Description

heckman fits regression models with selection by using either Heckman’s two-step consistent estimator or full maximum likelihood.

Quick start

Heckman model of $y$ on $x_1$ with $v_1$ predicting selection when binary variable selected indicates selection status

```
heckman y x1, select(selected = v1 x1)
```

As above, and generate $v_2$ containing the inverse Mills’s ratio

```
heckman y x1, select(selected = v1 x1) mills(v2)
```

Same as above

```
heckman y x1, select(selected = v1 x1) nshazard(v2)
```

Fit a Heckman model using the two-step estimation method

```
heckman y x1, select(selected = v1 x1) twostep
```

As above, and display first-stage probit estimates

```
heckman y x1, select(selected = v1 x1) twostep first
```

Menu

Statistics > Sample-selection models > Heckman selection model
Syntax

Basic syntax

```
heckman depvar [indepvars], select(varlist) [twostep]
```

or

```
heckman depvar [indepvars], select(depvars = varlist) [twostep]
```

Full syntax for maximum likelihood estimates only

```
heckman depvar [indepvars] [if] [in] [weight],
    select([depvars =] varlist [, noconstant offset(varname)])
        [heckman_ml_options]
```

Full syntax for Heckman’s two-step consistent estimates only

```
heckman depvar [indepvars] [if] [in], twostep
    select([depvars =] varlist [, noconstant]) [heckman_ts_options]
```
**heckman — Heckman selection model**

<table>
<thead>
<tr>
<th>heckman_ml_options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td></td>
</tr>
<tr>
<td><em>mle</em></td>
<td>use maximum likelihood estimator; the default</td>
</tr>
<tr>
<td>select()</td>
<td>specify selection equation: dependent and independent variables; whether to have constant term and offset variable</td>
</tr>
<tr>
<td><strong>noconstant</strong></td>
<td>suppress constant term</td>
</tr>
<tr>
<td>offset(varname)</td>
<td>include varname in model with coefficient constrained to 1</td>
</tr>
<tr>
<td>constraints(constraints)</td>
<td>apply specified linear constraints</td>
</tr>
<tr>
<td><strong>SE/Robust</strong></td>
<td></td>
</tr>
<tr>
<td>vce(vcetype)</td>
<td>vcetype may be oim, robust, cluster clustvar, opg, bootstrap, or jackknife</td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
<td></td>
</tr>
<tr>
<td>level(#)</td>
<td>set confidence level; default is level(95)</td>
</tr>
<tr>
<td>first</td>
<td>report first-step probit estimates</td>
</tr>
<tr>
<td>lrmodel</td>
<td>perform the likelihood-ratio model test instead of the default Wald test</td>
</tr>
<tr>
<td>nshazard(newvar)</td>
<td>generate nonselection hazard variable</td>
</tr>
<tr>
<td>mills(newvar)</td>
<td>synonym for nshazard()</td>
</tr>
<tr>
<td>nocnsreport</td>
<td>do not display constraints</td>
</tr>
<tr>
<td>display_options</td>
<td>control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling</td>
</tr>
<tr>
<td><strong>Maximization</strong></td>
<td></td>
</tr>
<tr>
<td>maximize_options</td>
<td>control the maximization process; seldom used</td>
</tr>
<tr>
<td>collinear</td>
<td>keep collinear variables</td>
</tr>
<tr>
<td>coeflegend</td>
<td>display legend instead of statistics</td>
</tr>
</tbody>
</table>

*select() is required.

The full specification is `select([ depvar s = ] varlist s [ , noconstant offset(varname o) ] )`. 
**heckman ts_options**

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>twostep</em></td>
<td>produce two-step consistent estimate</td>
</tr>
<tr>
<td><em>select()</em></td>
<td>specify selection equation: dependent and independent variables; whether to have constant term</td>
</tr>
<tr>
<td>noconstant</td>
<td>suppress constant term</td>
</tr>
<tr>
<td>rhosigma</td>
<td>truncate $\rho$ to $[-1,1]$ with consistent $\sigma$</td>
</tr>
<tr>
<td>rhotrunc</td>
<td>truncate $\rho$ to $[-1,1]$</td>
</tr>
<tr>
<td>rholimited</td>
<td>truncate $\rho$ in limited cases</td>
</tr>
<tr>
<td>rhoforce</td>
<td>do not truncate $\rho$</td>
</tr>
</tbody>
</table>

**SE**

vce(vcetype)

vcetype may be conventional, bootstrap, or jackknife

**Reporting**

level(#) set confidence level; default is level(95)

first report first-step probit estimates

nshazard(newvar) generate nonselection hazard variable

mills(newvar) synonym for nshazard()

display_options control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling

do not truncate $\rho$

**Options for Heckman selection model (ML)**

mle requests that the maximum likelihood estimator be used. This is the default.

select([*depvar_s =*] varlist_s [, noconstant offset(varname_o)]) specifies the variables and options for the selection equation. It is an integral part of specifying a Heckman model and is required. The selection equation should contain at least one variable that is not in the outcome equation.
If \emph{depvar}s is specified, it should be coded as 0 or 1, with 0 indicating an observation not selected and 1 indicating a selected observation. If \emph{depvar}s is not specified, observations for which \emph{depvar} is not missing are assumed selected, and those for which \emph{depvar} is missing are assumed not selected.

