963          0          5
        wage |      1,343    23.69217    6.305374    5.88497   45.80979

. notes

_dta:
  1.  Fictional data on 2,000 women, 1,343 of whom work.
  2.  age ....... age in years
  3.  educ ...... years of schooling
  4.  married ... 1 if married spouse present
  5.  children .. # of children under 12 years
  6.  wage ...... hourly wage (missing if not working)
```

See *Structural models 8: Dependencies between response variables* and *Structural models 9: Unobserved inputs, outputs, or both* in [SEM] **Intro 5** for background.

Remarks and examples

stata.com

Remarks are presented under the following headings:

The Heckman selection model as an SEM

We demonstrate below how `gsem` can be used to fit the Heckman selection model (Gronau 1974; Lewis 1974; Heckman 1976) and produce results comparable to those of Stata's dedicated `heckman` command; see [R] **heckman**.

Our purpose is not to promote `gsem` as an alternative to `heckman`. We have two other purposes.

One is to show that `gsem` can be used to generalize the Heckman selection model to response functions other than linear and, in addition or separately, to include multilevel effects when such effects are present.

The other is to show how Heckman selection models can be included in more complicated SEMs.

For those unfamiliar with this model, it deals with a continuous outcome that is observed only when another equation determines that the observation is selected, and the errors of the two equations are allowed to be correlated. Subjects often choose to participate in an event or medical trial or even the labor market, and thus the outcome of interest might be correlated with the decision to participate. Heckman won a Nobel Prize for this work.

The model is sometimes cast in terms of female labor supply, but it obviously has broader application. Nevertheless, we will consider a female labor-supply example.

Women are offered employment at a wage of w,

$$w_i = \mathbf{X}_i \boldsymbol{\beta} + \epsilon_i$$

Not all women choose to work, and w is observed only for those women who do work. Women choose to work if

$$\mathbf{Z}_i \gamma + \xi_i > 0$$

where

$$\begin{aligned} \epsilon_i &\sim N(0, \sigma^2) \\ \xi_i &\sim N(0, 1) \\ \text{corr}(\epsilon, \xi) &= \rho \end{aligned}$$

More generally, we can think of this model as applying to any continuously measured outcome w_i, which is observed only if $\mathbf{Z}_i \gamma + \xi_i > 0$. The important feature of the model is that the errors ξ_i of the selection equation and the errors ϵ_i of the observed-data equation are allowed to be correlated.

The Heckman selection model can be recast as a two-equation SEM—one linear regression (for the continuous outcome) and the other censored regression (for selection)—and with a latent variable L_i added to both equations. The latent variable is constrained to have variance 1 and to have coefficient 1 in the selection equation, leaving only the coefficient in the continuous-outcome equation to be estimated. For identification, the variance from the censored regression will be constrained to be equal to that of the linear regression. The results of doing this are the following:

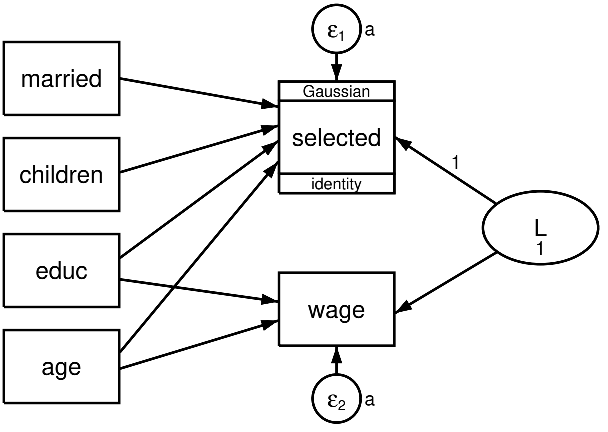
1. Latent variable L_i becomes the vehicle for carrying the correlation between the two equations.
2. All the parameters given above, namely, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2, and ρ, can be recovered from the SEM estimates.
3. If we call the estimated parameters in the SEM formulation $\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}^*$, and ${\sigma^2}^*$, and let κ denote the coefficient on L_i in the continuous-outcome equation, then

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{\beta}^* \\ \gamma &= \gamma^* / \sqrt{{\sigma^2}^* + 1} \\ \sigma^2 &= {\sigma^2}^* + \kappa^2 \\ \rho &= \kappa / \sqrt{({\sigma^2}^* + \kappa^2)({\sigma^2}^* + 1)} \end{aligned}$$

   This parameterization places no restriction on the range or sign of ρ. See Skrondal and Rabe-Hesketh (2004, 107–108).

Fitting the Heckman selection model as an SEM

We wish to fit the following Heckman selection model:



What makes this a Heckman selection model is

1. the inclusion of latent variable L in both the continuous-outcome (wage) equation and the censored-outcome selection equation;
2. constraining the selected <- L path coefficient to be 1;
3. constraining the variance of L to be 1; and
4. constraining the error variances to be equal.

Before we can fit this model, we need to create new variables selected and notselected. selected will equal 0 if the woman works (wage is not missing) and missing otherwise. notselected is the complement of selected: it equals 0 if the woman does not work (wage is missing) and missing otherwise. selected and notselected will be used as the dependent variables in the censored regression, providing the equivalent of a scaled probit regression.

