Title

stata.com

cpoisson — Censored Poisson regression

Description

`cpoisson` fits a Poisson model of a count dependent variable with some censored values. The command can be used when the dependent variable is left-censored (has a lower limit), is right-censored (has an upper limit), or is interval-censored (has a lower and an upper limit).

Quick start

Censored Poisson regression of `y` on `x` without options `ll()` and `ul()`, equivalent to Poisson regression

```
cpoisson y x
```

Add categorical variable `a` using factor-variable syntax, and specify censoring at an upper limit of 4

```
cpoisson y x i.a, ul(4)
```

Also specify a lower-censoring limit that varies across observations by using the variable `lower`

```
cpoisson y x i.a, ul(4) ll(lower)
```

Add offset variable `v`, and report results as incidence-rate ratios

```
cpoisson y x i.a, ul(4) ll(lower) offset(v) irr
```

Constrain the coefficient for `x` to 2

```
constraint define 1 x=2
cpoisson y x i.a, ul(4) constraints(1)
```

Menu

Statistics > Count outcomes > Censored Poisson regression

Syntax

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

options	Description
Model	
noconstant	suppress constant term
ll(# | *varname*)	specify lower limit for left-censoring
ul(# | *varname*)	specify upper limit for right-censoring
exposure(*varname*$_e$)	include ln(*varname*$_e$) 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*)	*vcetype* may be oim, robust, cluster *clustvar*, 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 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
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**.

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.

Options

Model

noconstant; see [R] **estimation options**.

ll(# | *varname*) and ul(# | *varname*) specify the lower and upper limits for censoring, respectively. You may specify one or both. Observations with *depvar* $\leq$ ll() are left-censored; observations with *depvar* $\geq$ ul() are right-censored; and remaining observations are not censored.

exposure(*varname*$_e$), offset(*varname*$_o$), 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**.

`irr` reports estimated coefficients transformed to incidence-rate ratios, that is, e^{β_i} rather than β_i. Standard errors and confidence intervals are similarly transformed. This option affects how results are displayed, not how they are estimated. `irr` may be specified at estimation or when replaying previously estimated results.

`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**. These options are seldom used.

Setting the optimization type to `technique(bhhh)` resets the default *vcetype* to `vce(opg)`.

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

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

Remarks and examples

stata.com

Censored Poisson regression is a method for analyzing censored count data. One of the most common sources of censored count data is top coding, data that record only the value x when x or greater is observed. Not observing subjects for a sufficient period of time is another common cause.

Censored count data models have been studied by Terza (1985) and Brännäs (1992), among others. For an introduction to censored Poisson regression, see Cameron and Trivedi (2005, 2013) and Winkelmann (2008). Raciborski (2011) discusses a command for right-censored Poisson regression and presented Monte Carlo results indicating that the estimator performs well in finite samples. See Creel and Loomis (1990) and Gurmu and Trivedi (1996) for some examples of survey applications with top coding.

Censored data can be right-censored, left-censored, or interval-censored. Right-censoring occurs when we observe the covariates but only observe that the dependent variable is greater than or equal to an upper limit. Left-censoring occurs when we observe the covariates but only observe that the dependent variable is less than or equal to a lower limit.

When the dependent variable is censored, we must use estimation methods that account for this limitation. If we do not account for censoring when our data are censored, our estimates will not converge to the true values. More formally, failure to control for censoring when it exists leads to inconsistent parameter estimation.

Censored Poisson regression provides an alternative to standard Poisson regression that produces consistent estimates when the dependent variable is censored. If the dependent variable is not censored, standard Poisson regression may be more appropriate; see [R] **poisson**.

Censoring differs from truncation. For censored observations, we observe complete covariate information but only a censored value of the dependent variable. When the data are truncated, we do not observe either the dependent variable or the covariates. Different research designs can give rise to censored data or truncated data.

For example, consider a study about the use of national parks. We could ask a random sample of people in the population how many national parks each has visited in the past year. Suppose we decide to record three for those individuals who visited three or more parks. In this case, individuals who visit four or more parks will have observations that are right-censored at three visits. Now suppose that instead of sending out surveys to a random sample from the population, we ask questions only of individuals who come to parks. We will have no information about individuals who do not visit at least one park, and the data will be truncated at zero visits.

Censoring and truncation are different statistical phenomenon and require different analytic methods. See [R] **tpoisson** for information on truncated Poisson regression.

▷ Example 1: Poisson model with top-coded data

Imagine that we have collected survey data about how many times a household has visited the ABC amusement park from a random sample of households in the state in which ABC is located. Respondents were asked about the number of visits to the park in the last year (`trips`), their income (`income`), and the number of children in the household (`children`). The number of trips recorded in `trips` was top coded at "three or more" visits.

