tpoisson — Truncated Poisson regression

Syntax

tpoisson depvar [ indepvars ] [ if ] [ in ] [ weight ] [ , options ]

options

Description

Model
noconstant
suppress constant term
ll(# | varname)
truncation point; default value is ll(0), zero truncation
exposure(varname_o)
include ln(varname_o) in model with coefficient constrained to 1
offset(varname_o)
include varname_o in model with coefficient constrained to 1
constraints(constraints)
apply specified linear constraints
collinear
keep collinear variables

SE/Robust
vce(vcetype)
vctype may be oim, robust, cluster clustvar, opg, bootstrap, or jackknife

Reporting
level(#)
set confidence level; default is level(95)
irr
report incidence-rate ratios
nocnsreport
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

indeps may contain factor variables; see [U] 11.4.3 Factor variables.
depvar and indeps 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.
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.

Menu

Statistics > Count outcomes > Truncated Poisson regression
tpoisson estimates the parameters of a truncated Poisson model by maximum likelihood. The dependent variable \texttt{depvar} is regressed on \texttt{indepvars}, where \texttt{depvar} is a positive count variable whose values are all above the truncation point.

**Options**

**Model**

- \texttt{noconstant}; see [R] estimation options.
- \texttt{ll(# | varname)} specifies the truncation point, which is a nonnegative integer. The default is zero truncation, \texttt{ll(0)}.
- \texttt{exposure(varname), offset(varname\_o), constraints(constraints), collinear}; see [R] estimation options.

**SE/Robust**

- \texttt{vce(vcetype)} specifies the type of standard error reported, which includes types that are derived from asymptotic theory (\texttt{opg}), that are robust to some kinds of misspecification (\texttt{robust}), that allow for intragroup correlation (\texttt{cluster clustvar}), and that use bootstrap or jackknife methods (\texttt{bootstrap}, \texttt{jackknife}); see [R] vce option.

**Reporting**

- \texttt{irr(#)}; see [R] estimation options.
- \texttt{irr} reports estimated coefficients transformed to incidence-rate ratios, that is, \(e^{\beta_i}\) rather than \(\beta_i\).
- Standard errors and confidence intervals are similarly transformed. This option affects how results are displayed, not how they are estimated. \texttt{irr} may be specified at estimation or when replaying previously estimated results.
- \texttt{nocnsreport}; see [R] estimation options.
- \texttt{display_options: noomitted, vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), fvwrapon(style), cformat(\%fmt), pformat(\%fmt), sformat(\%fmt), and nolstretch}; see [R] estimation options.

**Maximization**

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

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

- \texttt{coeflegend}; see [R] estimation options.
Remarks and examples

Truncated Poisson regression is used to model the number of occurrences of an event when that number is restricted to be above the truncation point. If the dependent variable is not truncated, standard Poisson regression may be more appropriate; see [R] poisson. Truncated Poisson regression was first proposed by Grogger and Carson (1991). For an introduction to Poisson regression, see Cameron and Trivedi (2005, 2010) and Long and Freese (2014). For an introduction to truncated Poisson models, see Cameron and Trivedi (2013) and Long (1997, chap. 8).

Suppose that the patients admitted to a hospital for a given condition form a random sample from a population of interest and that each admitted patient stays at least one day. You are interested in modeling the length of stay of patients in days. The sample is truncated at zero because you only have data on individuals who stayed at least one day. tpoisson accounts for the truncated sample, whereas poisson does not.

Truncation is not the same as censoring. Right-censored Poisson regression was implemented in Stata by Raciborski (2011).

Example 1

Consider the Simonoff (2003) dataset of running shoes for a sample of runners who registered an online running log. A running-shoe marketing executive is interested in knowing how the number of running shoes purchased relates to other factors such as gender, marital status, age, education, income, typical number of runs per week, average miles run per week, and the preferred type of running. These data are naturally truncated at zero. A truncated Poisson model is fit to the number of shoes owned on runs per week, miles run per week, gender, age, and marital status.