\texttt{noconstant} suppresses the selection constant term (intercept).

\texttt{offset(varname)} specifies that selection offset \emph{varname} be included in the model with the coefficient constrained to be 1.

\texttt{noconstant, offset(varname), constraints(constraints)}; see [R] \emph{Estimation options}.

\texttt{vce(vcetype)} specifies the type of standard error reported, which includes types that are derived from asymptotic theory (\emph{oim}, \emph{opg}), that are robust to some kinds of misspecification (\emph{robust}), that allow for intragroup correlation (\emph{cluster clustvar}), and that use bootstrap or jackknife methods (\emph{bootstrap}, \emph{jackknife}); see [R] \emph{vce option}.

\texttt{level(#)}; see [R] \emph{Estimation options}.

\texttt{first} specifies that the first-step probit estimates of the selection equation be displayed before estimation.

\texttt{lrmodel}; see [R] \emph{Estimation options}.

\texttt{nshazard(newvar) and mills(newvar)} are synonyms; either will create a new variable containing the nonselection hazard—what Heckman (1979) referred to as the inverse of the Mills ratio—from the selection equation. The nonselection hazard is computed from the estimated parameters of the selection equation.

\texttt{nocnsreport}; see [R] \emph{Estimation options}.

\texttt{display_options: noci, nopvalues, noomitted, vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), fvwrapon(style), cformat(\%fmt), pformat(\%fmt), sformat(\%fmt), and nolstretch}; see [R] \emph{Estimation options}.

\texttt{maximize_options: difficult, technique(algorithm_spec), iterate(#), [no]log, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), nrtolerance(#), nonrtolerance, and from(init_specs)}; see [R] \emph{Maximize}. These options are seldom used.

Setting the optimization type to \texttt{technique(bhhh)} resets the default \emph{vcetype} to \texttt{vce(opg)}.

The following options are available with \texttt{heckman} but are not shown in the dialog box:
\texttt{collinear}, \texttt{coeflegend}; see [R] \emph{Estimation options}.

**Options for Heckman selection model (two-step)**

\texttt{twostep} specifies that Heckman’s (1979) two-step efficient estimates of the parameters, standard errors, and covariance matrix be produced.
select([ depvar_s = ] varlist_s [, noconstant ] ) specifies the variables and options for the selection equation. It is an integral part of specifying a Heckman model and is required. The selection equation should contain at least one variable that is not in the outcome equation.

If depvar_s is specified, it should be coded as 0 or 1, with 0 indicating an observation not selected and 1 indicating a selected observation. If depvar_s is not specified, observations for which depvar is not missing are assumed selected, and those for which depvar is missing are assumed not selected.

noconstant suppresses the selection constant term (intercept).

noconstant; see [R] Estimation options.

rhosigma, rhotrunc, rholimited, and rhoforce are rarely used options to specify how the two-step estimator (option twostep) handles unusual cases in which the two-step estimate of \( \rho \) is outside the admissible range for a correlation, \([-1, 1]\). When \( \text{abs}(\rho) > 1 \), the two-step estimate of the coefficient variance–covariance matrix may not be positive definite and thus may be unusable for testing. The default is rhosigma.

rhosigma specifies that \( \rho \) be truncated, as with the rhotrunc option, and that the estimate of \( \sigma \) be made consistent with \( \hat{\rho} \), the truncated estimate of \( \rho \). So, \( \hat{\sigma} = \beta_m \hat{\rho} \); see Methods and formulas for the definition of \( \beta_m \). Both the truncated \( \rho \) and the new estimate of \( \hat{\sigma} \) are used in all computations to estimate the two-step covariance matrix.

rhotrunc specifies that \( \rho \) be truncated to lie in the range \([-1, 1]\). If the two-step estimate is less than \(-1 \), \( \rho \) is set to \(-1 \); if the two-step estimate is greater than \(1 \), \( \rho \) is set to \(1 \). This truncated value of \( \rho \) is used in all computations to estimate the two-step covariance matrix.

rholimited specifies that \( \rho \) be truncated only in computing the diagonal matrix \( D \) as it enters \( V_{\text{twostep}} \) and \( Q \); see Methods and formulas. In all other computations, the untruncated estimate of \( \rho \) is used.

rhoforce specifies that the two-step estimate of \( \rho \) be retained, even if it is outside the admissible range for a correlation. This option may, in rare cases, lead to a non–positive-definite covariance matrix.