We model right-censored `trips` as a function of `income` and `children`.

```
. use http://www.stata-press.com/data/r14/trips
(Visits to the ABC amusement park)
. cpoisson trips income children, ul(3)
initial:       log likelihood = -620.68749
rescale:       log likelihood = -620.68749
Iteration 0:   log likelihood = -620.68749
Iteration 1:   log likelihood = -600.96763
Iteration 2:   log likelihood = -600.78416
Iteration 3:   log likelihood = -600.78415
Censored Poisson regression                     Number of obs     =        500
                                                LR chi2(2)        =      49.29
Log likelihood = -600.78415                     Prob > chi2       =     0.0000
------------------------------------------------------------------------------
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0740477   .0137653     5.38   0.000     .0470683    .1010272
    children |   .1346922    .028617     4.71   0.000      .078604    .1907805
       _cons |   .0033918   .1455473     0.02   0.981    -.2818756    .2886592
------------------------------------------------------------------------------
           0  left-censored observations
         278     uncensored observations
         222 right-censored observations
```

Both income and the number of children have positive effects on the expected number of trips to the amusement park. The estimated parameters provide the sign, but not the magnitude of the effect, because the model is nonlinear; see [R] **cpoisson postestimation**.

◁

Stored results

cpoisson 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_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(ll_0)	log likelihood, constant-only model
e(N_clust)	number of clusters
e(chi2)	χ^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)	cpoisson
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(llopt)	contents of ll(), if specified
e(ulopt)	contents of ul(), if specified
e(wtype)	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 χ^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(footnote)	program and arguments to display footnote
e(marginsok)	predictions allowed by margins
e(marginsnotok)	predictions disallowed 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

We let y_j be the observed, interval-censored dependent variable for observation j and let y_j^* be the uncensored, latent dependent variable. When y_j is not censored, it is the same as y_j^*. When y_j is censored, only the censoring point is observed. Letting L_j denote the left-censoring point (lower limit) and U_j denote the right-censoring point (upper limit), we see that

$$y_j = \begin{cases} L_j & \text{if } y_j^* \leq L_j \\ y_j^* & \text{if } L_j < y_j^* < U_j \\ U_j & \text{if } y_j^* \geq U_j \end{cases}$$

Note that L_j and U_j may vary over the observations so that individuals may have different left- and right-censoring points.

Although `cpoisson` may be used with data that are left-censored, right-censored, or censored from both sides (which is known as interval-censored), we present the formulas for the interval-censored case because it applies to all three cases.

Let $f(y_j|\mathbf{x}_j)$ denote the probability mass function of the Poisson distribution. Defining $\xi_j = \mathbf{x}_j\boldsymbol{\beta} + \text{offset}_j$ implies that the conditional mean of the uncensored variable is given by $E(y_j^*|\mathbf{x}_j) = \exp(\xi_j)$. The log likelihood for observation j is given by

$$l_j = w_j \left[d_j\{-\exp(\xi_j) + y_j\xi_j - \ln(y_j!)\} + (1-d_j)\ln\left\{1 - \sum_{k=0}^{U_j-1} f(k|\mathbf{x}_j) + \sum_{k=0}^{L_j} f(k|\mathbf{x}_j)\right\}\right]$$

where d_j equals 1 when $L_j < y_j^* < U_j$ and equals 0 when $y_j^* \leq L_j$ or $y_j^* \geq U_j$. The log likelihood is thus

$$\ln L = \sum_{j=1}^{N} l_j$$

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*.

`cpoisson` also supports estimation with survey data. For details on variance–covariance estimates with survey data, see [SVY] **variance estimation**.

Acknowledgment

Rafal Raciborski of StataCorp previously implemented `rcpoisson`, a user-written Stata command for right-censored Poisson regression (Raciborski 2011).

References

Brännäs, K. 1992. Limited dependent Poisson regression. *Journal of the Royal Statistical Society, Series D* 41: 413–423.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

——. 2013. *Regression Analysis of Count Data*. 2nd ed. New York: Cambridge University Press.

Creel, M. D., and J. B. Loomis. 1990. Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. *American Journal of Agricultural Economics* 72: 434–441.

Farbmacher, H. 2011. Estimation of hurdle models for overdispersed count data. *Stata Journal* 11: 82–94.

Gurmu, S., and P. K. Trivedi. 1996. Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics* 14: 469–477.

Hilbe, J. M. 1998. sg91: Robust variance estimators for MLE Poisson and negative binomial regression. *Stata Technical Bulletin* 45: 26–28. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 177–180. College Station, TX: Stata Press.

Hilbe, J. M., and D. H. Judson. 1998. sg94: Right, left, and uncensored Poisson regression. *Stata Technical Bulletin* 46: 18–20. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 186–189. College Station, TX: Stata Press.

Raciborski, R. 2011. Right-censored Poisson regression model. *Stata Journal* 11: 95–105.

Terza, J. V. 1985. A tobit-type estimator for the censored Poisson regression model. *Economics Letters* 18: 361–365.

Winkelmann, R. 2008. *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer.

Also see

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

[R] **nbreg** — Negative binomial regression

[R] **poisson** — Poisson regression

[R] **tnbreg** — Truncated negative binomial regression

[R] **tpoisson** — Truncated Poisson 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**