No options are needed because zero truncation is the default for tpoisson.

```
. use http://www.stata-press.com/data/r13/runshoes
. tpoisson shoes rpweek mpweek male age married

Truncated Poisson regression
Truncation point:  0
LR chi2(5) = 22.75 Prob > chi2 = 0.0004
Log likelihood = -86.257994 Pseudo R2 = 0.1165

shoes             Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------ -------- -------- -------- -------- -------------------------
   rpweek        .1575811  .1097893    1.44    0.151    -.057602     .3727641
  mpweek        .0210673  .0091113    2.31    0.021     .0032094    .0389252
   male        -.0446134  .2444626   -0.18    0.855    -.4345246    .3647979
   age         -.1283912  .0137786   -9.35    0.000    -.1550012   -.1017812
  married       .0185565  .0137786    1.35    0.178     -.008449     .0455629
     _cons     -1.205844  .6619774   -1.82    0.069    -2.503296    .0916078
```

Using the zero-truncated Poisson regression with these data, only the coefficient on average miles per week is statistically significant at the 5% level.
Example 2

Semiconductor manufacturing requires that silicon wafers be coated with a layer of metal oxide. The depth of this layer is strictly controlled. In this example, a critical oxide layer is designed for $300 \pm 20$ angstroms (Å).

After the oxide layer is coated onto a wafer, the wafer enters a photolithography step in which the lines representing the electrical connections are printed on the oxide and later etched and filled with metal. The widths of these lines are measured. In this example, they are controlled to $90 \pm 5$ micrometers ($\mu$m).

After these and other steps, each wafer is electrically tested at probe. If too many failures are discovered, the wafer is rejected and sent for engineering analysis. In this example, the maximum number of probe failures tolerated for this product is 10.

A major failure at probe has been encountered—88 wafers had more than 10 failures each. The 88 wafers that failed were tested using 4 probe machines. The engineer suspects that the failures were a result of faulty probe machines, poor depth control, or poor line widths. The line widths and depths in these data are the actual measurement minus its specification target, 300 Å for the oxide depths and 90 $\mu$m for the line widths.

The following table tabulates the average failure rate for each probe using Stata’s `mean` command; see [R] mean.

```
. use http://www.stata-press.com/data/r13/probe
. mean failures, over(probe) nolog
```

<table>
<thead>
<tr>
<th>Over</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>failures</td>
<td>1</td>
<td>15.875</td>
<td>1.186293</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>14.95833</td>
<td>0.5912379</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>16.47069</td>
<td>0.9279866</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>23.09677</td>
<td>0.9451117</td>
</tr>
</tbody>
</table>

The 95% confidence intervals in this table suggest that there are about 5–11 additional failures per wafer on probe 4. These are unadjusted for varying line widths and oxide depths. Possibly, probe 4 received the wafers with larger line widths or extreme oxide depths.

Truncated Poisson regression more clearly identifies the root causes for the increased failures by estimating the differences between probes adjusted for the line widths and oxide depths. It also allows us to determine whether the deviations from specifications in line widths or oxide depths might be contributing to the problem.
. tpoisson failures i.probe depth width, ll(10) nolog

Truncated Poisson regression

Number of obs = 88
Truncation point: 10
LR chi2(5) = 73.70
Prob > chi2 = 0.0000
Log likelihood = -239.35746 Pseudo R2 = 0.1334

| failures | Coef. | Std. Err. | z    | P>|z|   | [95% Conf. Interval] |
|----------|-------|-----------|------|-------|---------------------|
| probe    |       |           |      |       |                     |
| 2        | -.1113037 | .1019786  | -1.09 | 0.275 | -.3111781 .0885707  |
| 3        | .0114339  | .1036032  | 0.11  | 0.912 | -.1916245 .2144924  |
| 4        | .4254115  | .0841277  | 5.06  | 0.000 | .2605242 .5902989   |
| depth    | -.0005034 | .0033375  | -0.15 | 0.880 | -.0070447 .006038   |
| width    | .0330225  | .015573   | 2.12  | 0.034 | .0025001 .063545    |
| _cons    | 2.714025  | .0752617  | 36.06 | 0.000 | 2.566515 2.861536   |

