

**heckman** — Heckman selection model[Description](#)[Menu](#)[Options for Heckman selection model \(ML\)](#)[Remarks and examples](#)[Methods and formulas](#)[Also see](#)[Quick start](#)[Syntax](#)[Options for Heckman selection model \(two-step\)](#)[Stored results](#)[References](#)

## 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  $x1$  with  $v1$  predicting selection when binary variable `selected` indicates selection status

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

As above, and generate  $v2$  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

### **heckman for maximum likelihood estimates**

Statistics > Sample-selection models > Heckman selection model (ML)

### **heckman for two-step consistent estimates**

Statistics > Sample-selection models > Heckman selection model (two-step)

## Syntax

### *Basic syntax*

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

or

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

### *Full syntax for maximum likelihood estimates only*

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

### *Full syntax for Heckman's two-step consistent estimates only*

```
heckman depvar [indepvars] [if] [in], twostep  
  select( [depvars =] varlists [, noconstant] ) [heckman_ts_options]
```

---

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

---

\*`select()` is required.

The full specification is `select( [depvars = ] varlists [ , noconstant offset(varnameo) ] )`.

<i>heckman_ts_options</i>	Description
Model	
* <u>select</u> ()	specify selection equation: dependent and independent variables; whether to have constant term
* <u>twostep</u>	produce two-step consistent estimate
<u>noconstant</u>	suppress constant term
<u>rhosigma</u>	truncate $\rho$ to $[-1, 1]$ with consistent $\sigma$
<u>hotrunc</u>	truncate $\rho$ to $[-1, 1]$
<u>holimited</u>	truncate $\rho$ in limited cases
<u>hoforce</u>	do not truncate $\rho$
SE	
<u>vce</u> ( <i>vcetype</i> )	<i>vcetype</i> may be conventional, <u>bootstrap</u> , or <u>jackknife</u>
Reporting	
<u>level</u> (#)	set confidence level; default is level(95)
<u>first</u>	report first-step probit estimates
<u>nshazard</u> ( <i>newvar</i> )	generate nonselection hazard variable
<u>mills</u> ( <i>newvar</i> )	synonym for <u>nshazard</u> ()
<u>display_options</u>	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
<u>coeflegend</u>	display legend instead of statistics

\* select() and twostep are required.

The full specification is select( [*depvar*<sub>s</sub> = ] *varlist*<sub>s</sub> [ , noconstant ] ).

*indepvars* and *varlist*<sub>s</sub> may contain factor variables; see [U] 11.4.3 **Factor variables**.

*depvar*, *indepvars*, *varlist*<sub>s</sub>, and *depvar*<sub>s</sub> may contain time-series operators; see [U] 11.4.4 **Time-series varlists**.

bootstrap, by, fp, jackknife, rolling, statsby, and svy are allowed; see [U] 11.1.10 **Prefix commands**.

Weights are not allowed with the bootstrap prefix; see [R] **bootstrap**.

aweights are not allowed with the jackknife prefix; see [R] **jackknife**.

twostep, vce(), first, noskip, and weights are not allowed with the svy prefix; see [SVY] **svy**.

pweights, aweights, fweights, and iweights are allowed with maximum likelihood estimation; see [U] 11.1.6 **weight**.

No weights are allowed if twostep is specified.

coeflegend does not appear in the dialog box.

See [U] 20 **Estimation and postestimation commands** for more capabilities of estimation commands.

## Options for Heckman selection model (ML)

Model

select( [*depvar*<sub>s</sub> = ] *varlist*<sub>s</sub> [ , noconstant offset(*varname*<sub>o</sub>) ] ) 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*<sub>s</sub> 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*<sub>s</sub> 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).

`offset(varnameo)` specifies that selection offset *varname<sub>o</sub>* be included in the model with the coefficient constrained to be 1.

`noconstant`, `offset(varname)`, `constraints(constraints)`, `collinear`; see [R] [estimation options](#).

#### SE/Robust

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

#### Reporting

`level(#)`; see [R] [estimation options](#).

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

`noskip` specifies that a full maximum-likelihood model with only a constant for the regression equation be fit. This model is not displayed but is used as the base model to compute a likelihood-ratio test for the model test statistic displayed in the estimation header. By default, the overall model test statistic is an asymptotically equivalent Wald test that all the parameters in the regression equation are zero (except the constant). For many models, this option can substantially increase estimation time.

`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.

`nocnsreport`; see [R] [estimation options](#).

`display_options`: `nocl`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(style)`, `cformat(%fmt)`, `pformat(%fmt)`, `sformat(%fmt)`, and `nolstretch`; see [R] [estimation options](#).

#### Maximization

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

Setting the optimization type to `technique(bhhh)` resets the default *vcetype* to `vce(opg)`.

