Right-censored Poisson regression model

Rafal Raciborski
StataCorp
College Station, TX
rraciborski@stata.com

Abstract. I present the `rcpoisson` command for right-censored count-data models with a constant (Terza 1985, *Economics Letters* 18: 361–365) and variable censoring threshold (Caudill and Mixon 1995, *Empirical Economics* 20: 183–196). I show the effects of censoring on estimation results by comparing the censored Poisson model with the uncensored one.

Keywords: st0219, rcpoisson, censoring, count data, Poisson model

1 Introduction

Models that adjust for censoring are required when the values of the dependent variable are available for a restricted range but the values of the independent variables are always observed; see Cameron and Trivedi (1998) and Winkelmann (2008). For example, a researcher who is interested in alcohol consumption patterns among male college students may define binge drinking as "five or more drinks in one sitting" and may code the dependent variable as 0, 1, 2, . . . , 5 or more drinks. In this case, the number of drinks consumed will be censored at five. Some students may have consumed more than five alcoholic beverages, but we will never know. Other examples of censoring include the number of shopping trips (Terza 1985), the number of children in the family (Caudill and Mixon 1995), and the number of doctor visits (Hilbe and Greene 2007).

Applying a traditional Poisson regression model to censored data will produce biased and inconsistent estimates; see Brännäs (1992) for details. Intuitively, when the data are right-censored, large values of the dependent variable are coded as small and the conditional mean of the dependent variable and the marginal effects will be attenuated.

In this article, I introduce the `rcpoisson` command for the estimation of right-censored count data. Section 2 describes the command, section 3 gives an example of cross-border shopping trips, section 4 presents postestimation commands, section 5 presents the results of a simulation study, section 6 presents methods and formulas, section 7 describes saved results, and the conclusion follows.

2 The rcpoisson command

The `rcpoisson` command fits right-censored count data models with a constant (Terza 1985) or a variable censoring threshold (Caudill and Mixon 1995). A variable censoring threshold allows the censoring value to differ across each individual or group—for

instance, in the example above we can add female college students to our study and define binge drinking for females as "3 or more drinks in one sitting".

2.1   Syntax

rcpoisson *depvar* $\lceil$ *indepvars* $\rceil$ $\lceil$ *if* $\rceil$ $\lceil$ *in* $\rceil$ $\lceil$ *weight* $\rceil$, ul$\lceil$ (*#* | *varname*) $\rceil$

   $\lceil$<u>nocon</u>stant <u>e</u>xposure(*varname_e*) <u>off</u>set(*varname_o*)

   <u>constr</u>aints(*constraints*) vce(*vcetype*) <u>l</u>evel(*#*) <u>irr</u> <u>nocnsr</u>eport

   <u>coefl</u>egend *display_options maximize_options* $\rceil$

depvar, *indepvars*, *varname_e*, and *varname_o* may contain time-series operators; see
   [U] **11.4.4 tsvarlist**.

indepvars may contain factor variables; see [U] **11.4.3 fvvarlist**.

fweights, iweights, and pweights are allowed; see [U] **11.1.6 weight**.

bootstrap, by, jackknife, mi estimate, nestreg, rolling, statsby, and stepwise
   are allowed; see [U] **11.1.10 prefix**.

   Weights are not allowed with the bootstrap prefix.

2.2   Options

ul$\lceil$ (*#* | *varname*) $\rceil$ indicates the upper (right) limit for censoring. Observations with
   depvar $\geq$ ul() are right-censored. A constant censoring limit is specified as ul(*#*),
   where *#* is a positive integer. A variable censoring limit is specified as ul(*varname*);
   varname should contain positive integer values. When the option is specified as ul,
   the upper limit is the maximum value of *depvar*. This is a required option.

noconstant suppresses constant terms.

exposure(*varname_e*) includes ln(*varname_e*) in the model with the coefficient con-
   strained to 1.

offset(*varname_o*) includes *varname_o* in the model with the coefficient constrained
   to 1.

constraints(*constraints*) applies specified linear constraints.

vce(*vcetype*) specifies the type of standard error reported. *vcetype* may be oim (the
   default), robust, or cluster *clustvar*.

level(*#*) sets the confidence level. The default is level(95).

irr reports incidence-rate ratios.

nocnsreport suppresses the display of constraints.