The coefficients listed for the probes are testing the null hypothesis: $H_0: \text{probe}_i = \text{probe}_1$, where $i$ equals 2, 3, and 4. Because the only coefficient that is statistically significant is the one for testing for $H_0: \text{probe}_4 = \text{probe}_1$, $p < 0.001$, and because the $p$-values for the other probes are not statistically significant, that is, $p \geq 0.275$, the implication is that there is a difference between probe 4 and the other machines. Because the coefficient for this test is positive, 0.425, the conclusion is that the average failure rate for probe 4, after adjusting for line widths and oxide depths, is higher than the other probes. Possibly, probe 4 needs calibration or the head used with this machine is defective.

Line-width control is statistically significant, $p = 0.034$, but variation in oxide depths is not causing the increased failure rate. The engineer concluded that the sudden increase in failures is the result of two problems. First, probe 4 is malfunctioning, and second, there is a possible lithography or etching problem.
Stated results

`tpoisson` 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(r2_p)`: pseudo-\( R^2 \)-squared
- `e(ll)`: log likelihood
- `e(ll_0)`: log likelihood, constant-only model
- `e(N_clust)`: number of clusters
- `e(chi2)`: \( \chi^2 \)
- `e(p)`: significance
- `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)`: `tpoisson`
- `e(cmdline)`: command as typed
- `e(depvar)`: name of dependent variable
- `e(llopt)`: contents of `ll()`, or 0 if not specified
- `e(type)`: weight type
- `e(wexp)`: weight expression
- `e(title)`: title in estimation output
- `e(clustvar)`: name of cluster variable
- `e(offset)`: linear offset variable
- `e(chi2type)`: Wald or LR; 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 \texttt{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 \texttt{fvset} as \texttt{asbalanced}
- `e(asobserved)`: factor variables \texttt{fvset} as \texttt{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

The conditional probability of observing \( y_j \) events given that \( y_j > \tau_j \), where \( \tau_j \) is the truncation point, is given by

\[
\Pr(Y = y_j | y_j > \tau_j, x_j) = \frac{\exp(-\lambda)\lambda^{y_j}}{y_j!Pr(Y > \tau_j | x_j)}
\]

The log likelihood (with weights \( w_j \) and offsets) is given by

\[
\xi_j = x_j\beta + \text{offset}_j
\]

\[
f(y_j) = \frac{\exp\{-\exp(\xi_j)\}\exp(\xi_j y_j)}{y_j!Pr(Y > \tau_j | \xi_j)}
\]

\[
\ln L = \sum_{j=1}^{n} w_j \left[ -\exp(\xi_j) + \xi_j y_j - \ln(y_j!) - \ln \{Pr(Y > \tau_j | \xi_j)\} \right]
\]

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 \texttt{P _robust}, particularly \texttt{Maximum likelihood estimators} and \texttt{Methods and formulas}.

tpoisson also supports estimation with survey data. For details on variance–covariance estimates with survey data, see \texttt{SVY variance estimation}.

Acknowledgment

We gratefully acknowledge the previous work by Joseph Hilbe (1999) at Arizona State University and past editor of the \textit{Stata Technical Bulletin} and coauthor of the Stata Press book \textit{Generalized Linear Models and Extensions}.

References


——. 2010. \textit{Microeconometrics Using Stata}. Rev. ed. College Station, TX: Stata Press.


Also see

[R] tpoisson postestimation — Postestimation tools for tpoisson
[R] poisson — Poisson regression
[R] nbreg — Negative binomial regression
[R] tnbreg — Truncated negative binomial regression
[R] zinb — Zero-inflated negative binomial regression
[R] zip — Zero-inflated Poisson regression
[SVY] svy estimation — Estimation commands for survey data
[XT] xtpoisson — Fixed-effects, random-effects, and population-averaged Poisson models
[U] 20 Estimation and postestimation commands