```
. generate selected = 0 if wage < .
(657 missing values generated)
. generate notselected = 0 if wage >= .
(1,343 missing values generated)
. tabulate selected notselected, missing

           |      notselected
  selected |         0          . |     Total
-----------+----------------------+----------
         0 |         0      1,343 |     1,343
         . |       657          0 |       657
-----------+----------------------+----------
     Total |       657      1,343 |     2,000
```

Old-time Stata users may be worried that because wage is missing in so many observations, namely, all those corresponding to nonworking women, there must be something special we need to do so that gsem uses all the data. There is nothing special we need to do. gsem counts missing values on an equation-by-equation basis, so it will use all the data for the censored regression part of the model while simultaneously using only the working-woman subsample for the continuous-outcome (wage) part of the model. We use all the data for the censored regression because gsem understands the meaning of missing values in the censored dependent variables so long as one of them is nonmissing.

To fit this model in command syntax, we type

```
. gsem (wage <- educ age  L)
>  (selected <- married children educ age L@1,
>   family(gaussian, udepvar(notselected))), var(L@1 e.wage@a e.selected@a)
Fitting fixed-effects model:
Iteration 0:   log likelihood = -5752.6506
Iteration 1:   log likelihood = -5260.9961
Iteration 2:   log likelihood = -5209.2571
Iteration 3:   log likelihood = -5208.9039
Iteration 4:   log likelihood = -5208.9038
Refining starting values:
Grid node 0:   log likelihood = -5208.7006
Fitting full model:
Iteration 0:   log likelihood = -5208.5322  (not concave)
Iteration 1:   log likelihood = -5208.0269
Iteration 2:   log likelihood =  -5202.872  (not concave)
Iteration 3:   log likelihood = -5202.0258
Iteration 4:   log likelihood = -5198.6178
Iteration 5:   log likelihood = -5193.0576
Iteration 6:   log likelihood = -5191.8655
Iteration 7:   log likelihood = -5178.5058
Iteration 8:   log likelihood = -5178.3095
Iteration 9:   log likelihood = -5178.3046
Iteration 10:  log likelihood = -5178.3046
Generalized structural equation model                Number of obs     = 2,000
Response:       wage                                 Number of obs     = 1,343
Family:         Gaussian
Link:           Identity
Lower response: selected                             Number of obs     = 2,000
Upper response: notselected                                Uncensored =      0
Family:         Gaussian                                Left-censored =    657
Link:           Identity                               Right-censored = 1,343
                                                       Interval-cens. =      0
Log likelihood = -5178.3046
 ( 1)  [selected]L = 1
 ( 2)  - [/]var(e.wage) + [/]var(e.selected) = 0
 ( 3)  [/]var(L) = 1
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
wage						
educ	.9899512	.0532552	18.59	0.000	.8855729	1.094329
age	.2131282	.020602	10.35	0.000	.172749	.2535074
L	5.923736	.1846818	32.08	0.000	5.561767	6.285706
_cons	.4859114	1.076865	0.45	0.652	-1.624705	2.596528
selected						
married	.6242746	.1054319	5.92	0.000	.4176319	.8309173
children	.6152095	.0652002	9.44	0.000	.4874196	.7429995
educ	.0781542	.0162868	4.80	0.000	.0462327	.1100757
age	.0511983	.006637	7.71	0.000	.0381901	.0642066
L	1	(constrained)				
_cons	-3.493217	.3730379	-9.36	0.000	-4.224357	-2.762076
var(L)	1	(constrained)				
var(e.wage)	.9664635	.2689653			.560141	1.66753
var(e.sele~d)	.9664635	.2689653			.560141	1.66753

