**tobit** — Tobit regression

**Description**

`tobit` fits models for continuous responses where the outcome variable is censored. Censoring limits may be fixed for all observations or vary across observations.

**Quick start**

Tobit regression of `y` on `x1` and `x2`, specifying that `y` is censored at the minimum of `y`
```
tobit y x1 x2, ll
```
Same as above, but where the lower-censoring limit is zero
```
tobit y x1 x2, ll(0)
```
Same as above, but specify the lower- and upper-censoring limits
```
tobit y x1 x2, ll(17) ul(34)
```
Same as above, but where `lower` and `upper` are variables containing the censoring limits
```
tobit y x1 x2, ll(lower) ul(upper)
```

**Menu**

Statistics > Linear models and related > Censored regression > Tobit regression
Syntax

```
tobit  depvar  [indepvars]  [if]  [in]  [weight]  [ , options]
```

**options**

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>noconstant</td>
</tr>
<tr>
<td>ll(varname</td>
</tr>
<tr>
<td>ul(varname</td>
</tr>
<tr>
<td>offset(varname)</td>
</tr>
<tr>
<td>constraints(constraints)</td>
</tr>
</tbody>
</table>

**SE/Robust**

```
vce(vcetype)
```

vcetype may be oim, opg, robust, cluster clustvar, bootstrap, or jackknife

**Reporting**

```
level(#)  
nocnsreport
display_options
```

set confidence level; default is level(95)
do not display constraints
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
coeflegend
```

keep collinear variables
display legend instead of statistics

**Model**

*indepvars* may contain factor variables; see [U] 11.4.3 Factor variables.
*depvar* and *indepvars* may contain time-series operators; see [U] 11.4.4 Time-series varlists.
*bayes*, *bootstrap*, *by*, *collect*, *fmm*, *fp*, *jackknife*, *nestreg*, *rolling*, *statsby*, *stepwise*, and *svy* are allowed; see [U] 11.1.10 Prefix commands. For more details, see [BAYES] bayes: tobit and [FMM] fmm: tobit.

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

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

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

aweights, fweights, iweights, and pweights are allowed; see [U] 11.1.6 weight.
collinear and coeflegend do not appear in the dialog box.

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

**Options**

*noconstant*; see [R] Estimation options.

ll(varname |#) and ul(varname |#) indicate the lower and upper limits for censoring, respectively. Observations with *depvar* ≤ ll() are left-censored; observations with *depvar* ≥ ul() are right-censored; and remaining observations are not censored. You do not have to specify the censoring values. If you specify ll, the lower limit is the minimum of *depvar*. If you specify ul, the upper limit is the maximum of *depvar*. 

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

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

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

aweights, fweights, iweights, and pweights are allowed; see [U] 11.1.6 weight.
collinear and coeflegend do not appear in the dialog box.

See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.
offset(varname), constraints(constraints); see [R] Estimation options.

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.

level(#), 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.

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.

The following options are available with tobit but are not shown in the dialog box: collinear, coeflegend; see [R] Estimation options.

Remarks and examples

stata.com

tobit fits a linear regression model for a censored continuous outcome. Censoring occurs when the dependent variable is observed only within a certain range of values. When it is not, we know only that it is either above (right-censoring) or below (left-censoring) the censoring value. Censoring differs from truncation. When the data are truncated, we do not observe either the dependent variable or the covariates; see [R] truncreg.

Censoring may result from study design or may be a result of how the outcome is measured. Right-censoring of data may occur, for example, in income surveys that top code the highest income category. Any respondent that earns the censoring limit or more reports only the value at the limit, and we do not know the respondent’s true income. Left-censoring arises naturally when measurements are obtained from an instrument or a laboratory procedure that has a limit of detection. If we observe a value at the measurement limit, we know the true value is at the limit or below it. tobit allows the censoring limits to be the same for all observations or to vary from observation to observation.

Tobin (1958) originally conceived the tobit model as one of consumption of consumer durables where purchases were left-censored at zero. Contemporary literature treats this and similar cases as a corner solution model. See Wooldridge (2020, sec. 17.2), Long (1997, 196–210), and Maddala and Lahiri (2006, 333–336) for an introduction to the tobit model. Wooldridge (2010, chap. 17 and 19) provides an advanced treatment of censored regression models. Cameron and Trivedi (2022, chap. 19) discuss the tobit model using Stata examples.

The tobit model can be written as the latent regression model $y = x\beta + \epsilon$ with a continuous outcome that is either observed or unobserved. Following Cong (2000), the observed outcome for observation $i$ is defined as
\[
y_i^* = \begin{cases} 
y_i & \text{if } a < y_i < b \\
a & \text{if } y_i \leq a \\
b & \text{if } y_i \geq b 
\end{cases}
\]

where \( a \) is the lower-censoring limit and \( b \) is the upper-censoring limit. The tobit model assumes that the error term is normally distributed; \( \epsilon \sim N(0, \sigma^2 I) \). Depending on the problem at hand, the quantity of interest in a tobit model may be the censored outcome, \( y_i^* \), or the uncensored outcome, \( y_i \). In the measurement instrument scenario above, we may wish to predict the values that fall below the measurement threshold. By contrast, in the consumption of consumer durables scenario above, the latent variable is an artificial construct and the variable of interest is the observed consumer expenditure.

### Example 1: Constant-censoring limit

University administrators want to know the relationship between high school grade point average (GPA) and students’ performance in college. \texttt{gpa.dta} contains fictional data on a cohort of 4,000 college students. College GPA (\texttt{gpa2}) and high school GPA (\texttt{hsgpa}) are measured on a continuous scale between zero and four. The outcome of interest is the student’s college GPA. But, for reasons of confidentiality, GPAs below 2.0 are reported as 2.0. In other words, the outcome is censored on the left.

We believe that GPA is also a function of the logarithm of income of the student’s parents (\texttt{pincome}) and whether or not the student participated in a study-skills program while in college (\texttt{program}).