`coeflegend` displays a coefficient legend instead of a coefficient table.

display_options control spacing and display of omitted variables and base and empty cells. The options include <u>noomit</u>ted, vsquish, <u>noempty</u>cells, <u>base</u>levels, and <u>allbase</u>levels; see [R] **estimation options**.

maximize_options control the maximization process. They are seldom used. The options include <u>diffi</u>cult, <u>tech</u>nique(*algorithm_spec*), <u>iter</u>ate(#), [<u>no</u>]<u>log</u>, <u>tra</u>ce, <u>grad</u>ient, showstep, <u>hess</u>ian, <u>showtol</u>erance, <u>tol</u>erance(#), <u>ltol</u>erance(#), <u>nrtol</u>erance(#), <u>nonrtol</u>erance, and from(*init_specs*); see [R] **maximize** for details. from() must be specified as a vector, for example, mat b0 = (0,0,0,0), rcpoisson ... from(b0).

2.3   Saved results

The censoring threshold will be returned in `e(ulopt)`. If a variable censoring threshold was specified, the macro will contain the name of the censoring variable; otherwise, the macro will contain the user-specified censoring value or the maximum value of the dependent variable.

Scalars
    e(N)   number of observations
    e(N_rc)   number of right-censored observations
    e(N_unc)   number of uncensored observations
    e(k)   number of parameters
    e(k_eq)   number of equations
    e(k_eq_model)   number of equations in model Wald test
    e(k_dv)   number of dependent variables
    e(k_autoCns)   number of base, empty, and omitted constraints
    e(df_m)   model 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) χ^2 statistic
    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)   `rcpoisson`
    e(cmdline)   command as typed
    e(depvar)   name of dependent variable
    e(ulopt)   contents of `ul()`
    e(wtype)   weight type
    e(wexp)   weight expression
    e(title)   title in estimation output
    e(clustvar)   name of cluster variable
    e(offset)   offset
    e(chi2type)   `Wald` or LR; type of model chi-squared 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(singularHmethod) `m-marquardt` or `hybrid`; method used when Hessian is singular
    e(crittype)   optimization criterion
    e(properties)   `b V`
    e(estat_cmd)   program used to implement `estat`
    e(predict)   program used to implement `predict`
    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

3 Example

To illustrate the effect of censoring, we use data on the frequency of cross-border shopping trips from Slovenia. The respondents were asked about the number of shopping trips they made to another European Union country in the previous 12-month period.[1] Out of 1,025 respondents, 780 made no cross-border shopping trips; 131 made one trip; and 114 made two or more trips. Thus a few more than 11% of observations are right-censored. Table 1 describes each variable used in the analysis.

Table 1. Independent variable description

female	1 if female, 0 otherwise
married	1 if married, remarried, or living with partner; 0 otherwise
under15	number of children under 15 years of age
age	1 if 15–24, 2 if 25–39, 3 if 40–54, 4 if 55+
city	1 if respondent lives in a large town, 0 otherwise
car	1 if respondent owns a car
internet	1 if respondent has an Internet connection at home