Notes:

1. Some of the estimated coefficients and parameters above will match those reported by the `heckman` command and others will not. The above parameters are in the transformed structural equation modeling metric. That metric can be transformed back to the Heckman metric and results will match. The relationship to the Heckman metric is

$$\boldsymbol{\beta} = \boldsymbol{\beta}^*$$
$$\gamma = \gamma^* / \sqrt{\sigma^{2^*} + 1}$$
$$\sigma^2 = \sigma^{2^*} + \kappa^2$$
$$\rho = \kappa / \sqrt{(\sigma^{2^*} + \kappa^2)(\sigma^{2^*} + 1)}$$

2. $\boldsymbol{\beta}$ refers to the coefficients on the continuous-outcome (wage) equation. We can read those coefficients directly, without transformation except that we ignore the `wage <- L` path:

$$\texttt{wage} = 0.9900\,\texttt{educ} + 0.2131\,\texttt{age} + 0.4859$$

3. γ refers to the selection equation, and because $\gamma = \gamma^* / \sqrt{\sigma^{2^*} + 1}$, we must divide the reported coefficients by the square root of $\sigma^{2^*} + 1$. What has happened here is that the scaled probit has variance $\sigma^{2^*} + 1$, and we are merely transforming back to the standard probit model, which has variance 1. The results are

$$\Pr(\texttt{selected} = 0) =$$
$$\Phi(0.4452\,\texttt{married} + 0.4387\,\texttt{children} + 0.0557\,\texttt{educ} + 0.0365\,\texttt{age} - 2.4910)$$

4. To calculate ρ, we first calculate $\sigma^2 = \sigma^{2^*} + \kappa^2$ and then calculate $\rho = \kappa / \sqrt{\sigma^2(\sigma^{2^*} + 1)}$:

$$\sigma^2 = 0.9664 + 5.9237^2 = 36.0571$$
$$\rho = 5.9237 / \sqrt{\sigma^2(.9664 + 1)} = 0.7035$$

5. These transformed results match the results that would have been reported had we typed

```
. heckman wage educ age, select(married children educ age)
  (output omitted)
```

6. There is an easier way to obtain the transformed results than by hand, and the easier way provides standard errors. That is the subject of the next section.

Transforming results and obtaining rho

We can use Stata's `nlcom` command to perform the transformations we made by hand above, and we can obtain standard errors.

Let's start by obtaining σ^2 and ρ. To remind you, the formulas are

$$\sigma^2 = \sigma^{2^*} + \kappa^2$$
$$\rho = \kappa / \sqrt{\sigma^2(\sigma^{2^*} + 1)}$$

We must describe these two formulas in a way that `nlcom` can understand. The Stata notation for parameters σ^{2^*} and κ fit by `gsem` is