These options have no effect when estimation is by maximum likelihood, the default. They also have no effect when the two-step estimate of \( \rho \) is in the range \([-1, 1]\).

vce(vcetype) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (conventional) and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] vce_option.

vce(conventional), the default, uses the two-step variance estimator derived by Heckman.

level(#); see [R] Estimation options.

first specifies that the first-step probit estimates of the selection equation be displayed before estimation.

nshazard(newvar) and mills(newvar) are synonyms; either will create a new variable containing the nonselection hazard—what Heckman (1979) referred to as the inverse of the Mills ratio—from the selection equation. The nonselection hazard is computed from the estimated parameters of the selection equation.
The following option is available with `heckman` but is not shown in the dialog box: `coeflegend`; see [R] Estimation options.

Remarks and examples

The Heckman selection model (Gronau 1974; Lewis 1974; Heckman 1976) assumes that there exists an underlying regression relationship,

\[ y_j = x_j \beta + u_{1j} \]

regression equation

The dependent variable, however, is not always observed. Rather, the dependent variable for observation \( j \) is observed if

\[ z_j \gamma + u_{2j} > 0 \]

selection equation

where

\[ u_1 \sim N(0, \sigma) \]
\[ u_2 \sim N(0, 1) \]
\[ \text{corr}(u_1, u_2) = \rho \]

When \( \rho \neq 0 \), standard regression techniques applied to the first equation yield biased results. `heckman` provides consistent, asymptotically efficient estimates for all the parameters in such models.

In one classic example, the first equation describes the wages of women. Women choose whether to work, and thus, from our point of view as researchers, whether we observe their wages in our data. If women made this decision randomly, we could ignore that not all wages are observed and use ordinary regression to fit a wage model. Such an assumption of random participation, however, is unlikely to be true; women who would have low wages may be unlikely to choose to work, and thus the sample of observed wages is biased upward. In the jargon of economics, women choose not to work when their personal reservation wage is greater than the wage offered by employers. Thus, women who choose not to work might have even higher offer wages than those who do work—they may have high offer wages, but they have even higher reservation wages. We could tell a story that competency is related to wages, but competency is rewarded more at home than in the labor force.

In any case, in this problem—which is the paradigm for most such problems—a solution can be found if there are some variables that strongly affect the chances for observation (the reservation wage) but not the outcome under study (the offer wage). Such a variable might be the number of children in the home. (Theoretically, we do not need such identifying variables, but without them, we depend on functional form to identify the model. It would be difficult for anyone to take such results seriously because the functional form assumptions have no firm basis in theory.)

Example 1

In the syntax for `heckman`, `depvar` and `indepvars` are the dependent variable and regressors for the underlying regression model to be fit \( (y = X\beta) \), and `varlist`s are the variables \( (Z) \) thought to determine whether `depvar` is observed or unobserved (selected or not selected). In our female wage example, the number of children at home would be included in the second list. By default, `heckman` assumes that missing values (see [U] 12.2.1 Missing values) of `depvar` imply that the dependent variable is
unobserved (not selected). With some datasets, it is more convenient to specify a binary variable (depvar$_s$) that identifies the observations for which the dependent is observed/selected (depvar$_s$ $\neq$ 0) or not observed (depvar$_s$ = 0); heckman will accommodate either type of data. Here we have a (fictional) dataset on 2,000 women, 1,343 of whom work:

```
use https://www.stata-press.com/data/r17/womenwk
summarize age educ married children wage
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>2,000</td>
<td>36.208</td>
<td>8.28656</td>
<td>20</td>
<td>59</td>
</tr>
<tr>
<td>education</td>
<td>2,000</td>
<td>13.084</td>
<td>3.045912</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>married</td>
<td>2,000</td>
<td>.6705</td>
<td>.4701492</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>children</td>
<td>2,000</td>
<td>1.6445</td>
<td>1.398963</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>wage</td>
<td>1,343</td>
<td>23.69217</td>
<td>6.305374</td>
<td>5.88497</td>
<td>45.80979</td>
</tr>
</tbody>
</table>

We will assume that the hourly wage is a function of education and age, whereas the likelihood of working (the likelihood of the wage being observed) is a function of marital status, the number of children at home, and (implicitly) the wage (via the inclusion of age and education, which we think determine the wage):

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

Iteration 0:  log likelihood = -5178.7009
Iteration 1:  log likelihood = -5178.3049
Iteration 2:  log likelihood = -5178.3045

**Heckman selection model**

Number of obs = 2,000
Selected = 1,343
Nonselected = 657
Wald chi2(2) = 508.44
Log likelihood = -5178.304
Prob > chi2 = 0.0000