In Stata, we fit the right-censored Poisson as follows:

```
. use eb
(Eurobarometer 69.1: Purchasing in the EU, February-March 2008)

. rcpoisson trips i.female i.married under15 i.age i.city i.car i.internet, ul
> nolog
Right-censored Poisson regression              Number of obs   =       1025
                                               LR chi2(9)      =     124.95
                                               Prob > chi2     =     0.0000
Log likelihood = -738.19296                    Pseudo R2       =     0.0780
```

trips	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
1.female	.1383446	.1084766	1.28	0.202	-.0742656	.3509547
1.married	.0546732	.1365587	0.40	0.689	-.2129769	.3223234
under15	.1104834	.0698027	1.58	0.113	-.0263275	.2472942
age						
2	-.2044181	.166461	-1.23	0.219	-.5306756	.1218393
3	-.6731548	.1896838	-3.55	0.000	-1.044928	-.3013814
4	-.8328144	.195075	-4.27	0.000	-1.215154	-.4504743
1.city	-.1590896	.1371978	-1.16	0.246	-.4279924	.1098133
1.car	.3853947	.2423583	1.59	0.112	-.0896188	.8604081
1.internet	.6339803	.1576153	4.02	0.000	.32506	.9429007
_cons	-1.454013	.2815106	-5.17	0.000	-2.005763	-.9022622

```
   Observation summary:      114 right-censored observations (11.1 percent)
                             911     uncensored observations
```

[1]. We use question QC2_1 from the Eurobarometer 69.1, Inter-university Consortium for Political and Social Research, Study No. 25163.

In this case, the `ul` option is equivalent to `ul(2)`—`ul` with no argument tells Stata to treat the maximum value of the dependent variable as the censoring value.

The interpretation of parameters in the censored Poisson model is exactly the same as in the uncensored model. For example, the frequency of trips for people who have an Internet connection at home is `exp(.634)` or 1.89 times larger than for those with no Internet connection. Those rates can be obtained by specifying the `irr` option. Alternatively, we can calculate the percent change in the number of expected trips, which is `(exp(.634)-1)*100` or 89%.

We can compare the censored model with the uncensored one:[2]

```
. poisson trips i.female i.married under15 i.age i.city i.car i.internet, nolog

Poisson regression                              Number of obs   =      1025
                                                LR chi2(9)      =    116.04
                                                Prob > chi2     =    0.0000
Log likelihood = -756.63717                     Pseudo R2       =    0.0712
```

trips	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
1.female	.1309279	.1081285	1.21	0.226	-.0810001	.3428559
1.married	.0524751	.1361917	0.39	0.700	-.2144556	.3194059
under15	.0998008	.0698594	1.43	0.153	-.0371211	.2367228
age						
2	-.179913	.1660251	-1.08	0.279	-.5053162	.1454901
3	-.6324862	.1888077	-3.35	0.001	-1.002542	-.2624299
4	-.788986	.1944563	-4.06	0.000	-1.170113	-.4078586
1.city	-.1525254	.13705	-1.11	0.266	-.4211385	.1160876
1.car	.3942109	.2417194	1.63	0.103	-.0795505	.8679723
1.internet	.6157459	.1574147	3.91	0.000	.3072188	.924273
_cons	-1.517786	.2811689	-5.40	0.000	-2.068867	-.9667054

As can be seen, all coefficients but one returned by uncensored Poisson are smaller than those of the censored Poisson, and so are the marginal effects. The estimate for `internet` falls slightly to 0.616, which translates to an 85% increase in the number of expected trips for respondents with an Internet connection at home.

4   Model evaluation

`rcpoisson` supports all the postestimation commands available to `poisson`, including `estat` and `predict`. The only difference is that, by default, `predict` returns the predicted number of events from the right-censored Poisson distribution. If you want to obtain the predicted number of events from the underlying uncensored distribution, specify the `np` option.

---

2. Equivalently, you can use `rcpoisson` with a `ul()` value greater than the maximum value of the dependent variable. Point estimates will be identical to those obtained with `poisson`, but if you use `margins`, the marginal effects will differ; therefore, I do not advise using `rcpoisson` for the estimation of uncensored models.

