Title

stata.com

poisson — Poisson regression

Syntax

`poisson` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
`noconstant`	suppress constant term
`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*, `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

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

depvar, *indepvars*, *varname*$_e$, and *varname*$_o$ may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

`bootstrap`, `by`, `fp`, `jackknife`, `mfp`, `mi estimate`, `nestreg`, `rolling`, `statsby`, `stepwise`, and `svy` are allowed; see **[U] 11.1.10 Prefix commands**.

`vce(bootstrap)` and `vce(jackknife)` are not allowed with the `mi estimate` prefix; see **[MI] mi estimate**.

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

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

`fweight`s, `iweight`s, and `pweight`s 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 > Poisson regression

Description

poisson fits a Poisson regression of *depvar* on *indepvars*, where *depvar* is a nonnegative count variable.

If you have panel data, see [XT] **xtpoisson**.

Options

Model

noconstant, 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 or stored. irr may be specified at estimation or when replaying previously estimated results.

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

display_options: 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 poisson but is not shown in the dialog box:

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

Remarks and examples

stata.com

The basic idea of Poisson regression was outlined by Coleman (1964, 378–379). See Cameron and Trivedi (2013; 2010, chap. 17) and Johnson, Kemp, and Kotz (2005, chap. 4) for information about the Poisson distribution. See Cameron and Trivedi (2013), Long (1997, chap. 8), Long and Freese (2014, chap. 9), McNeil (1996, chap. 6), and Selvin (2011, chap. 6) for an introduction to Poisson regression. Also see Selvin (2004, chap. 5) for a discussion of the analysis of spatial distributions, which includes a discussion of the Poisson distribution. An early example of Poisson regression was Cochran (1940).

Poisson regression fits models of the number of occurrences (counts) of an event. The Poisson distribution has been applied to diverse events, such as the number of soldiers kicked to death by horses in the Prussian army (von Bortkiewicz 1898); the pattern of hits by buzz bombs launched against London during World War II (Clarke 1946); telephone connections to a wrong number (Thorndike 1926); and disease incidence, typically with respect to time, but occasionally with respect to space. The basic assumptions are as follows:

1. There is a quantity called the *incidence rate* that is the rate at which events occur. Examples are 5 per second, 20 per 1,000 person-years, 17 per square meter, and 38 per cubic centimeter.
2. The incidence rate can be multiplied by *exposure* to obtain the expected number of observed events. For example, a rate of 5 per second multiplied by 30 seconds means that 150 events are expected; a rate of 20 per 1,000 person-years multiplied by 2,000 person-years means that 40 events are expected; and so on.
3. Over very small exposures ϵ, the probability of finding more than one event is small compared with ϵ.
4. Nonoverlapping exposures are mutually independent.

With these assumptions, to find the probability of k events in an exposure of size E, you divide E into n subintervals $E_1, E_2, \dots, E_n$, and approximate the answer as the binomial probability of observing k successes in n trials. If you let $n \to \infty$, you obtain the Poisson distribution.

In the Poisson regression model, the incidence rate for the jth observation is assumed to be given by

$$r_j = e^{\beta_0+\beta_1 x_{1,j}+\cdots+\beta_k x_{k,j}}$$

If E_j is the exposure, the expected number of events, C_j, will be

$$\begin{aligned} C_j &= E_j e^{\beta_0+\beta_1 x_{1,j}+\cdots+\beta_k x_{k,j}} \\ &= e^{\ln(E_j)+\beta_0+\beta_1 x_{1,j}+\cdots+\beta_k x_{k,j}} \end{aligned}$$

This model is fit by `poisson`. Without the `exposure()` or `offset()` options, E_j is assumed to be 1 (equivalent to assuming that exposure is unknown), and controlling for exposure, if necessary, is your responsibility.