```
| wage | Coefficient | Std. err. | z     | P>|z| | [95% conf. interval] |
|------|-------------|-----------|-------|-----|----------------------|
| wage |             |           |       |   |                      |
| educ | .9899537    | .0532565  | 18.59 | 0.000 | .8855729 1.094334   |
| age  | .2131294    | .0206031  | 10.34 | 0.000 | .1727481 0.2535108 |
| _cons| .4857752    | 1.077037  | 0.45  | 0.652 | -1.625179 2.59673  |
|      |             |           |       |   |                      |
| select |         |           |       |   |                      |
| married| .4451721    | .0673954  | 6.61  | 0.000 | .3130794 0.5772647 |
| children| .4387068    | .0277828  | 15.79 | 0.000 | .3842534 0.4931601 |
| education| .0557318    | .0107349  | 5.19  | 0.000 | .0346917 0.0767718 |
| age  | .0365098    | .0041533  | 8.79  | 0.000 | .0283694 0.0446502 |
| _cons| -2.491015   | .1893402  | -13.16 | 0.000 | -2.862115 -2.119915 |
|      |             |           |       |   |                      |
| /athrho| .8742086    | .1014225  | 8.62  | 0.000 | .6754241 1.072993  |
| /lnsigma| 1.792559    | .027598   | 64.95 | 0.000 | 1.738468 1.84665   |
|      |             |           |       |   |                      |
| rho  | .7035061    | .0512264  | 13.16 | 0.000 | .5885365 .7905862  |
| sigma| 6.004797    | .1657202  | 36.20 | 0.000 | 5.68862 6.338548   |
| lambda| 4.224412    | .3992265  | 10.34 | 0.000 | 3.441942 5.006881  |
```

LR test of indep. eqns. (rho = 0): chi2(1) = 61.20 Prob > chi2 = 0.0000

heckman assumes that wage is the dependent variable and that the first variable list (educ and age) are the determinants of wage. The variables specified in the select() option (married, children, educ, and age) are assumed to determine whether the dependent variable is observed (the selection equation). Thus, we fit the model

\[
\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{age} + u_1
\]
and we assumed that wage is observed if

$$\gamma_0 + \gamma_1 \text{married} + \gamma_2 \text{children} + \gamma_3 \text{educ} + \gamma_4 \text{age} + u_2 > 0$$

where $u_1$ and $u_2$ have correlation $\rho$.

The reported results for the wage equation are interpreted exactly as though we observed wage data for all women in the sample; the coefficients on age and education level represent the estimated marginal effects of the regressors in the underlying regression equation. The results for the two ancillary parameters require some explanation. heckman does not directly estimate $\rho$; to constrain $\rho$ within its valid limits, and for numerical stability during optimization, it estimates the inverse hyperbolic tangent of $\rho$:

$$\text{atanh} \rho = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right)$$

This estimate is reported as /athrho. In the bottom panel of the output, heckman undoes this transformation for you: the estimated value of $\rho$ is 0.7035061. The standard error for $\rho$ is computed using the delta method, and its confidence intervals are the transformed intervals of /athrho.

Similarly, $\sigma$, the standard error of the residual in the wage equation, is not directly estimated; for numerical stability, heckman instead estimates $\ln \sigma$. The untransformed sigma is reported at the end of the output: 6.004797.

Finally, some researchers—especially economists—are used to the selectivity effect summarized not by $\rho$ but by $\lambda = \rho \sigma$. heckman reports this, too, along with an estimate of the standard error and confidence interval.

Technical note

If each of the equations in the model had contained many regressors, the heckman command could have become long. An alternate way of specifying our wage model would be to use Stata’s global macros. The following lines are an equivalent way of specifying our model:

```
   . global wageeq "wage educ age"
   . global seleq "married children educ age"
   . heckman $wageeq, select($seleq)
```

(output omitted)

Technical note

The reported model $\chi^2$ test is a Wald test that all coefficients in the regression model (except the constant) are 0. heckman is an estimation command, so you can use test, testnl, or lrtest to perform tests against whatever nested alternate model you choose; see [R] test, [R] testnl, and [R] lrtest.

The estimation of $\rho$ and $\sigma$ in the forms atanh $\rho$ and $\ln \sigma$ extends the range of these parameters to infinity in both directions, thus avoiding boundary problems during the maximization. Tests of $\rho$ must be made in the transformed units. However, because atanh(0) = 0, the reported test for atanh $\rho = 0$ is equivalent to the test for $\rho = 0$. 


The likelihood-ratio test reported at the bottom of the output is an equivalent test for \( \rho = 0 \) and is computationally the comparison of the joint likelihood of an independent probit model for the selection equation and a regression model on the observed wage data against the Heckman model likelihood. Because \( \chi^2 = 61.20 \), this clearly justifies the Heckman selection equation with these data.

Example 2

`heckman` supports the Huber/White/sandwich estimator of variance under the `vce(robust)` and `vce(cluster clustvar)` options or when `pweights` are used for population-weighted data; see [U] 20.22 Obtaining robust variance estimates. We can obtain robust standard errors for our wage model by specifying clustering on county of residence (the county variable).