Continuing with the censored example above, we can obtain predicted values and marginal effects by typing

```
. predict n
(option n assumed; predicted number of events)
. margins

Predictive margins                              Number of obs   =       1025
Model VCE    : OIM

Expression   : Predicted number of events, predict()
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf. Interval]	
_cons	.3563163	.0176632	20.17	0.000	.321697	.3909356

```
. margins, dydx(_all)

Average marginal effects                        Number of obs   =       1025
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.female 1.married under15 2.age 3.age 4.age 1.city 1.car
               1.internet
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf. Interval]	
1.female	.0458264	.0356648	1.28	0.199	-.0240753	.1157281
1.married	.0181764	.04526	0.40	0.688	-.0705315	.1068843
under15	.0368447	.02326	1.58	0.113	-.008744	.0824335
age						
2	-.0929382	.0774597	-1.20	0.230	-.2447564	.0588801
3	-.2549375	.0769697	-3.31	0.001	-.4057953	-.1040797
4	-.2963611	.0767083	-3.86	0.000	-.4467065	-.1460157
1.city	-.0510675	.0423248	-1.21	0.228	-.1340225	.0318875
1.car	.1124467	.0612255	1.84	0.066	-.0075531	.2324465
1.internet	.1896224	.0423883	4.47	0.000	.1065428	.2727019

```
Note: dy/dx for factor levels is the discrete change from the base level.
```

Thus having an Internet connection at home increases the expected number of cross-border trips by 0.19, holding all the other variables constant.

5   Simulation study

To compare the uncensored and censored Poisson models, we perform a simulation study. True values of the dependent variable y, obtained from a Poisson distribution, are censored at two constant points, and we attempt to recover the true parameters used in the data-generating process. The variables x_1 and x_2 are generated as uncorrelated standard normal variates, and we fix the values of β_1 and β_2 at 1 and -1, respectively. We perform 1,000 replications on sample sizes of 250, 500, and 1,000, and we choose the censoring constants such that the percentages of censored y-values are roughly 10% and 39%. Tables 2 and 3 present the results:

Table 2. Simulation results with about 10% censoring

	n	Mean		Std. Dev.		Std. Err.		Rej. Rate	
		poi	rcpoi	poi	rcpoi	poi	rcpoi	poi	rcpoi
β_{x_1}	250	0.700	1.004	0.072	0.069	0.048	0.069	0.994	0.036
	500	0.694	1.003	0.054	0.048	0.033	0.048	1.00	0.050
	1000	0.689	1.001	0.036	0.034	0.024	0.034	1.00	0.052
β_{x_2}	250	-0.698	-1.003	0.071	0.070	0.048	0.070	0.995	0.036
	500	-0.693	-1.002	0.052	0.049	0.034	0.048	1.00	0.050
	1000	-0.689	-1.000	0.036	0.034	0.023	0.034	1.00	0.045

`poi` and `rcpoi` stand for uncensored and censored Poisson, respectively.

Table 3. Simulation results with about 39% censoring

	n	Mean		Std. Dev.		Std. Err.		Rej. Rate	
		poi	rcpoi	poi	rcpoi	poi	rcpoi	poi	rcpoi
β_{x_1}	250	0.456	1.010	0.053	0.110	0.065	0.110	1.00	0.044
	500	0.453	1.006	0.038	0.083	0.045	0.077	1.00	0.075
	1000	0.453	1.004	0.025	0.053	0.032	0.054	1.00	0.048
β_{x_2}	250	-0.456	-1.012	0.052	0.109	0.064	0.110	1.00	0.039
	500	-0.455	-1.007	0.037	0.081	0.045	0.077	1.00	0.060
	1000	-0.453	-1.004	0.025	0.053	0.032	0.054	1.00	0.045

`poi` and `rcpoi` stand for uncensored and censored Poisson, respectively.

The "Mean" columns report the average of the estimated coefficients over 1,000 simulation runs. The "Std. Dev." columns report the standard deviation of the estimated coefficients, while the "Std. Err." columns report the mean of the standard error of the true parameters. Finally, the "Rej. Rate" columns report the rate at which the true null hypothesis was rejected at the 0.05 level. With the exception of one run, the coverage

is effectively 95% or better for the censored Poisson model. The coverage for the un-censored Poisson model is essentially zero. As can be seen, the bias for the uncensored Poisson model is substantial, even with a small amount of censoring.

