### Syntax

```stata
heckprobit depvar indepvars [if] [in] [weight],
    select([depvars =] varlist [, noconstant offset(varname)]) [options]
```

#### options
- **Model**
  - `*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
  - `collinear` keep collinear variables

#### 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
- `noskip` perform likelihood-ratio test
- `nocoins` do not display constraints
- `display_options` control 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
- `coeflegend` display legend instead of statistics

*select(*) is required.

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

`indepvars` and `varlist` may contain factor variables; see [U] 11.4.3 Factor variables.

`depvar`, `indepvars`, `depvars`, and `varlist` 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.

`vcetype`, `first`, `noconstant`, and `weights` are not allowed with the `svy` prefix; see [SVY] svy.

`pweights`, `fweights`, and `iweights` are allowed; see [U] 11.1.6 weight.

`coeflegend` does not appear in the dialog box.

See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.
heckprobit — Probit model with sample selection

Menu

Statistics > Sample-selection models > Probit model with selection

Description

heckprobit fits maximum-likelihood probit models with sample selection.

heckprob is a synonym for heckprobit.

Options

Model

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 selection 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, 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), 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.

nocnsreport; see [R] estimation options.

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

Maximization

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

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

The following option is available with \texttt{heckprobit} but is not shown in the dialog box: \texttt{coeflegend}; see \texttt{[R] estimation options}.

Remarks and examples

The probit model with sample selection (Van de Ven and Van Pragg 1981) assumes that there exists an underlying relationship

\[ y_{ij}^* = x_{ij} \beta + u_{1j} \]

such that we observe only the binary outcome

\[ y_{ij}^{\text{probit}} = (y_{ij}^* > 0) \]

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

\[ y_{ij}^{\text{select}} = (z_{ij} \gamma + u_{2j} > 0) \]

where

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

When \( \rho \neq 0 \), standard probit techniques applied to the first equation yield biased results. \texttt{heckprobit} provides consistent, asymptotically efficient estimates for all the parameters in such models.

For the model to be well identified, the selection equation should have at least one variable that is not in the probit equation. Otherwise, the model is identified only by functional form, and the coefficients have no structural interpretation.

Example 1

We use the data from Pindyck and Rubinfeld (1998). In this dataset, the variables are whether children attend private school (\texttt{private}), number of years the family has been at the present residence (\texttt{years}), log of property tax (\texttt{logptax}), log of income (\texttt{loginc}), and whether one voted for an increase in property taxes (\texttt{vote}).

In this example, we alter the meaning of the data. Here we assume that we observe whether children attend private school only if the family votes for increasing the property taxes. This assumption is not true in the dataset, and we make it only to illustrate the use of this command.

We observe whether children attend private school only if the head of household voted for an increase in property taxes. We assume that the vote is affected by the number of years in residence, the current property taxes paid, and the household income. We wish to model whether children are sent to private school on the basis of the number of years spent in the current residence and the current property taxes paid.
. use http://www.stata-press.com/data/r13/school
. heckprob private years logptax, select(vote=years loginc logptax)
Fitting probit model:
Iteration 0:  log likelihood = -17.122381
Iteration 1:  log likelihood = -16.243974
(output omitted)
Iteration 5:  log likelihood = -15.883655
Fitting selection model:
Iteration 0:  log likelihood = -63.036914
Iteration 1:  log likelihood = -58.534843
Iteration 2:  log likelihood = -58.497292
Iteration 3:  log likelihood = -58.497288
Comparison:  log likelihood = -74.380943
Fitting starting values:
Iteration 0:  log likelihood = -40.895684
Iteration 1:  log likelihood = -16.654497
(output omitted)
Iteration 6:  log likelihood = -15.753765
Fitting full model:
Iteration 0:  log likelihood = -75.010619  (not concave)
Iteration 1:  log likelihood = -74.287786
Iteration 2:  log likelihood = -74.250137
Iteration 3:  log likelihood = -74.245088
Iteration 4:  log likelihood = -74.244973
Iteration 5:  log likelihood = -74.244973
Probit model with sample selection
<table>
<thead>
<tr>
<th>Number of obs</th>
<th>Censored obs</th>
<th>Uncensored obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>36</td>
<td>59</td>
</tr>
<tr>
<td>Wald chi2(2)</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-74.24497</td>
<td></td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.5935</td>
<td></td>
</tr>
</tbody>
</table>

|               | Coef.       | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|---------------|-------------|-----------|-------|------|---------------------|
| private       |             |           |       |      |                     |
| years         | -.1142597   | .1461717  | -.78  | 0.434| -.400751 .1722317   |
| logptax       | .3516098    | 1.016485  | 0.35  | 0.729| -1.640665 2.343884  |
| _cons         | -2.780665   | 6.905838  | -.40  | 0.687| -16.31586 10.75453  |
| vote          |             |           |       |      |                     |
| years         | -.0167511   | .0147735  | -1.13 | 0.257| -.0457067 .0122045  |
| loginc        | .9923024    | .4430009  | 2.24  | 0.025| .1240366 1.860568   |
| logptax       | -1.278732   | .5717545  | -2.24 | 0.025| -.2399401 -.1581647 |
| _cons         | -.545821    | 4.070418  | -.13  | 0.893| -8.523694 7.432052  |
| /athrho       | -.8663156   | 1.450028  | -.60  | 0.550| -3.708318 1.975687  |
| rho           | -.6994973   | .7405343  | -.9987984 | .962269 |
| LR test of indep. eqns. (rho = 0): | chi2(1) = | 0.27 | Prob > chi2 = | 0.6020 |
The final iteration log is for fitting the full probit model with sample selection. A likelihood-ratio test of the log likelihood for this model and the comparison log likelihood is presented at the end of the output. If we had specified the vce(robust) option, this test would be presented as a Wald test instead of as a likelihood-ratio test.