```
. heckman wage educ age, select(married children educ age) vce(cluster county)
Iteration 0:  log pseudolikelihood = -5178.7009
Iteration 1:  log pseudolikelihood = -5178.3049
Iteration 2:  log pseudolikelihood = -5178.3045
Heckman selection model
(regression model with sample selection)  Number of obs = 2,000
               Selected = 1,343
               Nonselected = 657
          Wald chi2(1) = .
(Std. err. adjusted for 10 clusters in county)
Robust

|        | Coefficient | std. err. | z    | P>|z| | [95% conf. interval] |
|--------|-------------|-----------|------|------|----------------------|
| wage   |             |           |      |      |                      |
| education | .9899537   | .0600061  | 16.50| 0.000| .8723438             | 1.107564 |
| age    | .2131294   | .020995   | 10.15| 0.000| .17198               | .2542789|
| _cons | .4857752   | 1.302103  | 0.37 | 0.709| -2.066299            | 3.03785  |
| select |             |           |      |      |                      |
| married | .4451721   | .0731472  | 6.09 | 0.000| .3018062             | .5885379 |
| children | .4387068  | .0312386  | 14.04| 0.000| .3774802             | .4999333 |
| education | .0557318  | .0110039  | 5.06 | 0.000| .0341645             | .0772991 |
| age    | .0365098   | .004038   | 9.04 | 0.000| .0285954             | .0444242 |
| _cons | -2.491015  | .1153305  | -21.60| 0.000| -2.717059            | -2.264972|
| /athrho | .8742086   | .1403337  | 6.23 | 0.000| .5991596             | 1.149258 |
| /lnsigma | 1.792559  | .0258458  | 69.36| 0.000| 1.741902             | 1.843216 |
| rho    | .7035061   | .0708796  | 5.36 | 0.000| .5364513             | .817508  |
| sigma  | 6.004797   | .155199   | 5.71 | 0.000| 5.318899             | 6.316818 |
| lambda | 4.224412   | .5186709  | 8.13 | 0.000| 3.707835             | 5.240988 |
```

Wald test of indep. eqns. (rho = 0): chi2(1) = 38.81  Prob > chi2 = 0.0000

The robust standard errors tend to be a bit larger, but we notice no systematic differences. This finding is not surprising because the data were not constructed to have any county-specific correlations or any other characteristics that would deviate from the assumptions of the Heckman model.
Example 3

Stata also produces Heckman’s (1979) two-step efficient estimator of the model with the `twostep` option. Maximum likelihood estimation of the parameters can be time consuming with large datasets, and the two-step estimates may provide a good alternative in such cases. Continuing with the women’s wage model, we can obtain the two-step estimates with Heckman’s consistent covariance estimates by typing

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

Heckman selection model -- two-step estimates  Number of obs = 2,000
(regression model with sample selection)  Selected = 1,343
Nonselected = 657  Wald chi2(2) = 442.54
Prob > chi2 = 0.0000

|        | Coefficient | Std. err. | z     | P>|z|  | [95% conf. interval] |
|--------|-------------|-----------|-------|------|---------------------|
| wage   |             |           |       |      |                     |
| education | 0.9825259  | 0.0538821 | 18.23 | 0.00 | 0.8769189 1.088133 |
| age    | 0.2118695  | 0.0220511 | 9.61  | 0.00 | 0.1686502 0.2550888 |
| _cons  | 0.7340391  | 1.248331  | 0.59  | 0.56 | -1.712645 3.180723 |
| select |             |           |       |      |                     |
| married | 0.4308575  | 0.074208  | 5.81  | 0.00 | 0.2854125 0.5763025 |
| children | 0.4473249  | 0.0287417 | 15.56 | 0.00 | 0.3909922 0.5036576 |
| education | 0.0583645  | 0.0109742 | 5.32  | 0.00 | 0.0368555 0.0798735 |
| age    | 0.0347211  | 0.0042293 | 8.21  | 0.00 | 0.0264318 0.0430105 |
| _cons  | -2.467365  | 1.925635  | -12.81| 0.00 | -2.844782 -2.089948 |
| /mills |             |           |       |      |                     |
| lambda | 4.001615    | 0.6065388 | 6.60  | 0.00 | 2.812821 5.19041   |
| rho    | 0.67284     |           |       |      |                     |
| sigma  | 5.9473529   |           |       |      |                     |

Technical note

The Heckman selection model depends strongly on the model being correct, much more so than ordinary regression. Running a separate probit or logit for sample inclusion followed by a regression, referred to in the literature as the two-part model (Manning, Duan, and Rogers 1987)—not to be confused with Heckman’s two-step procedure—is an especially attractive alternative if the regression part of the model arose because of taking a logarithm of zero values. When the goal is to analyze an underlying regression model or to predict the value of the dependent variable that would be observed in the absence of selection, however, the Heckman model is more appropriate. When the goal is to predict an actual response, the two-part model is usually the better choice.