6 Methods and formulas

The basics of the censored Poisson model are presented in Cameron and Trivedi (1998) and Winkelmann (2008). Here we assume that the censoring mechanism is independent of the count variable (see Brännäs [1992]).

Consider the probability function of the Poisson random variable:

$$f(y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \qquad i = 1, \dots, n$$

where $\mu_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$, $\mathbf{x}_i$ is a vector of exogenous variables, and $\boldsymbol{\beta}$ is a vector of unknown parameters. In a traditional Poisson setting, we observe all y_i exactly; however, in a censored Poisson model, we observe the true $y_i^\star$ only below a censoring point c_i. Thus

$$y_i = \begin{cases} y_i^\star, & \text{if } y_i^\star < c_i \\ c_i, & \text{if } y_i^\star \geq c_i \end{cases}$$

The censoring point c_i can vary for each observation (Caudill and Mixon 1995). If c is a constant, we have a model with a constant censoring threshold (Terza 1985).

If y_i is censored, we know that

$$\Pr(y_i \geq c_i) = \sum_{j=c_i}^{\infty} \Pr(y_i = j) = \sum_{j=c_i}^{\infty} f(j) = 1 - \sum_{j=0}^{c_i-1} f(j) = 1 - F(c_i - 1)$$

We define an indicator variable d_i such that

$$d_i = \begin{cases} 1, & \text{if } y_i^\star \geq c_i \\ 0, & \text{otherwise} \end{cases}$$

Then the log likelihood function of the sample can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = \log \left[\prod_{i=1}^{n} \{f(y_i)\}^{1-d_i} \{1 - F(c_i - 1)\}^{d_i} \right]$$

$$= \sum_{i=1}^{n} \left[(1 - d_i) \log f(y_i) + d_i \log \{1 - F(c_i - 1)\} \right]$$

The gradient is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ (1 - d_i)(y_i - \mu_i) - d_i c_i \phi_i \right\} \mathbf{x}_i'$$

and the Hessian is

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}^2} = -\sum_{i=1}^{n} \left[(1 - d_i) \mu_i - d_i c_i \left\{ (c_i - \mu_i) \phi_i - c_i \phi_i^2 \right\} \right] \mathbf{x}_i' \mathbf{x}_i$$

where $\phi_i = f(c_i)/1 - F(c_i - 1)$.

Estimation is by maximum likelihood. The initial values are taken from the uncensored Poisson model unless the user provides initial values using `from()`.

The conditional mean is given by

$$E(y_i; \mu_i) = c_i - \sum_{j=0}^{c_i - 1} f(j)(c_i - j)$$

Numerical derivatives of the conditional mean are used for the calculation of the marginal effects and their standard errors.

7 Conclusion

In this article, I introduced the `rcpoisson` command for censored count data. I illustrated the usage on censored survey responses and provided a comparison with the uncensored Poisson model. I showed, through simulation, that the uncensored Poisson model is unable to recover the true values of the parameters from the underlying distribution—something the censored Poisson model was quite successful with.

8 Acknowledgments

I am grateful to David Drukker for his guidance and support. I also thank Jeff Pitblado for his comments and suggestions. Any remaining errors and omissions are mine.

9 References

Brännäs, K. 1992. Limited dependent Poisson regression. *Statistician* 41: 413–423.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Caudill, S. B., and F. G. Mixon, Jr. 1995. Modeling household fertility decisions: Estimation and testing of censored regression models for count data. *Empirical Economics* 20: 183–196.

Hilbe, J. M., and W. H. Greene. 2007. Count response regression models. In *Handbook of Statistics 27: Epidemiology and Medical Statistics*, ed. C. R. Rao, J. P. Miller, and D. C. Rao, 210–252. Amsterdam: Elsevier.

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.

About the author

Rafal Raciborski is an econometrician at StataCorp. In the summer of 2009, he worked as an intern at StataCorp. He produced this project during his internship.