```
. use https://www.stata-press.com/data/r18/gpa
(High school GPA and performance in college)
. tobit gpa2 hsgpa pincome program, ll
Refining starting values:
Grid node 0:  Log likelihood = -2551.3989
Fitting full model:
Iteration 0:  Log likelihood = -2551.3989
Iteration 1:  Log likelihood = -2065.4023
Iteration 2:  Log likelihood = -2015.8135
Iteration 3:  Log likelihood = -2015.1281
Iteration 4:  Log likelihood = -2015.1258
Iteration 5:  Log likelihood = -2015.1258

Tobit regression
Number of obs = 4,000
Uncensored = 2,794
Limits: Lower =  2
Upper = +inf
LR chi2(3) = 4712.61
Prob > chi2 = 0.0000
Pseudo R2 = 0.5390

| Coefficient | Std. err. | t     | P>|t|   | [95% conf. interval] |
|-------------|-----------|-------|-------|----------------------|
| hsgpa       | .6586311  | .0128699 | 51.18 | 0.000                | .633399   | .6838632 |
| pincome     | .3159297  | .0074568 | 42.37 | 0.000                | .3013103  | .3305491 |
| program     | .5554416  | .0147468 | 37.67 | 0.000                | .5265297  | .5843535 |
| _cons       | -.8902578 | .0478484 | -18.61| 0.000                | -.9840673 | -.7964482 |

Tobit reports the coefficients for the latent regression model. Thus, we can interpret the coefficients just as we would the coefficients from OLS. For example, participation in a study-skills program increases the expected uncensored GPA by 0.56 points.
Example 2: Tobit model for a corner solution

Suppose that we are interested in the number of hours married women spend working for wages, and we treat observations recording zero hours as observed, per the corner-solution approach discussed in Wooldridge (2010, chap. 16). We use the labor supply data extracted by Mroz (1987) from the 1975 PSID for 753 married women. The variable whrs75 records the annual number of hours worked. Forty-three percent of the surveyed women worked zero hours, and the remaining women worked on average 1,303 hours a year.

We regress hours worked on household income excluding wife’s income (nwinc), years of schooling (wedyrs), years of labor market experience (wexper) and its square, age (wifeage), an indicator for the presence of children under 6 years of age at home (k16), and an indicator for the presence of children from 6 to 18 years old at home (k618).

```
. use https://www.stata-press.com/data/r18/mroz87
(1975 PSID data from Mroz, 1987)
. tobit whrs75 nwinc wedyrs wexper c.wexper#c.wexper wifeage k16 k618, ll(0)
Refining starting values:
Grid node 0:  Log likelihood = -3961.1577
Fitting full model:
Iteration 0:  Log likelihood = -3961.1577
Iteration 1:  Log likelihood = -3836.8928
Iteration 2:  Log likelihood = -3819.2637
Iteration 3:  Log likelihood = -3819.0948
Iteration 4:  Log likelihood = -3819.0946
```

```
Tobit regression                Number of obs = 753
                                Uncensored = 428
                                Limits: Lower = 0
                                Left-censored = 325
                                Upper = +inf
                                Right-censored = 0
                                LR chi2(7) = 271.59
                                Prob > chi2 = 0.0000
                                Log likelihood = -3819.0946
                                Pseudo R2 = 0.0343

                                   whrs75
                               Coefficient  Std. err.     t    P>|t|     [95% conf. interval]
                            nwinc    -8.814227    4.459089   -1.98    0.048    -17.56808   -.0603708
                          wedyrs     80.64541    21.58318    3.74    0.000    38.27441   123.0164
                           wexper    131.5640    17.27935    7.61    0.000     97.64211   165.4860
                      c.wexper#c.wexper  -1.864153    .5376606   -3.47    0.001   -2.919661  -.8086455
                          wifeage   -54.40491     7.418483   -7.33    0.000    -68.9685   -39.84133
                          k16    -894.0202    111.8777   -7.99    0.000   -1113.653   -674.3875
                          k618   -16.21805    38.64130   -0.42    0.675    -92.07668   59.64057
                     _cons     965.3068    446.4351     2.16    0.031    88.88827  1841.7250

                      var(e.whrs75)  1258927    93304.48  1088458  1456093
```