The Heckman selection model can be unstable when the model is not properly specified or if a specific dataset simply does not support the model’s assumptions. For example, let’s examine the solution to another simulated problem.
. use https://www.stata-press.com/data/r17/twopart
. heckman yt x1 x2 x3, select(z1 z2) nonrtol

Iteration 0: log likelihood = -111.94996
Iteration 1: log likelihood = -110.82258
Iteration 2: log likelihood = -110.17707
Iteration 3: log likelihood = -107.70663 (not concave)
Iteration 4: log likelihood = -107.07729 (not concave)
(output omitted)
Iteration 36: log likelihood = -104.0825

Heckman selection model
(regression model with sample selection)
Number of obs = 150
Selected = 63
Nonselected = 87
Wald chi2(3) = 8.84e+08
Log likelihood = -104.0825
Prob > chi2 = 0.0000

|         | Coefficient | Std. err. | z      | P>|z| | 95% conf. interval |
|---------|-------------|-----------|--------|-----|-------------------|
| yt      |             |           |        |     |                   |
| x1      | 0.8974192   | 0.0002164 | 4146.52| 0.000| 0.896995 - 0.8978434 |
| x2      | -2.525303   | 0.0001244 | -2.0e+04| 0.000| -2.525546 - 2.525059 |
| x3      | 2.855786    | 0.0002695 | 1.1e+04| 0.000| 2.855258 - 2.856314 |
| _cons   | 0.6975003   | 0.0907873 | 7.68   | 0.000| 0.5195604 - 0.8754402 |
| select  |             |           |        |     |                   |
| z1      | -0.6826482  | 0.089871  | -7.67  | 0.000| -0.8570598 - 0.5082367 |
| z2      | 1.003678    | 0.1308344 | 7.67   | 0.000| 0.7472471 - 1.260108 |
| _cons   | -0.3605665  | 0.1219011 | -2.96  | 0.003| -0.5994883 - 0.1216447 |
| /athrho | 16.11489    | 260.7581  | 0.06   | 0.951| -494.9617 - 527.1914 |
| /lnsigma| -.5396877   | 0.1303548 | -4.14  | 0.000| -0.7951785 - 0.284197 |
| rho     | 1           | 1.05e-11  |        |     | -1                |
| sigma   | 0.5829302   | 0.0759878 | 4.51   | 0.000| 0.4515006 - 0.7526184 |
| lambda  | 0.5829302   | 0.0759878 | 4.33   | 0.000| 0.433997 - 0.7318635 |

LR test of indep. eqns. (rho = 0): chi2(1) = 25.67 Prob > chi2 = 0.0000

The model has converged to a value of \( \rho \) that is 1.0—within machine-rounding tolerances. Given the form of the likelihood for the Heckman selection model, this implies a division by zero, and it is surprising that the model solution turns out as well as it does. Reparameterizing \( \rho \) has allowed the estimation to converge, but we clearly have problems with the estimates. Moreover, if this had occurred in a large dataset, waiting for convergence might take considerable time.

This dataset was not intentionally developed to cause problems. It is actually generated by a “Heckman process” and when generated starting from different random values can be easily estimated. The luck of the draw here merely led to data that, despite the source, did not support the assumptions of the Heckman model.

The two-step model is generally more stable when the data are problematic. It even tolerates estimates of \( \rho \) less than −1 and greater than 1. For these reasons, the two-step model may be preferred when exploring a large dataset. Still, if the maximum likelihood estimates cannot converge, or converge to a value of \( \rho \) that is at the boundary of acceptable values, there is scant support for fitting a Heckman selection model on the data. Heckman (1979) discusses the implications of \( \rho \) being exactly 1 or 0, together with the implications of other possible covariance relationships among the model’s determinants.
Stored results

`heckman` (maximum likelihood) stores the following in `e()`:

**Scalars**
- `e(N)` number of observations
- `e(N_selected)` number of selected observations
- `e(N_nonselected)` number of nonselected observations
- `e(k)` number of parameters
- `e(k_eq)` number of equations in `e(b)`
- `e(k_eq_model)` number of equations in overall model test
- `e(k_aux)` number of auxiliary parameters
- `e(k_dv)` number of dependent variables
- `e(df_m)` model degrees of freedom
- `e(ll)` log likelihood
- `e(ll_0)` log likelihood, constant-only model
- `e(N_clust)` number of clusters
