Title

stata.com

tnbreg — Truncated negative binomial regression

Syntax
Menu
Description
Options
Remarks and examples
Stored results
Methods and formulas
Acknowledgment
References
Also see

Syntax

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

options	Description
Model	
`noconstant`	suppress constant term
`ll(`#|*varname*`)`	truncation point; default value is `ll(0)`, zero truncation
`dispersion(mean)`	parameterization of dispersion; the default
`dispersion(constant)`	constant dispersion for all observations
`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)`
`nolrtest`	suppress likelihood-ratio test
`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 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**.

`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 > Truncated negative binomial regression

Description

tnbreg estimates the parameters of a truncated negative binomial model by maximum likelihood. The dependent variable *depvar* is regressed on *indepvars*, where *depvar* is a positive count variable whose values are all above the truncation point.

Options

Model

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

ll(*#* | *varname*) specifies the truncation point, which is a nonnegative integer. The default is zero truncation, ll(0).

dispersion(mean | constant) specifies the parameterization of the model. dispersion(mean), the default, yields a model with dispersion equal to $1+\alpha \exp(\mathbf{x}_j\boldsymbol{\beta}+\mathrm{offset}_j)$; that is, the dispersion is a function of the expected mean: $\exp(\mathbf{x}_j\boldsymbol{\beta}+\mathrm{offset}_j)$. dispersion(constant) has dispersion equal to $1+\delta$; that is, it is a constant for all observations.

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**.

nolrtest suppresses fitting the Poisson model. Without this option, a comparison Poisson model is fit, and the likelihood is used in a likelihood-ratio test of the null hypothesis that the dispersion parameter is zero.

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 tnbreg but is not shown in the dialog box:

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

Remarks and examples

stata.com

Grogger and Carson (1991) showed that overdispersion causes inconsistent estimation of the mean in the truncated Poisson model. To solve this problem, they proposed using the truncated negative binomial model as an alternative. If data are truncated but do not exhibit overdispersion, the truncated Poisson model is more appropriate; see [R] **tpoisson**. For an introduction to negative binomial regression, see Cameron and Trivedi (2005, 2010) and Long and Freese (2014). For an introduction to truncated negative binomial models, see Cameron and Trivedi (2013) and Long (1997, chap. 8).

tnbreg fits the mean-dispersion and the constant-dispersion parameterizations of truncated negative binomial models. These parameterizations extend those implemented in nbreg; see [R] **nbreg**.

▷ Example 1

We illustrate the truncated negative binomial model using the 1997 MedPar dataset (Hilbe 1999). The data are from 1,495 patients in Arizona who were assigned to a diagnostic-related group (DRG) of patients having a ventilator. Length of stay (los), the dependent variable, is a positive integer; it cannot have zero values. The data are truncated because there are no observations on individuals who stayed for zero days.

The objective of this example is to determine whether the length of stay was related to the binary variables: died, hmo, type1, type2, and type3.

The died variable was recorded as a 0 unless the patient died, in which case, it was recorded as a 1. The other variables also adopted this encoding. The hmo variable was set to 1 if the patient belonged to a health maintenance organization (HMO).

The type1–type3 variables indicated the type of admission used for the patient. The type1 variable indicated an emergency admit. The type2 variable indicated an urgent admit—that is, the first available bed. The type3 variable indicated an elective admission. Because type1–type3 were mutually exclusive, only two of the three could be used in the truncated negative binomial regression shown below.

```
. use http://www.stata-press.com/data/r13/medpar
. tnbreg los died hmo type2-type3, vce(cluster provnum) nolog
Truncated negative binomial regression
Truncation point: 0                              Number of obs   =       1495
Dispersion      = mean                           Wald chi2(4)    =      36.01
Log likelihood = -4737.535                       Prob > chi2     =     0.0000
                              (Std. Err. adjusted for 54 clusters in provnum)
```

los	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
died	-.2521884	.061533	-4.10	0.000	-.3727908	-.1315859
hmo	-.0754173	.0533132	-1.41	0.157	-.1799091	.0290746
type2	.2685095	.0666474	4.03	0.000	.137883	.3991359
type3	.7668101	.2183505	3.51	0.000	.338851	1.194769
_cons	2.224028	.034727	64.04	0.000	2.155964	2.292091
/lnalpha	-.630108	.0764019			-.779853	-.480363
alpha	.5325343	.0406866			.4584734	.6185588

Because observations within the same hospital (`provnum`) are likely to be correlated, we specified the `vce(cluster provnum)` option. The results show that whether the patient died in the hospital and the type of admission have significant effects on the patient's length of stay.

◁

▷ Example 2

To illustrate truncated negative binomial regression with more complex data than the previous example, similar data were created from 100 hospitals. Each hospital had its own way of tracking patient data. In particular, hospitals only recorded data from patients with a minimum length of stay, denoted by the variable `minstay`.

Definitions for minimum length of stay varied among hospitals, typically, from 5 to 18 days. The objective of this example is the same as before: to determine whether the length of stay, recorded in `los`, was related to the binary variables: `died`, `hmo`, `type1`, `type2`, and `type3`.