σ^{2^*}: `_b[/var(e.wage)]`

κ: `_b[wage:L]`

We cannot remember that notation; however, we can type `gsem, coeflegend` to be reminded. We now have all that we need to obtain the estimates of σ^2 and ρ. Because `heckman` reports σ rather than σ^2, we will tell `nlcom` to report the `sqrt(`σ^2`)`:

```
. nlcom (sigma: sqrt(_b[/var(e.wage)] +_b[wage:L]^2))
>       (rho: _b[wage:L]/(sqrt((_b[/var(e.wage)]+1)*(_b[/var(e.wage)]
> + _b[wage:L]^2))))
     sigma: sqrt(_b[/var(e.wage)] +_b[wage:L]^2)
       rho: _b[wage:L]/(sqrt((_b[/var(e.wage)]+1)*(_b[/var(e.wage)]
> + _b[wage:L]^2)))
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
sigma	6.004758	.1656471	36.25	0.000	5.680095	6.32942
rho	.703489	.0511861	13.74	0.000	.603166	.8038119

The output above nearly matches what `heckman` reports. `heckman` does not report the test statistics and p-values for these two parameters. In addition, the confidence interval that `heckman` reports for ρ will differ slightly from the above and is better. `heckman` uses a method that will not allow ρ to be outside of -1 and 1, whereas `nlcom` is simply producing a confidence interval for the calculation we requested and in absence of the knowledge that the calculation corresponds to a correlation coefficient. The same applies to the confidence interval for σ, where the bounds are 0 and infinity.

To obtain the coefficients and standard errors for the selection equation, we type

```
. nlcom (married:  _b[selected:married]/sqrt(_b[/var(e.wage)]+1))
>       (children: _b[selected:children]/sqrt(_b[/var(e.wage)]+1))
>       (educ: _b[selected:educ]/sqrt(_b[/var(e.wage)]+1))
>       (age: _b[selected:age]/sqrt(_b[/var(e.wage)]+1))
   married: _b[selected:married]/sqrt(_b[/var(e.wage)]+1)
  children: _b[selected:children]/sqrt(_b[/var(e.wage)]+1)
      educ: _b[selected:educ]/sqrt(_b[/var(e.wage)]+1)
       age: _b[selected:age]/sqrt(_b[/var(e.wage)]+1)
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
married	.445177	.0673953	6.61	0.000	.3130847	.5772693
children	.4387126	.0277788	15.79	0.000	.3842671	.4931581
educ	.0557326	.0107348	5.19	0.000	.0346927	.0767725
age	.0365101	.0041534	8.79	0.000	.0283696	.0446505

The above output matches what `heckman` reports.

Fitting the model with the Builder

Use the diagram in *Fitting the Heckman selection model as an SEM* above for reference.

1. Open the dataset and create the selection variable.

   In the Command window, type

   ```
   . use https://www.stata-press.com/data/r17/gsem_womenwk
   . generate selected = 0 if wage < .
   . generate notselected = 0 if wage >= .
   ```

2. Open a new Builder diagram.

   Select menu item **Statistics > SEM (structural equation modeling) > Model building and estimation**.

3. Put the Builder in gsem mode by clicking on the G/SEM button.
4. Create the independent variables.

   Select the Add observed variables set tool, , and then click in the diagram about one-fourth of the way in from the left and one-fourth of the way up from the bottom.

   In the resulting dialog box,

   a. select the *Select variables* radio button (it may already be selected);
   b. use the *Variables* control to select the variables `married`, `children`, `educ`, and `age` in this order;
   c. select `Vertical` in the *Orientation* control;
   d. click on **OK**.