Comparing rates is most easily done by calculating *incidence-rate ratios* (IRRs). For instance, what is the relative incidence rate of chromosome interchanges in cells as the intensity of radiation increases; the relative incidence rate of telephone connections to a wrong number as load increases; or the relative incidence rate of deaths due to cancer for females relative to males? That is, you want to hold all the x's in the model constant except one, say, the ith. The IRR for a one-unit change in x_i is

$$\frac{e^{\ln(E)+\beta_1 x_1+\cdots+\beta_i(x_i+1)+\cdots+\beta_k x_k}}{e^{\ln(E)+\beta_1 x_1+\cdots+\beta_i x_i+\cdots+\beta_k x_k}} = e^{\beta_i}$$

More generally, the IRR for a Δx_i change in x_i is $e^{\beta_i \Delta x_i}$. The `lincom` command can be used after `poisson` to display incidence-rate ratios for any group relative to another; see [R] **lincom**.

▷ Example 1

Chatterjee and Hadi (2012, 174) give the number of injury incidents and the proportion of flights for each airline out of the total number of flights from New York for nine major U.S. airlines in one year:

```
. use http://www.stata-press.com/data/r13/airline
. list

     +-------------------------------------+
     | airline   injuries        n   XYZowned |
     |-------------------------------------|
  1. |       1         11   0.0950         1 |
  2. |       2          7   0.1920         0 |
  3. |       3          7   0.0750         0 |
  4. |       4         19   0.2078         0 |
  5. |       5          9   0.1382         0 |
     |-------------------------------------|
  6. |       6          4   0.0540         1 |
  7. |       7          3   0.1292         0 |
  8. |       8          1   0.0503         0 |
  9. |       9          3   0.0629         1 |
     +-------------------------------------+
```

To their data, we have added a fictional variable, `XYZowned`. We will imagine that an accusation is made that the airlines owned by XYZ Company have a higher injury rate.

```
. poisson injuries XYZowned, exposure(n) irr
Iteration 0:   log likelihood = -23.027197
Iteration 1:   log likelihood = -23.027177
Iteration 2:   log likelihood = -23.027177
Poisson regression                                Number of obs   =          9
                                                  LR chi2(1)      =       1.77
                                                  Prob > chi2     =     0.1836
Log likelihood = -23.027177                       Pseudo R2       =     0.0370

------------------------------------------------------------------------------
    injuries |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    XYZowned |   1.463467    .406872     1.37   0.171     .8486578    2.523675
       _cons |   58.04416   8.558145    27.54   0.000     43.47662    77.49281
       ln(n) |          1  (exposure)
------------------------------------------------------------------------------
```

We specified `irr` to see the IRRs rather than the underlying coefficients. We estimate that XYZ Airlines' injury rate is 1.46 times larger than that for other airlines, but the 95% confidence interval is 0.85 to 2.52; we cannot even reject the hypothesis that XYZ Airlines has a lower injury rate.

◁

❑ Technical note

In example 1, we assumed that each airline's exposure was proportional to its fraction of flights out of New York. What if "large" airlines, however, also used larger planes, and so had even more passengers than would be expected, given this measure of exposure? A better measure would be each airline's fraction of passengers on flights out of New York, a number that we do not have. Even so, we suppose that `n` represents this number to some extent, so a better estimate of the effect might be

```
. gen lnN=ln(n)
. poisson injuries XYZowned lnN
Iteration 0:   log likelihood = -22.333875
Iteration 1:   log likelihood = -22.332276
Iteration 2:   log likelihood = -22.332276
Poisson regression                              Number of obs     =          9
                                                LR chi2(2)        =      19.15
                                                Prob > chi2       =     0.0001
Log likelihood = -22.332276                     Pseudo R2         =     0.3001

------------------------------------------------------------------------------
    injuries |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    XYZowned |   .6840667   .3895877     1.76   0.079    -.0795111    1.447645
         lnN |   1.424169   .3725155     3.82   0.000     .6940517    2.154285
       _cons |   4.863891   .7090501     6.86   0.000     3.474178    6.253603
------------------------------------------------------------------------------
```

Here rather than specifying the `exposure()` option, we explicitly included the variable that would normalize for exposure in the model. We did not specify the `irr` option, so we see coefficients rather than IRRs. We started with the model

$$\text{rate} = e^{\beta_0+\beta_1\text{XYZowned}}$$

The observed counts are therefore

$$\text{count} = ne^{\beta_0+\beta_1\text{XYZowned}} = e^{\ln(n)+\beta_0+\beta_1\text{XYZowned}}$$

which amounts to constraining the coefficient on ln(`n`) to 1. This is what was estimated when we specified the `exposure(n)` option. In the above model, we included the normalizing exposure ourselves and, rather than constraining the coefficient to be 1, estimated the coefficient.