Example 2

In example 1, we could have obtained robust standard errors by specifying the vce(robust) option. We do this here and also eliminate the iteration logs by using the nolog option:

```
. heckprob private years logptax, sel(vote=years loginc logptax) vce(robust) nolog
```

```
Probit model with sample selection
Number of obs = 95
Censored obs = 36
Uncensored obs = 59
Wald chi2(2) =  2.55
Log pseudolikelihood = -74.24497 Prob > chi2 = 0.2798

Robust
Coef. Std. Err.  z  P>|z|  [95% Conf. Interval]
private
years  -.1142597  .1113977 -1.03 0.305  -.3325951   .1040758
logptax  .3516098  .7358265  0.48 0.633  -1.090584   1.793803
_cons  -2.780665  4.786678 -0.58 0.561  -12.16238   6.601051
vote
years  -.0167511  .0173344 -0.97 0.334  -.0507259   .0172237
loginc  .9923024  .4228044  2.35 0.019   .1636209   1.820984
logptax  -1.278783  .5095156 -2.51 0.012  -2.277415  -.2801508
_cons  -.545821  4.543892 -0.12 0.904  -9.451686   8.360044
/athrho  -.8663156  1.630643 -0.53 0.595  -4.062318   2.329687
rho  -.6994973  .8327753 -.83 0.407  -.2310276   .884469
```

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.28  Prob > chi2 = 0.6952

Regardless of whether we specify the vce(robust) option, the outcome is not significantly different from the outcome obtained by fitting the probit and selection models separately. This result is not surprising because the selection mechanism estimated was invented for the example rather than borne from any economic theory.
heckprobit stores the following in e():

Scalars

- e(N): number of observations
- e(N_cens): number of censored 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(ll_c): log likelihood, comparison model
- e(N_clust): number of clusters
- e(chi2): \( \chi^2 \)
- e(chi2_c): \( \chi^2 \) for comparison test
- e(p_c): p-value for comparison test
- e(p): significance of 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): heckprobit
- 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(clustvar): name of cluster variable
- e(offset1): offset for regression equation
- e(offset2): offset for selection equation
- e(chi2type): Wald or LR; type of model \( \chi^2 \) test
- e(chi2_ct): type of comparison \( \chi^2 \) test
- 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(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(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
### Methods and formulas

Van de Ven and Van Pragg (1981) provide an introduction and an explanation of this model. The probit equation is

\[ y_j = (x_j\beta + u_{1j} > 0) \]

The selection equation is

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

where

\[ u_1 \sim N(0, 1) \]

\[ u_2 \sim N(0, 1) \]

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

The log likelihood is

\[
\ln L = \sum_{j \in S} w_j \ln \left\{ \Phi_2 \left( x_j\beta + \text{offset}_j^\beta, z_j\gamma + \text{offset}_j^\gamma, \rho \right) \right\} \\
+ \sum_{j \in S} w_j \ln \left\{ \Phi_2 \left( -x_j\beta + \text{offset}_j^\beta, z_j\gamma + \text{offset}_j^\gamma, -\rho \right) \right\} \\
+ \sum_{j \not\in S} w_j \ln \left\{ 1 - \Phi \left( z_j\gamma + \text{offset}_j^\gamma \right) \right\}
\]

where \( S \) is the set of observations for which \( y_j \) is observed, \( \Phi_2(\cdot) \) is the cumulative bivariate normal distribution function (with mean \( [0 \ 0]^\prime \)), \( \Phi(\cdot) \) is the standard cumulative normal, and \( w_j \) is an optional weight for observation \( j \).

In the maximum likelihood estimation, \( \rho \) is not directly estimated. Directly estimated is \( \text{atanh} \rho \):

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

From the form of the likelihood, it is clear that if \( \rho = 0 \), the log likelihood for the probit model with sample selection is equal to the sum of the probit model for the outcome \( y \) and the selection model. We can perform a likelihood-ratio test by comparing the likelihood of the full model with the sum of the log likelihoods for the probit and selection models.

This command supports the Huber/White/sandwich estimator of the variance and its clustered version using \texttt{vce(robust)} and \texttt{vce(cluster clustvar)}, respectively. See [P] \texttt{robust}, particularly \textit{Maximum likelihood estimators} and \textit{Methods and formulas}.

\texttt{heckprobit} also supports estimation with survey data. For details on VCEs with survey data, see [SVY] \textit{variance estimation}.

### References

Baum, C. F. 2006. \textit{An Introduction to Modern Econometrics Using Stata}. College Station, TX: Stata Press.


Also see

[R] **heckprobit postestimation** — Postestimation tools for heckprobit

[R] **heckman** — Heckman selection model

[R] **heckoprobit** — Ordered probit model with sample selection

[R] **probit** — Probit regression

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

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

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