The binary variables encode the same information as in example 1 above. The `minstay` variable was used to allow for varying truncation points.

```
. use http://www.stata-press.com/data/r13/medproviders
. tnbreg los died hmo type2-type3, ll(minstay) vce(cluster hospital) nolog
Truncated negative binomial regression
Truncation points: minstay                      Number of obs   =       2144
Dispersion      = mean                          Wald chi2(4)    =      15.22
Log likelihood = -7864.0928                     Prob > chi2     =     0.0043
                             (Std. Err. adjusted for 100 clusters in hospital)
```

los	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
died	.0781044	.0303596	2.57	0.010	.0186006	.1376081
hmo	-.0731128	.0368897	-1.98	0.047	-.1454152	-.0008104
type2	.0294136	.0390167	0.75	0.451	-.0470578	.1058849
type3	.0626352	.054012	1.16	0.246	-.0432265	.1684969
_cons	3.014964	.0290886	103.65	0.000	2.957951	3.071977
/lnalpha	-.9965131	.082867			-1.158929	-.8340967
alpha	.3691645	.0305916			.313822	.4342666

In this analysis, two variables have a statistically significant relationship with length of stay. On average, patients who died in the hospital had longer lengths of stay ($p = 0.01$). Because the coefficient for HMO is negative, that is, $b_{\text{HMO}} = -0.073$, on average, patients who were insured by an HMO had shorter lengths of stay ($p = 0.047$). The type of admission was not statistically significant ($p > 0.05$).

◁

Stored results

tnbreg stores the following in e():

Scalars

e(N)	number of observations
e(k)	number of parameters
e(k_aux)	number of auxiliary 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(ll_c)	log likelihood, comparison model
e(alpha)	value of alpha
e(N_clust)	number of clusters
e(chi2)	χ^2
e(chi2_c)	χ^2 for comparison test
e(p)	significance
e(rank)	rank of e(V)
e(rank0)	rank of e(V) for constant-only model
e(ic)	number of iterations
e(rc)	return code
e(converged)	1 if converged, 0 otherwise

Macros

e(cmd)	tnbreg
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(llopt)	contents of ll(), or 0 if not 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(chi2_ct)	Wald or LR; type of model χ^2 test corresponding to e(chi2_c)
e(dispers)	mean or constant
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(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

Methods and formulas are presented under the following headings:

Mean-dispersion model
Constant-dispersion model

Mean-dispersion model

A negative binomial distribution can be regarded as a gamma mixture of Poisson random variables. The number of times an event occurs, y_j, is distributed as Poisson$(\nu_j \mu_j)$. That is, its conditional likelihood is

$$f(y_j \mid \nu_j) = \frac{(\nu_j \mu_j)^{y_j} e^{-\nu_j \mu_j}}{\Gamma(y_j + 1)}$$

where $\mu_j = \exp(\mathbf{x}_j \boldsymbol{\beta} + \text{offset}_j)$ and ν_j is an unobserved parameter with a Gamma$(1/\alpha, \alpha)$ density:

$$g(\nu) = \frac{\nu^{(1-\alpha)/\alpha} e^{-\nu/\alpha}}{\alpha^{1/\alpha} \Gamma(1/\alpha)}$$

This gamma distribution has a mean of 1 and a variance of α, where α is our ancillary parameter.

The unconditional likelihood for the jth observation is therefore

$$f(y_j) = \int_0^\infty f(y_j \mid \nu) g(\nu)\, d\nu = \frac{\Gamma(m + y_j)}{\Gamma(y_j + 1)\Gamma(m)}\, p_j^m (1 - p_j)^{y_j}$$

where $p_j = 1/(1 + \alpha \mu_j)$ and $m = 1/\alpha$. Solutions for α are handled by searching for $\ln\alpha$ because α must be greater than zero. The conditional probability of observing y_j events given that y_j is greater than the truncation point τ_j is

$$\Pr(Y = y_j \,|\, y_j > \tau_j, \mathbf{x}_j) = \frac{f(y_j)}{\Pr(Y > \tau_j \,|\, \mathbf{x}_j)}$$

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

$$m = 1/\alpha \qquad p_j = 1/(1 + \alpha \mu_j) \qquad \mu_j = \exp(\mathbf{x}_j \boldsymbol{\beta} + \text{offset}_j)$$

$$\begin{aligned} \ln L = \sum_{j=1}^{n} w_j \Big[& \ln\{\Gamma(m + y_j)\} - \ln\{\Gamma(y_j + 1)\} \\ & - \ln\{\Gamma(m)\} + m \ln(p_j) + y_j \ln(1 - p_j) - \ln\{\Pr(Y > \tau_j \,|\, p_j, m)\} \Big] \end{aligned}$$

Constant-dispersion model

The constant-dispersion model assumes that y_j is conditionally distributed as Poisson(μ_j^*), where $\mu_j^* \sim \text{Gamma}(\mu_j/\delta, \delta)$ for some dispersion parameter δ [by contrast, the mean-dispersion model assumes that $\mu_j^* \sim \text{Gamma}(1/\alpha, \alpha\mu_j)$]. The log likelihood is given by

$$m_j = \mu_j/\delta \qquad p = 1/(1+\delta)$$

$$\ln L = \sum_{j=1}^{n} w_j \Big[\ln\{\Gamma(m_j + y_j)\} - \ln\{\Gamma(y_j+1)\} \\ - \ln\{\Gamma(m_j)\} + m_j \ln(p) + y_j \ln(1-p) - \ln\{\Pr(Y > \tau_j \,|\, p, m_j)\} \Big]$$

with everything else defined as shown above in the calculations for the mean-dispersion model.

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*.

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

Acknowledgment

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

References

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

——. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.

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

Grogger, J. T., and R. T. Carson. 1991. Models for truncated counts. *Journal of Applied Econometrics* 6: 225–238.

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.

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

Long, J. S., and J. Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd ed. College Station, TX: Stata Press.

Simonoff, J. S. 2003. *Analyzing Categorical Data*. New York: Springer.

Also see

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

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

[R] **poisson** — Poisson 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] **xtnbreg** — Fixed-effects, random-effects, & population-averaged negative binomial models

[U] 20 Estimation and postestimation commands