The estimated coefficient is 1.42, a respectable distance away from 1, and is consistent with our speculation that larger airlines also use larger airplanes. With this small amount of data, however, we also have a wide confidence interval that includes 1.

Our estimated *coefficient* on `XYZowned` is now 0.684, and the implied IRR is $e^{0.684} \approx 1.98$ (which we could also see by typing `poisson, irr`). The 95% confidence interval for the coefficient still includes 0 (the interval for the IRR includes 1), so although the point estimate is now larger, we still cannot be certain of our results.

Our expert opinion would be that, although there is not enough evidence to support the charge, there is enough evidence to justify collecting more data.

❑

▷ Example 2

In a famous age-specific study of coronary disease deaths among male British doctors, Doll and Hill (1966) reported the following data (reprinted in Rothman, Greenland, and Lash [2008, 264]):

	Smokers		Nonsmokers	
Age	Deaths	Person-years	Deaths	Person-years
35–44	32	52,407	2	18,790
45–54	104	43,248	12	10,673
55–64	206	28,612	28	5,710
65–74	186	12,663	28	2,585
75–84	102	5,317	31	1,462

The first step is to enter these data into Stata, which we have done:

```
. use http://www.stata-press.com/data/r13/dollhill3, clear
. list
```

	agecat	smokes	deaths	pyears
1.	35-44	1	32	52,407
2.	45-54	1	104	43,248
3.	55-64	1	206	28,612
4.	65-74	1	186	12,663
5.	75-84	1	102	5,317
6.	35-44	0	2	18,790
7.	45-54	0	12	10,673
8.	55-64	0	28	5,710
9.	65-74	0	28	2,585
10.	75-84	0	31	1,462

The most "natural" analysis of these data would begin by introducing indicator variables for each age category and one indicator for smoking:

```
. poisson deaths smokes i.agecat, exposure(pyears) irr
Iteration 0:   log likelihood = -33.823284
Iteration 1:   log likelihood = -33.600471
Iteration 2:   log likelihood = -33.600153
Iteration 3:   log likelihood = -33.600153
Poisson regression                                Number of obs   =         10
                                                  LR chi2(5)      =     922.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -33.600153                       Pseudo R2       =     0.9321
```

deaths	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smokes	1.425519	.1530638	3.30	0.001	1.154984	1.759421
agecat						
45-54	4.410584	.8605197	7.61	0.000	3.009011	6.464997
55-64	13.8392	2.542638	14.30	0.000	9.654328	19.83809
65-74	28.51678	5.269878	18.13	0.000	19.85177	40.96395
75-84	40.45121	7.775511	19.25	0.000	27.75326	58.95885
_cons	.0003636	.0000697	-41.30	0.000	.0002497	.0005296
ln(pyears)	1	(exposure)				

In the above, we specified `irr` to obtain IRRs. We estimate that smokers have 1.43 times the mortality rate of nonsmokers. See, however, example 1 in [R] **poisson postestimation**.

◁

Stored results

poisson 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-squared
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)	poisson
e(cmdline)	command as typed
e(depvar)	name of dependent variable
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(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

Methods and formulas

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

$$\Pr(Y=y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

$$\xi_j = \mathbf{x}_j\boldsymbol{\beta} + \mathrm{offset}_j$$

$$f(y_j) = \frac{e^{-\exp(\xi_j)}e^{\xi_j y_j}}{y_j!}$$

$$\ln L = \sum_{j=1}^{n} w_j \left\{ -e^{\xi_j} + \xi_j y_j - \ln(y_j!) \right\}$$

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*.

`poisson` also supports estimation with survey data. For details on VCEs with survey data, see [SVY] **variance estimation**.