- `e(lambda)` \( \lambda \)
- `e(selambda)` standard error of \( \lambda \)
- `e(sigma)` sigma
- `e(chi2)` \( \chi^2 \)
- `e(chi2_c)` \( \chi^2 \) for comparison test
- `e(p)` \( p \)-value for model test
- `e(p_c)` \( p \)-value for comparison test
- `e(rho)` \( \rho \)
- `e(rank)` rank of `e(V)`
- `e(rank0)` rank of `e(V)` for constant-only model
- `e(ic)` number of iterations
- `e(rc)` return code
- `e(converged)` 1 if converged, 0 otherwise

**Macros**
- `e(cmd)` `heckman`
- `e(cmdline)` command as typed
- `e(depvar)` names of dependent variables
- `e(wtype)` weight type
- `e(wexp)` weight expression
- `e(title)` title in estimation output
- `e(title2)` secondary title in estimation output
- `e(clustvar)` name of cluster variable
- `e(offset1)` offset for regression equation
- `e(offset2)` offset for selection equation
- `e(mills)` variable containing nonselection hazard (inverse of Mills’s ratio)
- `e(chi2type)` Wald or LR; type of model \( \chi^2 \) test
- `e(chi2_ct)` Wald or LR; type of model \( \chi^2 \) test corresponding to `e(chi2_c)`
- `e(vce)` `vcetype` specified in `vce()`
- `e(vcetype)` title used to label Std. err.
- `e(opt)` type of optimization
- `e(which)` `max` or `min`; whether optimizer is to perform maximization or minimization
- `e(method)` `ml`
- `e(ml_method)` type of `ml` method
- `e(user)` name of likelihood-evaluator program
- `e(technique)` maximization technique
- `e(properties)` `b V`
- `e(predict)` program used to implement `predict`
- `e(marginsok)` predictions allowed by `margins`
- `e(marginsnotok)` predictions disallowed by `margins`
- `e(asbalanced)` factor variables `fvset` as `asbalanced`
- `e(asobserved)` factor variables `fvset` as `asobserved`
Matrices
- `e(b)` coefficient vector
- `e(Cns)` constraints matrix
- `e(ilog)` iteration log (up to 20 iterations)
- `e(gradient)` gradient vector
- `e(V)` variance–covariance matrix of the estimators
- `e(V_modelbased)` model-based variance

Functions
- `e(sample)` marks estimation sample

In addition to the above, the following is stored in `r()`:

Matrices
- `r(table)` matrix containing the coefficients with their standard errors, test statistics, \( p \)-values, and confidence intervals

Note that results stored in `r()` are updated when the command is replayed and will be replaced when any \( r \)-class command is run after the estimation command.

`heckman` (two-step) stores the following in \( e() \):

Scalars
- `e(N)` number of observations
- `e(N_selected)` number of selected observations
- `e(N_nonselected)` number of nonselected observations
- `e(df_m)` model degrees of freedom
- `e(lambda)` \( \lambda \)
- `e(selambda)` standard error of \( \lambda \)
- `e(sigma)` sigma
- `e(chi2)` \( \chi^2 \)
- `e(p)` \( p \)-value for comparison test
- `e(rho)` \( \rho \)
- `e(rank)` rank of \( e(V) \)

Macros
- `e(cmd)` `heckman`
- `e(cmdline)` command as typed
- `e(depvar)` names of dependent variables
- `e(title)` title in estimation output
- `e(title2)` secondary title in estimation output
- `e(mills)` variable containing nonselection hazard (inverse of Mills's ratio)
- `e(chi2type)` Wald or LR; type of model \( \chi^2 \) test
- `e(vce)` `vctype` specified in `vce()`
- `e(rhometh)` rhosigma, rhotrunc, rholimited, or rhoforce
- `e(method)` `twostep`
- `e(properties)` `b V`
- `e(predict)` program used to implement `predict`
- `e(marginsok)` predictions allowed by `margins`
- `e(marginsnotok)` predictions disallowed by `margins`
- `e(asbalanced)` factor variables `fvset` as `asbalanced`
- `e(asobserved)` factor variables `fvset` as `asobserved`

Matrices
- `e(b)` coefficient vector
- `e(V)` variance–covariance matrix of the estimators

Functions
- `e(sample)` marks estimation sample

In addition to the above, the following is stored in `r()`:

Matrices
- `r(table)` matrix containing the coefficients with their standard errors, test statistics, \( p \)-values, and confidence intervals
Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

James Joseph Heckman (1944– ) was born in Chicago in 1944 and studied mathematics at Colorado College and economics at Princeton. He has taught economics at Columbia and (since 1973) at the University of Chicago. He has worked on developing a scientific basis for economic policy evaluation, with emphasis on models of individuals or disaggregated groups and the problems and possibilities created by heterogeneity, diversity, and unobserved counterfactual states. In 2000, he shared the Nobel Prize in Economics with Daniel L. McFadden.

