Title

stata.com

heckman — Heckman selection model

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)
```

Same as above, and generate `v2` containing the inverse of 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
```

Same 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_s) [twostep]
```

or

```
heckman depvar [indepvars], select(depvar_s = varlist_s) [twostep]
```

Full syntax for maximum likelihood estimates only

```
heckman depvar [indepvars] [if] [in] [weight],
    select([depvar_s =] varlist_s [, noconstant offset(varname_o)])
    [heckman_ml_options]
```

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

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

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

* select( ) is required.

The full specification is select([$depvar_s$ =] $varlist_s$ [, noconstant offset($varname_o$)]).

heckman_ts_options	Description
Model	
* twostep	produce two-step consistent estimate
* select()	specify selection equation: dependent and independent variables; whether to have constant term
noconstant	suppress constant term
rhosigma	truncate ρ to $[-1, 1]$ with consistent σ
rhotrunc	truncate ρ to $[-1, 1]$
rholimited	truncate ρ in limited cases
rhoforce	do not truncate ρ
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
coeflegend	display legend instead of statistics

*twostep and select( ) are required.

The full specification is select([$depvar_s$ =] $varlist_s$ [, noconstant]).

indepvars and $varlist_s$ may contain factor variables; see **[U] 11.4.3 Factor variables**.

depvar, *indepvars*, $varlist_s$, and $depvar_s$ may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

bayes, bootstrap, by, collect, fp, jackknife, rolling, statsby, and svy are allowed; see **[U] 11.1.10 Prefix commands**. For more details, see [BAYES] **bayes: heckman**.

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

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

pweights, fweights, and iweights are allowed with maximum likelihood estimation; see **[U] 11.1.6 weight**. No weights are allowed if twostep is specified.

collinear and coeflegend do 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

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 *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).

`offset(`*varname*$_o$`)` specifies that selection offset *varname*$_o$ be included in the model with the coefficient constrained to be 1.

`noconstant`, `offset(`*varname*`)`, `constraints(`*constraints*`)`; 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.

`lrmodel`; see [R] **Estimation options**.

`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: `noci`, `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 options are available with `heckman` but are not shown in the dialog box:

`collinear`, `coeflegend`; see [R] **Estimation options**.

Options for Heckman selection model (two-step)

Model

`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 ρ 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 ρ be truncated, as with the `rhotrunc` option, and that the estimate of σ be made consistent with $\widehat{\rho}$, the truncated estimate of ρ. So, $\widehat{\sigma} = \boldsymbol{\beta}_m\widehat{\rho}$; see *Methods and formulas* for the definition of $\boldsymbol{\beta}_m$. Both the truncated ρ and the new estimate of $\widehat{\sigma}$ are used in all computations to estimate the two-step covariance matrix.

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

`rholimited` specifies that ρ be truncated only in computing the diagonal matrix $\mathbf{D}$ as it enters $\mathbf{V}_{\text{twostep}}$ and $\mathbf{Q}$; see *Methods and formulas*. In all other computations, the untruncated estimate of ρ is used.

`rhoforce` specifies that the two-step estimate of ρ 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 ρ is in the range $[-1, 1]$.

SE

`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.

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

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

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

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

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

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 ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$), and *varlist*$_s$ 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 ($depvar_s$) 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/r18/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)              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						
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
select						
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

$$\texttt{wage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{age} + u_1$$

and we assumed that wage is observed if

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

where u_1 and u_2 have correlation ρ.

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 ρ; to constrain ρ within its valid limits, and for numerical stability during optimization, it estimates the inverse hyperbolic tangent of ρ:

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

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

Similarly, σ, 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 ρ 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 χ^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 ρ and σ in the forms atanh ρ and $\ln \sigma$ extends the range of these parameters to infinity in both directions, thus avoiding boundary problems during the maximization. Tests of ρ must be made in the transformed units. However, because $\operatorname{atanh}(0) = 0$, 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

This command 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                         Number of obs     =      2,000
(regression model with sample selection)              Selected    =      1,343
                                                      Nonselected =        657
                                                Wald chi2(1)      =          .
Log pseudolikelihood = -5178.304                Prob > chi2       =          .
                                (Std. err. adjusted for 10 clusters in county)
```