The following option is available with `heckman` but is not shown in the dialog box:

`coeflegend`; see [R] [estimation options](#).

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

## Model

`select( [ depvars = ] varlists [ , 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*<sub>s</sub> 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*<sub>s</sub> 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).

`twostep` specifies that Heckman's (1979) two-step efficient estimates of the parameters, standard errors, and covariance matrix be produced.

`noconstant`; see [R] [estimation options](#).

`rhosigma`, `rot trunc`, `rholimited`, and `rho force` 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 `rot trunc` 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.

`rot trunc` 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**<sub>twostep</sub> and **Q**; see [Methods and formulas](#). In all other computations, the untruncated estimate of  $\rho$  is used.

`rho force` 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  $[-1, 1]$ .

## SE

`vce(vctype)` 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.

## Reporting

`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.

*display\_options*: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(style)`, `cformat(%fmt)`, `pformat(%fmt)`, `sformat(%fmt)`, and `nolstretch`; see [R] [estimation options](#).

The following option is available with `heckman` but is not shown in the dialog box:

`coeflegend`; see [R] [estimation options](#).

## Remarks and examples

[stata.com](http://www.stata.com)

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

$$y_j = \mathbf{x}_j\boldsymbol{\beta} + u_{1j} \qquad \text{regression equation}$$

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

$$\mathbf{z}_j\boldsymbol{\gamma} + u_{2j} > 0 \qquad \text{selection equation}$$

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 ( $y = \mathbf{X}\beta$ ), and `varlists` are the variables ( $\mathbf{Z}$ ) 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 (`depvars`) that identifies the observations for which the dependent is observed/selected (`depvars ≠ 0`) or not observed (`depvars = 0`); `heckman` will accommodate either type of data. Here we have a (fictional) dataset on 2,000 women, 1,343 of whom work:

```
. use http://www.stata-press.com/data/r14/womenwk
. summarize age educ married children wage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	2,000	36.208	8.28656	20	59
education	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

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
(regression model with sample selection)  Censored obs   =       657
                                           Uncensored obs =     1,343
                                           Wald chi2(2)   =     508.44
Log likelihood = -5178.304              Prob > chi2    =     0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
<b>wage</b>						
wage						
education	.9899537	.0532565	18.59	0.000	.8855729	1.094334
age	.2131294	.0206031	10.34	0.000	.1727481	.2535108
_cons	.4857752	1.077037	0.45	0.652	-1.625179	2.59673
<b>select</b>						
married	.4451721	.0673954	6.61	0.000	.3130794	.5772647
children	.4387068	.0277828	15.79	0.000	.3842534	.4931601
education	.0557318	.0107349	5.19	0.000	.0346917	.0767718
age	.0365098	.0041533	8.79	0.000	.0283694	.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			.5885365	.7905862
sigma	6.004797	.1657202			5.68862	6.338548
lambda	4.224412	.3992265			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  $\operatorname{atanh} \rho$  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  $\operatorname{atanh}(0) = 0$ , the reported test for  $\operatorname{atanh} \rho = 0$  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] 20.21 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      Number of obs      =      2,000
(regression model with sample selection)  Censored obs       =        657
                                          Uncensored obs     =     1,343
                                          Wald chi2(1)       =          .
                                          Prob > chi2        =          .

Log pseudolikelihood = -5178.304          Prob > chi2        =          .
                                          (Std. Err. adjusted for 10 clusters in county)
```

	Robust				[95% Conf. Interval]	
wage	Coef.	Std. Err.	z	P> z		
<b>wage</b>						
education	.9899537	.0600061	16.50	0.000	.8723438	1.107564
age	.2131294	.0209995	10.15	0.000	.17198	.2542789
_cons	.4857752	1.302103	0.37	0.709	-2.066299	3.03785
<b>select</b>						
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			.5364513	.817508
sigma	6.004797	.155199			5.708189	6.316818
lambda	4.224412	.5186709			3.207835	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)          Censored obs    =       657
                                                  Uncensored obs  =     1,343
                                                  Wald chi2(2)    =     442.54
                                                  Prob > chi2     =      0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
<b>wage</b>						
education	.9825259	.0538821	18.23	0.000	.8769189	1.088133
age	.2118695	.0220511	9.61	0.000	.1686502	.2550888
_cons	.7340391	1.248331	0.59	0.557	-1.712645	3.180723
<b>select</b>						
married	.4308575	.074208	5.81	0.000	.2854125	.5763025
children	.4473249	.0287417	15.56	0.000	.3909922	.5036576
education	.0583645	.0109742	5.32	0.000	.0368555	.0798735
age	.0347211	.0042293	8.21	0.000	.0264318	.0430105
_cons	-2.467365	.1925635	-12.81	0.000	-2.844782	-2.089948
<b>mills</b>						
lambda	4.001615	.6065388	6.60	0.000	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 http://www.stata-press.com/data/r14/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 33: log likelihood = -104.0825 (not concave)
Iteration 34: log likelihood = -104.0825