Methods and formulas


Regression estimates using the nonselection hazard (Heckman 1979) provide starting values for maximum likelihood estimation.

The regression equation is

$$y_j = x_j \beta + u_1$$

The selection equation is

$$z_j \gamma + u_2 > 0$$

where

$$u_1 \sim N(0, \sigma)$$

$$u_2 \sim N(0, 1)$$

$$\text{corr}(u_1, u_2) = \rho$$

The log likelihood for observation $j$, $\ln L_j = l_j$, is

$$l_j = \begin{cases} 
  w_j \ln \Phi \left( \frac{z_j \gamma + (y_j - x_j \beta) \rho/\sigma}{\sqrt{1 - \rho^2}} \right) - \frac{w_j}{2} \left( \frac{y_j - x_j \beta}{\sigma} \right)^2 - w_j \ln(\sqrt{2\pi} \sigma) & y_j \text{ observed} \\
  w_j \ln \Phi(-z_j \gamma) & y_j \text{ not observed}
\end{cases}$$

where $\Phi(\cdot)$ is the standard cumulative normal and $w_j$ is an optional weight for observation $j$.

In the maximum likelihood estimation, $\sigma$ and $\rho$ are not directly estimated. Directly estimated are $\ln \sigma$ and $\text{atanh} \, \rho$:

$$\text{atanh} \, \rho = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right)$$

The standard error of $\lambda = \rho \sigma$ is approximated through the propagation of error (delta) method; that is,

$$\text{Var}(\lambda) \approx D \text{Var}\{\text{atanh} \, \rho, \ln \sigma\} D'$$
where $D$ is the Jacobian of $\lambda$ with respect to $\text{atanh} \, \rho$ and $\ln \sigma$.

With maximum likelihood estimation, this command supports the Huber/White/sandwich estimator of the variance and its clustered version using `vce(robust)` and `vce(cluster clustvar)`, respectively. See `[P] _robust`, particularly *Maximum likelihood estimators* and *Methods and formulas*.

The maximum likelihood version of `heckman` also supports estimation with survey data. For details on VCEs with survey data, see `[SVY] Variance estimation`.

The two-step estimates are computed using Heckman’s (1979) procedure.

Probit estimates of the selection equation

$$\Pr(y_j \text{ observed} \mid z_j) = \Phi(z_j \gamma)$$

are obtained. From these estimates, the nonselection hazard—that what Heckman (1979) referred to as the inverse of the Mills ratio, $m_j$—for each observation $j$ is computed as

$$m_j = \frac{\phi(z_j \hat{\gamma})}{\Phi(z_j \hat{\gamma})}$$

where $\phi$ is the normal density. We also define

$$\delta_j = m_j(m_j + \hat{\gamma} z_j)$$

Following Heckman, the two-step parameter estimates of $\beta$ are obtained by augmenting the regression equation with the nonselection hazard $m$. Thus, the regressors become $[X \ m]$, and we obtain the additional parameter estimate $\beta_m$ on the variable containing the nonselection hazard.

A consistent estimate of the regression disturbance variance is obtained using the residuals from the augmented regression and the parameter estimate on the nonselection hazard,

$$\hat{\sigma}^2 = \frac{e'e + \beta_m^2 \sum_{j=1}^{N} \delta_j}{N}$$

The two-step estimate of $\rho$ is then

$$\hat{\rho} = \frac{\beta_m}{\hat{\sigma}}$$

Heckman derived consistent estimates of the coefficient covariance matrix on the basis of the augmented regression.

Let $W = [X \ m]$ and $R$ be a square, diagonal matrix of dimension $N$, with $(1 - \hat{\rho}^2 \delta_j)$ as the diagonal elements. The conventional VCE is

$$V_{\text{twostep}} = \hat{\sigma}^2 (W'W)^{-1}(W'RW + Q)(W'W)^{-1}$$

where

$$Q = \hat{\rho}^2 (W'DZ)V_p(Z'DW)$$

where $D$ is the square, diagonal matrix of dimension $N$ with $\delta_j$ as the diagonal elements; $Z$ is the data matrix of selection equation covariates; and $V_p$ is the variance–covariance estimate from the probit estimation of the selection equation.
References


Baum, C. F. 2006. *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.


[https://doi.org/10.1086/260267](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.


Also see

[R] heckman postestimation — Postestimation tools for heckman

[R] heckoprobit — Ordered probit model with sample selection

[R] heckpoisson — Poisson regression with sample selection

[R] heckprobit — Probit model with sample selection

[R] regress — Linear regression

[R] tobit — Tobit regression

[BAYES] bayes: heckman — Bayesian Heckman selection model

[ERM] eregress — Extended linear regression

[SVY] svy estimation — Estimation commands for survey data

[TE] etregress — Linear regression with endogenous treatment effects

[XT] xheckman — Random-effects regression with sample selection

[U] 20 Estimation and postestimation commands