wage	Coefficient	Robust 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			.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)              Selected    =      1,343
                                                      Nonselected =        657
                                                Wald chi2(2)      =     442.54
                                                Prob > chi2       =     0.0000
```

wage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
wage						
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
select						
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
/mills						
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 https://www.stata-press.com/data/r18/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                         Number of obs     =        150
(regression model with sample selection)              Selected    =         63
                                                      Nonselected =         87
                                                Wald chi2(3)      =   8.84e+08
Log likelihood = -104.0825                      Prob > chi2       =     0.0000
```

yt	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
yt						
x1	.8974192	.0002164	4146.52	0.000	.896995	.8978434
x2	-2.525303	.0001244	-2.0e+04	0.000	-2.525546	-2.525059
x3	2.855786	.0002695	1.1e+04	0.000	2.855258	2.856314
_cons	.6975003	.0907873	7.68	0.000	.5195604	.8754402
select						
z1	-.6826482	.0889871	-7.67	0.000	-.8570598	-.5082367
z2	1.003678	.1308344	7.67	0.000	.7472471	1.260108
_cons	-.3605665	.1219011	-2.96	0.003	-.5994883	-.1216447
/athrho	16.11489	260.7581	0.06	0.951	-494.9617	527.1914
/lnsigma	-.5396877	.1303548	-4.14	0.000	-.7951785	-.284197
rho	1	1.05e-11			-1	1
sigma	.5829302	.0759878			.4515006	.7526184
lambda	.5829302	.0759878			.433997	.7318635

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

The model has converged to a value of ρ 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 ρ 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 ρ 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 ρ 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 ρ 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)	λ
e(selambda)	standard error of λ
e(sigma)	sigma
e(chi2)	χ^2
e(chi2_c)	χ^2 for comparison test
e(p)	p-value for model test
e(p_c)	p-value for comparison test
e(rho)	ρ
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 χ^2 test
e(chi2_ct)	Wald or LR; type of model χ^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)	λ
e(selambda)	standard error of λ
e(sigma)	sigma
e(chi2)	χ^2
e(p)	p-value for comparison test
e(rho)	ρ
e(rank)	rank of e(V)
e(fconverged)	1 if first-stage model converged, 0 otherwise

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 χ^2 test
e(vce)	*vcetype* 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

Cameron and Trivedi (2022, 974–981) and Greene (2018, 950–957) 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

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

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

In the maximum likelihood estimation, σ and ρ are not directly estimated. Directly estimated are $\ln\sigma$ and atanh ρ:

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

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

$$\mathrm{Var}(\lambda) \approx \mathbf{D}\,\mathrm{Var}\{(\mathrm{atanh}\,\rho \;\; \ln\sigma)\}\,\mathbf{D}'$$

where $\mathbf{D}$ is the Jacobian of λ with respect to atanh ρ 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

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

where ϕ is the normal density. We also define

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

Following Heckman, the two-step parameter estimates of $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}_m$ 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,

$$\widehat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e} + \boldsymbol{\beta}_m^2 \sum_{j=1}^{N} \delta_j}{N}$$

The two-step estimate of ρ is then

$$\widehat{\rho} = \frac{\boldsymbol{\beta}_m}{\widehat{\sigma}}$$

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 $(1 - \widehat{\rho}^2\delta_j)$ as the diagonal elements. The conventional VCE is

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

where

$$\mathbf{Q} = \widehat{\rho}^{\,2}(\mathbf{W}'\mathbf{D}\mathbf{Z})\mathbf{V}_{\mathbf{P}}(\mathbf{Z}'\mathbf{D}\mathbf{W})$$

where $\mathbf{D}$ is the square, diagonal matrix of dimension N with δ_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. 2022. *Microeconometrics Using Stata*. 2nd 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.

Cook, J. A., J.-S. Lee, and N. Newberger. 2021. On identification and estimation of Heckman models. *Stata Journal* 21: 972–998.

D'Haultfœuille, X., A. Maurel, X. Qiu, and Y. Zhang. 2020. Estimating selection models without an instrument with Stata. *Stata Journal* 20: 297–308.

Greene, W. H. 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

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

Heckman, J. 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. https://doi.org/10.2307/1912352.

Jones, A. M. 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. https://doi.org/10.1016/0304-4076(87)90081-9.

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] **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

[CAUSAL] **etregress** — Linear regression with endogenous treatment effects

[ERM] **eregress** — Extended linear regression

[SVY] **svy estimation** — Estimation commands for survey data

[XT] **xtheckman** — Random-effects regression with sample selection

[U] **20 Estimation and postestimation commands**