Heckman selection model          Number of obs   =       150
(regression model with sample selection)  Censored obs   =        87
                                           Uncensored obs =        63
                                           Wald chi2(3)   =    8.64e+08
Log likelihood = -104.0825        Prob > chi2     =     0.0000

```

	yt	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
yt							
	x1	.8974192	.0002247	3994.69	0.000	.8969789	.8978595
	x2	-2.525303	.0001472	-1.7e+04	0.000	-2.525591	-2.525014
	x3	2.855786	.0004181	6829.86	0.000	2.854966	2.856605
	_cons	.6975442	.0920515	7.58	0.000	.5171265	.8779619
select							
	z1	-.6825988	.0900159	-7.58	0.000	-.8590267	-.5061709
	z2	1.003605	.132347	7.58	0.000	.7442097	1.263
	_cons	-.3604652	.1232778	-2.92	0.003	-.6020852	-.1188452
	/athrho	16.19193	280.9822	0.06	0.954	-534.523	566.9069
	/lnsigma	-.5396153	.1318714	-4.09	0.000	-.7980786	-.2811521
	rho	1	9.73e-12			-1	1
	sigma	.5829725	.0768774			.4501931	.7549135
	lambda	.5829725	.0768774			.4322955	.7336494

```
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

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

### Macros

<code>e(cmd)</code>	<code>heckman</code>
<code>e(cmdline)</code>	command as typed
<code>e(depvar)</code>	names of dependent variables
<code>e(wtype)</code>	weight type
<code>e(wexp)</code>	weight expression
<code>e(title)</code>	title in estimation output
<code>e(title2)</code>	secondary title in estimation output
<code>e(clustvar)</code>	name of cluster variable
<code>e(offset1)</code>	offset for regression equation
<code>e(offset2)</code>	offset for selection equation
<code>e(mills)</code>	variable containing nonselection hazard (inverse of Mills's ratio)
<code>e(chi2type)</code>	Wald or LR; type of model $\chi^2$ test
<code>e(chi2_ct)</code>	Wald or LR; type of model $\chi^2$ test corresponding to <code>e(chi2_c)</code>
<code>e(vce)</code>	<code>vcetype</code> specified in <code>vce()</code>
<code>e(vcetype)</code>	title used to label Std. Err.
<code>e(opt)</code>	type of optimization
<code>e(which)</code>	max or min; whether optimizer is to perform maximization or minimization
<code>e(method)</code>	m1
<code>e(ml_method)</code>	type of ml method
<code>e(user)</code>	name of likelihood-evaluator program
<code>e(technique)</code>	maximization technique
<code>e(properties)</code>	b V
<code>e(predict)</code>	program used to implement <code>predict</code>
<code>e(marginsok)</code>	predictions allowed by <code>margins</code>
<code>e(asbalanced)</code>	factor variables <code>fvset</code> as <code>asbalanced</code>
<code>e(asobserved)</code>	factor variables <code>fvset</code> as <code>asobserved</code>

### Matrices

<code>e(b)</code>	coefficient vector
<code>e(Cns)</code>	constraints matrix
<code>e(iilog)</code>	iteration log (up to 20 iterations)
<code>e(gradient)</code>	gradient vector
<code>e(V)</code>	variance-covariance matrix of the estimators
<code>e(V_modelbased)</code>	model-based variance

