**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 $x_1$ and $x_2$, specifying that $y$ is censored at the minimum of $y$

```
tobit y x1 x2, ll
```

As above, but where the lower-censoring limit is zero

```
tobit y x1 x2, ll(0)
```

As above, but specify the lower- and upper-censoring limits

```
tobit y x1 x2, ll(17) ul(34)
```

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>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>noconstant</td>
<td>suppress constant term</td>
</tr>
<tr>
<td>ll `(varname</td>
<td>#)`</td>
</tr>
<tr>
<td>ul `(varname</td>
<td>#)`</td>
</tr>
<tr>
<td>offset(varname)</td>
<td>include varname in model with coefficient constrained to 1</td>
</tr>
<tr>
<td>constraints(constraints)</td>
<td>apply specified linear constraints</td>
</tr>
<tr>
<td>collinear</td>
<td>keep collinear variables</td>
</tr>
</tbody>
</table>

### SE/Robust

| vce(vcetype)           | vcetype may be oim, robust, cluster clustvar, bootstrap, or jackknife |

### Reporting

<table>
<thead>
<tr>
<th>level(#)</th>
<th>set confidence level; default is level(95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nocnsreport</td>
<td>do not display constraints</td>
</tr>
<tr>
<td>display_options</td>
<td>control columns and column formats, row spacing, line width,</td>
</tr>
<tr>
<td></td>
<td>display of omitted variables and base and empty cells, and factor-variable labeling</td>
</tr>
</tbody>
</table>

### Maximization

<table>
<thead>
<tr>
<th>maximize_options</th>
<th>control the maximization process; seldom used</th>
</tr>
</thead>
</table>

| coeflegend            | display legend instead of statistics                                     |

### 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, 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. |

| coeflegend does 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`. 
offset(varname), constraints(constraints), collinear; see [R] estimation options.

vce(vcetype) specifies the type of standard error reported, which includes types that are derived
from asymptotic theory (oim), 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, nolabel, 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,
gradiant, showstep, hessian, showtolerance, tolerance(#), ltolerance(#),
nrtolerance(#), and nonrtolerance, and from(init_specs); see [R] maximize. These options
are seldom used.

The following option is available with tobit but is not shown in the dialog box:
coeflegend; see [R] estimation options.

Remarks and 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

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 (2016, sec. 17.2), Long (1997, 196–210), and Maddala and
provides an advanced treatment of censored regression models. Cameron and Trivedi (2010, chap. 16)
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

stata.com
\[
y_i^* = \begin{cases} 
y_i & \text{if } a < y_i < b \\
a & \text{if } y_i \leq a \\
b & \text{if } y_i \geq b
\end{cases}
\]

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 http://www.stata-press.com/data/r15/gpa
    (College GPA)
. 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 Left-censored = 1,206
          upper = +inf Right-censored = 0
LR chi2(3) = 4712.61
Prob > chi2 = 0.0000
Log likelihood = -2015.1258
Pseudo R2 = 0.5390

|          | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------|--------|-----------|-------|------|----------------------|
| gpa2     | 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 |
| var(e.gpa2)| .161703| .0044004  | .1533019 | .1705645|
```

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 http://www.stata-press.com/data/r15/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:

<table>
<thead>
<tr>
<th>Iteration</th>
<th>log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-3961.1577</td>
</tr>
<tr>
<td>1</td>
<td>-3836.8928</td>
</tr>
<tr>
<td>2</td>
<td>-3819.2637</td>
</tr>
<tr>
<td>3</td>
<td>-3819.0948</td>
</tr>
<tr>
<td>4</td>
<td>-3819.0946</td>
</tr>
</tbody>
</table>

Tobit regression

| Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-------|-----------|------|-------|----------------------|
| nwinc | -8.814227 | 4.459089 | -1.98 | 0.048 | -17.56808 | -0.0603708 |
| wedyrs| 80.64541  | 21.58318 | 3.74  | 0.000 | 38.27441 | 123.0164 |
| wexper| 131.564   | 17.27935 | 7.61  | 0.000 | 97.64211 | 165.486 |
| c.wexper#c.wexper | -1.864153 | .5376606 | -3.47 | 0.001 | -2.919661 | -0.8086455 |
| 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 | 38.6413 | -0.42 | 0.675 | -92.07668 | 59.64057 |
| _cons  | 965.3068  | 446.4351 | 2.16  | 0.031 | 88.88827 | 1841.725 |

Log likelihood = -3819.0946

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.
. margins, dydx(wedyrs) predict(ystar(0,.))

Average marginal effects
Model VCE : OIM
Expression : E(whrs75*|whrs75>0), predict(ystar(0,.))
dy/dx w.r.t. : wedyrs

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
</tr>
<tr>
<td>wedyrs</td>
<td>47.47306</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_lc)` 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(k_dv)` number of dependent variables
- `e(df_m)` model degrees of freedom
- `e(df_r)` residual degrees of freedom
- `e(r2_p)` pseudo-$R$-squared
- `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)` minimum 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 χ² 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(method)` estimation 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(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

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


