# Title

stata.com

**ivfprobit —** Fractional probit model with continuous endogenous covariates



## Description

`ivfprobit` fits a model for a fractional dependent variable, such as a rate or proportion, where one or more of the covariates are endogenous. The dependent variable must be greater than or equal to 0 and less than or equal to 1. `ivfprobit` assumes all endogenous covariates are continuous.

## Quick start

Fractional probit regression of `y1` on `x` and endogenous regressor `y2` that is instrumented using `z`

```
ivfprobit y1 x (y2 = z)
```

Same as above, but with endogenous regressors `y2` and `y3`, using `z1`, `z2`, and `z3` as instruments

```
ivfprobit y1 x (y2 y3 = z1 z2 z3)
```

## Menu

Statistics > Endogenous covariates > Fractional probit with endogenous covariates

# Syntax

```
ivfprobit depvar [varlist1] (varlist2 = varlistiv) [if] [in] [weight] [, options]
```

*varlist*$_1$ is the list of exogenous variables.

*varlist*$_2$ is the list of endogenous variables.

*varlist*$_{\text{iv}}$ is the list of exogenous variables used with *varlist*$_1$ as instruments for *varlist*$_2$.

| *options* | Description |
|---|---|
| **Model** | |
| noconstant | suppress constant term |
| constraints(*numlist*) | apply specified linear constraints |
| **SE/Robust** | |
| vce(*vcetype*) | *vcetype* may be robust, cluster *clustvar*, bootstrap, or jackknife |
| **Reporting** | |
| level(#) | set confidence level; default is level(95) |
| first | report first-stage regression |
| 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 |
| coeflegend | display legend instead of statistics |

*varlist*$_1$ and *varlist*$_{\text{iv}}$ may contain factor variables; see **[U] 11.4.3 Factor variables**.

*depvar*, *varlist*$_1$, *varlist*$_2$, and *varlist*$_{\text{iv}}$ may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

bootstrap, by, collect, 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**.

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

fweights, iweights, and pweights 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.

# Options

Model

`noconstant`, `constraints(`*numlist*`)`; see [R] **Estimation options**.

SE/Robust

`vce(`*vcetype*`)` specifies the type of standard error reported, which includes types 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` requests that the parameters for the reduced-form equations showing the relationships between the endogenous variables and instruments be displayed. The default is not to show these parameter estimates.

`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**.

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

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

# Remarks and examples

stata.com

Remarks are presented under the following headings:

*Model setup*
*Model identification*

## Model setup

`ivfprobit` fits models for fractional dependent variables when one or more of the covariates is endogenous. Fractional variables can take any value in the interval $[0, 1]$; thus, `ivfprobit` is useful for modeling outcomes such as rates and proportions.

Formally, we can write the model fit by `ivfprobit` as

$$E(y_{1i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{y}_{2i}, u_i) = \Phi(\mathbf{y}_{2i}\boldsymbol{\beta} + \mathbf{x}_{1i}\boldsymbol{\gamma} + u_i) \tag{1}$$

$$\mathbf{y}_{2i} = \mathbf{x}_{i1}\boldsymbol{\Pi}_1 + \mathbf{x}_{i2}\boldsymbol{\Pi}_2 + v_i$$

where subscript $i$ denotes the observation, $\mathbf{y}_{2i}$ is a $1 \times p$ vector of continuous endogenous variables, $\mathbf{x}_{1i}$ is a $1 \times k_1$ vector of exogenous covariates, and $\mathbf{x}_{2i}$ is a $1 \times k_2$ vector of excluded instruments. Endogeneity arises from the possible correlation between $u_i$ and $v_i$. The coefficients in vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the parameters of interest. Matrices $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ contain the coefficients of the first stage for the reduced-form equation.

To obtain parameter estimates, `ivfprobit` maximizes the same likelihood as `ivprobit` but does not require a binary dependent variable and does not require the joint density of the errors in the model to be specified correctly; see [R] **ivprobit** for more information. `ivfprobit` fits the model via quasilikelihood estimation rather than maximum likelihood estimation. The key insight behind quasilikelihood estimation is that we do not need to know the true distribution of the entire model to obtain consistent parameter estimates. In fact, the only requirement is the correct specification of the conditional mean given in (1) after integrating over $u_i$. Specifying the full distribution of the model correctly is required only if we want to obtain asymptotically efficient standard errors from maximum likelihood estimation.

`ivfprobit` does not assume that the true model is a probit model that accounts for endogeneity, such as the one fit by `ivprobit`. Therefore, the standard errors provided by maximum likelihood estimation are not appropriate. Instead, `ivfprobit` takes the maximum quasilikelihood approach and reports robust standard errors by default.

For further discussion on quasilikelihood estimation in the context of fractional regression, see Papke and Wooldridge (1996) and Wooldridge (2010).

## ▷ Example 1

We use a corporate 401(k) participation dataset and fit a fractional probit regression of the 401(k) participation rate (`prate`), on corporate employment size (`ltotemp`) and its square, an indicator of whether the 401(k) is the sole pension plan (`sole`), and plan matching rate (`mrate`). The plan matching rate is endogenous and is instrumented using the age of the plan (`age`) and its square.