Functions	
<code>e(sample)</code>	marks estimation sample
<b>heckman</b> (two-step) stores the following in <code>e()</code> :	
Scalars	
<code>e(N)</code>	number of observations
<code>e(N_cens)</code>	number of censored observations
<code>e(df_m)</code>	model degrees of freedom
<code>e(lambda)</code>	$\lambda$
<code>e(selambda)</code>	standard error of $\lambda$
<code>e(sigma)</code>	$\sigma$
<code>e(chi2)</code>	$\chi^2$
<code>e(p)</code>	significance of comparison test
<code>e(rho)</code>	$\rho$
<code>e(rank)</code>	rank of <code>e(V)</code>
Macros	
<code>e(cmd)</code>	<b>heckman</b>
<code>e(cmdline)</code>	command as typed
<code>e(depvar)</code>	names of dependent variables
<code>e(title)</code>	title in estimation output
<code>e(title2)</code>	secondary title in estimation output
<code>e(mills)</code>	variable containing nonselection hazard (inverse of Mills's ratio)
<code>e(chi2type)</code>	Wald or LR; type of model $\chi^2$ test
<code>e(vce)</code>	<i>vctype</i> specified in <code>vce()</code>
<code>e(vctype)</code>	title used to label Std. Err.
<code>e(rhometh)</code>	<code>rhosigma</code> , <code>rho trunc</code> , <code>rho limited</code> , or <code>rho force</code>
<code>e(method)</code>	<code>twostep</code>
<code>e(properties)</code>	<code>b V</code>
<code>e(predict)</code>	program used to implement <code>predict</code>
<code>e(marginsok)</code>	predictions allowed by <code>margins</code>
<code>e(marginsnotok)</code>	predictions disallowed by <code>margins</code>
<code>e(asbalanced)</code>	factor variables <code>fvset</code> as <code>asbalanced</code>
<code>e(asobserved)</code>	factor variables <code>fvset</code> as <code>asobserved</code>
Matrices	
<code>e(b)</code>	coefficient vector
<code>e(Cns)</code>	constraints matrix
<code>e(V)</code>	variance–covariance matrix of the estimators
Functions	
<code>e(sample)</code>	marks estimation sample

James Joseph Heckman 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

Cameron and Trivedi (2010, 556–562) and Greene (2012, 873–880) provide good introductions to the Heckman selection model. Adkins and Hill (2011, sec. 16.8) describe the two-step estimator with an application using Stata. Jones (2007, 35–40) illustrates Heckman estimation with an application to health economics.

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

The regression equation is

$$y_j = \mathbf{x}_j\boldsymbol{\beta} + u_{1j}$$

The selection equation is

$$\mathbf{z}_j\boldsymbol{\gamma} + u_{2j} > 0$$

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

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

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,

$$\text{Var}(\lambda) \approx \mathbf{D} \text{Var}\{(\text{atanh } \rho \quad \ln \sigma)\} \mathbf{D}'$$

where  $\mathbf{D}$  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

$$\Pr(y_j \text{ observed} \mid \mathbf{z}_j) = \Phi(\mathbf{z}_j\boldsymbol{\gamma})$$

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

$$m_j = \frac{\phi(\mathbf{z}_j\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z}_j\hat{\boldsymbol{\gamma}})}$$

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

$$\delta_j = m_j(m_j + \hat{\gamma} \mathbf{z}_j)$$

Following Heckman, the two-step parameter estimates of  $\beta$  are obtained by augmenting the regression equation with the nonselection hazard  $\mathbf{m}$ . Thus the regressors become  $[\mathbf{X} \ \mathbf{m}]$ , 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,

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

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

$$\mathbf{V}_{\text{twostep}} = \hat{\sigma}^2 (\mathbf{W}'\mathbf{W})^{-1} (\mathbf{W}'\mathbf{R}\mathbf{W} + \mathbf{Q}) (\mathbf{W}'\mathbf{W})^{-1}$$

where

$$\mathbf{Q} = \hat{\rho}^2 (\mathbf{W}'\mathbf{D}\mathbf{Z}) \mathbf{V}_p (\mathbf{Z}'\mathbf{D}\mathbf{W})$$

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

## References

- Adkins, L. C., and R. C. Hill. 2011. *Using Stata for Principles of Econometrics*. 4th ed. Hoboken, NJ: Wiley.
- Baum, C. F. 2006. *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.
- Cameron, A. C., and P. K. Trivedi. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.
- Chiburis, R., and M. Lokshin. 2007. Maximum likelihood and two-step estimation of an ordered-probit selection model. *Stata Journal* 7: 167–182.
- Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.
- Gronau, R. 1974. Wage comparisons: A selectivity bias. *Journal of Political Economy* 82: 1119–1143.
- 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.
- . 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.
- Jones, A. 2007. *Applied Econometrics for Health Economists: A Practical Guide*. 2nd ed. Abingdon, UK: Radcliffe.
- Lewis, H. G. 1974. Comments on selectivity biases in wage comparisons. *Journal of Political Economy* 82: 1145–1155.
- Manning, W. G., N. Duan, and W. H. Rogers. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 35: 59–82.
- Tauchmann, H. 2014. Lee (2009) treatment-effect bounds for nonrandom sample selection. *Stata Journal* 14: 884–894.



## Also see

[R] [heckman postestimation](#) — Postestimation tools for heckman

[R] [heckoprobit](#) — Ordered probit model with sample selection

[R] [heckprobit](#) — Probit model with sample selection

[R] [regress](#) — Linear regression

[R] [tobit](#) — Tobit regression

[SVY] [svy estimation](#) — Estimation commands for survey data

[TE] [etregress](#) — Linear regression with endogenous treatment effects

[U] [20 Estimation and postestimation commands](#)