Siméon-Denis Poisson (1781–1840) was a French mathematician and physicist who contributed to several fields: his name is perpetuated in Poisson brackets, Poisson's constant, Poisson's differential equation, Poisson's integral, and Poisson's ratio. Among many other results, he produced a version of the law of large numbers. His rather misleadingly titled *Recherches sur la probabilité des jugements* embraces a complete treatise on probability, as the subtitle indicates, including what is now known as the Poisson distribution. That, however, was discovered earlier by the Huguenot–British mathematician Abraham de Moivre (1667–1754).

References

Bru, B. 2001. Siméon-Denis Poisson. In *Statisticians of the Centuries*, ed. C. C. Heyde and E. Seneta, 123–126. New York: Springer.

Cameron, A. C., and P. K. Trivedi. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.

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

Chatterjee, S., and A. S. Hadi. 2012. *Regression Analysis by Example*. 5th ed. New York: Hoboken, NJ.

Clarke, R. D. 1946. An application of the Poisson distribution. *Journal of the Institute of Actuaries* 72: 481.

Cochran, W. G. 1940. The analysis of variance when experimental errors follow the Poisson or binomial laws. *Annals of Mathematical Statistics* 11: 335–347.

——. 1982. *Contributions to Statistics*. New York: Wiley.

Coleman, J. S. 1964. *Introduction to Mathematical Sociology*. New York: Free Press.

Doll, R., and A. B. Hill. 1966. Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. *Journal of the National Cancer Institute, Monographs* 19: 205–268.

Gould, W. W. 2011. Use poisson rather than regress; tell a friend. The Stata Blog: Not Elsewhere Classified. http://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/.

Harris, T., Z. Yang, and J. W. Hardin. 2012. Modeling underdispersed count data with generalized Poisson regression. *Stata Journal* 12: 736–747.

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.

——. 1999. sg102: Zero-truncated Poisson and negative binomial regression. *Stata Technical Bulletin* 47: 37–40. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 233–236. 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.

Johnson, N. L., A. W. Kemp, and S. Kotz. 2005. *Univariate Discrete Distributions*. 3rd ed. New York: Wiley.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Long, J. S., and J. Freese. 2001. Predicted probabilities for count models. *Stata Journal* 1: 51–57.

——. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd ed. College Station, TX: Stata Press.

McNeil, D. 1996. *Epidemiological Research Methods*. Chichester, UK: Wiley.

Miranda, A., and S. Rabe-Hesketh. 2006. Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata Journal* 6: 285–308.

Newman, S. C. 2001. *Biostatistical Methods in Epidemiology*. New York: Wiley.

Poisson, S. D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Paris: Bachelier.

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

Rodríguez, G. 1993. sbe10: An improvement to poisson. *Stata Technical Bulletin* 11: 11–14. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 94–98. College Station, TX: Stata Press.

Rogers, W. H. 1991. sbe1: Poisson regression with rates. *Stata Technical Bulletin* 1: 11–12. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 62–64. College Station, TX: Stata Press.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

Rutherford, E., J. Chadwick, and C. D. Ellis. 1930. *Radiations from Radioactive Substances*. Cambridge: Cambridge University Press.

Rutherford, M. J., P. C. Lambert, and J. R. Thompson. 2010. Age–period–cohort modeling. *Stata Journal* 10: 606–627.

Sasieni, P. D. 2012. Age–period–cohort models in Stata. *Stata Journal* 12: 45–60.

Schonlau, M. 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. *Stata Journal* 5: 330–354.

Selvin, S. 2004. *Statistical Analysis of Epidemiologic Data*. 3rd ed. New York: Oxford University Press.

——. 2011. *Statistical Tools for Epidemiologic Research*. New York: Oxford University Press.

Thorndike, F. 1926. Applications of Poisson's probability summation. *Bell System Technical Journal* 5: 604–624.

Tobías, A., and M. J. Campbell. 1998. sg90: Akaike's information criterion and Schwarz's criterion. *Stata Technical Bulletin* 45: 23–25. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 174–177. College Station, TX: Stata Press.

von Bortkiewicz, L. 1898. *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

Also see

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

[R] **glm** — Generalized linear models

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

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

[R] **zip** — Zero-inflated Poisson regression

[ME] **mepoisson** — Multilevel mixed-effects Poisson regression

[MI] **estimation** — Estimation commands for use with mi estimate

[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