Unlike in example 1, we are interested in the marginal effect of the covariates on the observed outcome. We can use `margins` to estimate, for example, the average marginal effect of years of education on the expected value of the actual hours worked.
Average marginal effects
Model VCE: OIM
Expression: E(whrs75|whrs75>0), predict(ystar(0,))
dy/dx wrt: wedyrs

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
<td>std. err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>wedyrs</td>
<td>47.47306</td>
<td>12.6214</td>
<td>3.76</td>
<td>0.000</td>
<td>22.73558</td>
</tr>
</tbody>
</table>

The average marginal effect of years of education on the actual hours worked is 47.47. See [R] tobit postestimation for more examples using margins.

James Tobin (1918–2002) was an American economist who after education and research at Harvard moved to Yale, where he was on the faculty from 1950 to 1988. He made many outstanding contributions to economics and was awarded the Nobel Prize in 1981 “for his analysis of financial markets and their relations to expenditure decisions, employment, production and prices”. He trained in the U.S. Navy with the writer, Herman Wouk, who later fashioned a character after Tobin in the novel The Caine Mutiny (1951): “A mandarin-like midshipman named Tobit, with a domed forehead, measured quiet speech, and a mind like a sponge, was ahead of the field by a spacious percentage.”

### Stored results

tobit stores the following in e():

Scalars

- e(N) number of observations
- e(N_unc) number of uncensored observations
- e(N_l0c) number of left-censored observations
- e(N_rc) number of right-censored observations
- e(k) number of parameters
- e(k_eq) number of equations in e(b)
- e(k_aux) number of auxiliary parameters
- e(df_m) model degrees of freedom
- e(df_r) residual degrees of freedom
- e(r2_p) pseudo-$R^2$
- e(ll) log likelihood
- e(ll_0) log likelihood, constant-only model
- e(N_clust) number of clusters
- e(chi2) $\chi^2$ statistic
- e(F) $F$ statistic
- e(p) $p$-value for model test
- 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)` tobit
- `e(cmdline)` command as typed
- `e(depvar)` name of dependent variable
- `e(llopt)` minimum of `depvar` or contents of `ll()`
- `e(ulopt)` maximum of `depvar` or contents of `ul()`
- `e(wtype)` weight type
- `e(wexp)` weight expression
- `e(covariates)` list of covariates
- `e(title)` title in estimation output
- `e(clustvar)` name of cluster variable
- `e(offset)` linear offset variable
- `e(chi2type)` type of model $\chi^2$ test
- `e(vcetype)` vcetype specified in `vce()`
- `e(vcetitle)` 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)` estimation method:
- `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(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.

Methods and formulas

See Methods and formulas in [R] intreg.

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.

tobit also supports estimation with survey data. For details on VCEs with survey data, see [SVY] Variance estimation.
References


Cameron, A. C., and P. K. Trivedi. 2022. *Microeconometrics Using Stata*. 2nd ed. College Station, TX: Stata Press.


Also see

[R] tobit postestimation — Postestimation tools for tobit
[R] heckman — Heckman selection model
[R] intreg — Interval regression
[R] ivtobit — Tobit model with continuous endogenous covariates
[R] regress — Linear regression
[R] truncreg — Truncated regression
[BAYES] bayes: tobit — Bayesian tobit regression
[FMM] fmm: tobit — Finite mixtures of tobit regression models
[ERM] eintreg — Extended interval regression
[ME] metobit — Multilevel mixed-effects tobit regression
[SVY] svy estimation — Estimation commands for survey data
[XT] xtintreg — Random-effects interval-data regression models
[XT] xttobit — Random-effects tobit models
[U] 20 Estimation and postestimation commands