Our outcome variable `prate` has values between 0 and 1, including 1,351 firms with participation rates of 1. We assume that the functional form of the expected participation rate, after integrating over $u_i$, is a cumulative normal density as in (1).

We use ivfprobit to fit the fractional probit model, accounting for endogeneity of mrate.

```
. use https://www.stata-press.com/data/r18/401k
(Firm-level data on 401k participation)
. ivfprobit prate c.ltotemp##c.ltotemp i.sole (mrate = c.age##c.age)
Fitting exogenous fractional probit model:
Iteration 0:  Log pseudolikelihood = -1769.7046
Iteration 1:  Log pseudolikelihood = -1675.4223
Iteration 2:  Log pseudolikelihood = -1674.7663
Iteration 3:  Log pseudolikelihood = -1674.7661
Iteration 4:  Log pseudolikelihood = -1674.7661
Fitting full model:
Iteration 0:  Log pseudolikelihood =  -3712.498
Iteration 1:  Log pseudolikelihood = -3712.4767
Iteration 2:  Log pseudolikelihood = -3712.4767
Fractional probit model with endogenous regressors
                                                  Number of obs =  4,075
                                                  Wald chi2(4)  = 907.06
Log pseudolikelihood = -3712.4767                 Prob > chi2   = 0.0000
```

| | Coefficient | Robust std. err. | z | P>\|z\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| mrate | 1.907922 | .0946094 | 20.17 | 0.000 | 1.722491 | 2.093353 |
| ltotemp | -.4229273 | .0744177 | -5.68 | 0.000 | -.5687833 | -.2770713 |
| c.ltotemp# c.ltotemp | .0217492 | .0046476 | 4.68 | 0.000 | .01264 | .0308583 |
| sole Only plan | -.1733119 | .0366136 | -4.73 | 0.000 | -.2450733 | -.1015504 |
| _cons | 1.904103 | .3199032 | 5.95 | 0.000 | 1.277104 | 2.531102 |
| corr(e.mrate, e.prate) | -.5690386 | .0431738 | | | -.6476498 | -.4784406 |
| sd(e.mrate) | .3989664 | .0061807 | | | .3870345 | .4112661 |

```
Wald test of exogeneity: chi2(1) = 102.40                 Prob > chi2 = 0.0000
Endogenous: mrate
Exogenous:  ltotemp c.ltotemp#c.ltotemp 1.sole age c.age#c.age
```

We find a positive effect of mrate on the participation rate. Additionally, we see that the correlation between the unobservables, corr(e.mrate, e.prate), is different from zero. This means there is evidence to support our endogeneity conjecture.

◁

## Model identification

As in the basic linear instrumental-variables model, the order condition for identification requires that the number of excluded exogenous variables (that is, the additional instruments) be at least as great as the number of included endogenous variables ($k_2 \geq p$). ivfprobit checks this for you and issues an error message if the order condition is not met.

ivfprobit, like probit and ivprobit, checks the exogenous and endogenous variables to see if any of them predict the outcome variable perfectly. It will drop any offending variables and observations and then fit the model on the remaining data. Instruments that are perfect predictors do not affect estimation, so they are not checked. See *Model identification* in [R] **probit** for more information.

# Stored results

ivfprobit stores the following in e():

Scalars

| | |
|---|---|
| e(N) | number of 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_dv) | number of dependent variables |
| e(df_m) | model degrees of freedom |
| e(ll) | log likelihood |
| e(N_clust) | number of clusters |
| e(endog_ct) | number of endogenous covariates |
| e(p) | model Wald $p$-value |
| e(p_exog) | exogeneity test Wald $p$-value |
| e(chi2) | model Wald $\chi^2$ |
| e(chi2_exog) | Wald $\chi^2$ test of exogeneity |
| e(rank) | rank of e(V) |
| e(ic) | number of iterations |
| e(rc) | return code |
| e(converged) | 1 if converged, 0 otherwise |

Macros

| | |
|---|---|
| e(cmd) | ivfprobit |
| e(cmdline) | command as typed |
| e(depvar) | name of dependent variable |
| e(endog) | names of endogenous variables |
| e(exog) | names of exogenous variables |
| e(wtype) | weight type |
| e(wexp) | weight expression |
| e(title) | title in estimation output |
| e(clustvar) | name of cluster variable |
| e(chi2type) | Wald; type of model $\chi^2$ 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(estat_cmd) | program used to implement estat |
| e(predict) | program used to implement predict |
| e(footnote) | program used to implement the footnote display |
| e(marginsok) | predictions allowed 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(Sigma) | $\widehat{\mathbf{\Sigma}}$ |
| 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.

## Methods and formulas

See *Methods and formulas* in [R] **ivprobit**.

## References

Papke, L. E., and J. M. Wooldridge. 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11: 619–632. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6⟨619::AID-JAE418⟩3.0.CO;2-1.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

## Also see

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

[R] **fracreg** — Fractional response regression

[R] **gmm** — Generalized method of moments estimation

[R] **ivprobit** — Probit model with continuous endogenous covariates

[R] **ivregress** — Single-equation instrumental-variables regression

[R] **ivtobit** — Tobit model with continuous endogenous covariates

[R] **probit** — Probit regression

[ERM] **eprobit** — Extended probit regression

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

[XT] **xtprobit** — Random-effects and population-averaged probit models

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