   If you wish, move the set of variables by clicking on any variable and dragging it.
5. Create the generalized response for selection.
   a. Select the Add generalized response variable tool, .
   b. Click about one-third of the way in from the right side of the diagram, to the right of the `married` rectangle.
   c. In the Contextual Toolbar, select `Gaussian, Identity` in the *Family/Link* control (it may already be selected).
   d. In the Contextual Toolbar, select `selected` in the *Variable* control.
   e. In the Contextual Toolbar, click on the **Properties...** button.
   f. In the resulting *Variable properties* dialog box, click on the **Censoring...** button in the **Variable** tab.
   g. In the resulting *Censoring* dialog box, select the *Interval-measured, depvar is lower boundary* radio button. In the resulting *Interval-measured* box below, use the *Upper bound* control to select the variable `notselected`.
   h. Click on **OK** in the *Censoring* dialog box, and then click on **OK** in the *Variable properties* dialog box. The Details pane will now show `selected` as the lower bound and `notselected` as the upper bound of our interval measure.
6. Create the endogenous `wage` variable.
   a. Select the Add observed variable tool, , and then click about one-third of the way in from the right side of the diagram, to the right of the `age` rectangle.
   b. In the Contextual Toolbar, select `wage` with the *Variable* control.
7. Create paths from the independent variables to the dependent variables.
   a. Select the Add path tool, →.
   b. Click in the right side of the `married` rectangle (it will highlight when you hover over it), and drag a path to the left side of the `selected` rectangle (it will highlight when you can release to connect the path).
   c. Continuing with the → tool, create the following paths by clicking first in the right side of the rectangle for the independent variable and dragging it to the left side of the rectangle for the dependent variable:

```
children -> selected
educ -> selected
age -> selected
educ -> wage
age -> wage
```

8. Clean up the direction of the error terms.

   We want the error for `selected` to be above the rectangle and the error for `wage` to be below the rectangle, but it is likely they have been created in other directions.

   a. Choose the Select tool, .
   b. Click in the `selected` rectangle.
   c. Click on one of the **Error rotation** buttons, , in the Contextual Toolbar until the error is above the rectangle.
   d. Click in the `wage` rectangle.
   e. Click on one of the **Error rotation** buttons, , in the Contextual Toolbar until the error is below the rectangle.

9. Create the latent variable.

   a. Select the Add latent variable tool, , and then click at the far right of the diagram and vertically centered between the `selected` and `wage` variables.
   b. In the Contextual Toolbar, type `L` in the *Name* control and press *Enter*.

10. Draw paths from the latent variable to each endogenous variable.

    a. Select the Add path tool, .
    b. Click in the upper left quadrant of the `L` oval, and drag a path to the right side of the `selected` rectangle.
    c. Continuing with the tool, create another path by clicking first in the lower-left quadrant of the `L` oval and dragging a path to the right side of the `wage` rectangle.

11. Place constraints on the variances and on the path from `L` to `selected`.

    a. Choose the Select tool, .
    b. Click on the `L` oval. In the Contextual Toolbar, type `1` in the σ^2 box and press *Enter*.
    c. Click on the error oval attached to the `wage` rectangle. In the Contextual Toolbar, type `a` in the σ^2 box and press *Enter*.
    d. Click on the error oval attached to the `selected` rectangle. In the Contextual Toolbar, type `a` in the σ^2 box and press *Enter*.
    e. Click on the path from `L` to `selected`. In the Contextual Toolbar, type `1` in the β box and press *Enter*.

12. Clean up the location of the paths.

    If you do not like where a path has been connected to its variables, use the Select tool, , to click on the path, and then simply click on where it connects to a rectangle and drag the endpoint.

13. Estimate.

    Click on the **Estimate** button, , in the Standard Toolbar, and then click on **OK** in the resulting *GSEM estimation options* dialog box.

You can open a completed diagram in the Builder by typing

```
. webgetsem gsem_select
```

References

Gronau, R. 1974. Wage comparisons: A selectivity bias. *Journal of Political Economy* 82: 1119–1143. https://doi.org/10.1086/260267.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–492.

Lewis, H. G. 1974. Comments on selectivity biases in wage comparisons. *Journal of Political Economy* 82: 1145–1155.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Also see

[SEM] **Example 34g** — Combined models (generalized responses)

[SEM] **Example 46g** — Endogenous treatment-effects model

[SEM] **Intro 5** — Tour of models

[SEM] **gsem** — Generalized structural equation model estimation command