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.

Quick start

Truncated negative binomial regression of y on x with truncation at 0

tnbreg y x

Report incidence-rate ratios

tnbreg y x, irr

Add categorical variable a using factor variable syntax

tnbreg y x i.a

As above, but specify a constant truncation point of 2

tnbreg y x i.a, ll(2)

With exposure variable exp

  tnbreg y x i.a, exposure(exp)

As above, but specifying a variable truncation point stored in variable min

  tnbreg y x i.a, exposure(exp) ll(min)

With cluster–robust standard errors clustering by the levels of cvar

  tnbreg y x i.a, exposure(exp) ll(min) vce(cluster cvar)

Menu

Statistics > Count outcomes > Truncated negative binomial regression
Syntax

\texttt{tnbreg depvar [ indepvars ] [ if ] [ in ] [ weight ] [ , options ]}

\textbf{options} Description

\textbf{Model}

\texttt{noconstant} suppress constant term
\texttt{ll(# | varname)} truncation point; default value is \texttt{ll(0)}, zero truncation
\texttt{dispersion(mean)} parameterization of dispersion; the default
\texttt{dispersion(constant)} constant dispersion for all observations
\texttt{exposure(varname\_e)} include \texttt{ln(varname\_e)} in model with coefficient constrained to 1
\texttt{offset(varname\_o)} include \texttt{varname\_o} in model with coefficient constrained to 1
\texttt{constraints(constraints)} apply specified linear constraints

\textbf{SE/Robust}

\texttt{vce(vcetype)} \texttt{vcetype} may be \texttt{oim}, \texttt{robust}, \texttt{cluster clustvar}, \texttt{opg}, \texttt{bootstrap}, or \texttt{jackknife}

\textbf{Reporting}

\texttt{level(#)} set confidence level; default is \texttt{level(95)}
\texttt{nolrtest} suppress likelihood-ratio test
\texttt{irr} report incidence-rate ratios
\texttt{noconsreport} do not display constraints
\texttt{display_options} control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling

\textbf{Maximization}

\texttt{maximize_options} control the maximization process; seldom used
\texttt{collinear} keep collinear variables
\texttt{coeflegend} display legend instead of statistics

\textit{indepvars} may contain factor variables; see \texttt{[U] 11.4.3 Factor variables}.
\textit{depvar} and \textit{indepvars} may contain time-series operators; see \texttt{[U] 11.4.4 Time-series varlists}.
\texttt{bayes}, \texttt{bootstrap}, \texttt{by}, \texttt{collect}, \texttt{fp}, \texttt{jackknife}, \texttt{rolling}, \texttt{statsby}, and \texttt{svy} are allowed; see \texttt{[U] 11.1.10 Prefix commands}. For more details, see \texttt{[BAYES] bayes: tnbreg}.

Weights are not allowed with the \texttt{bootstrap} prefix; see \texttt{[R] bootstrap}.
\texttt{vce()} and weights are not allowed with the \texttt{svy} prefix; see \texttt{[SVY] svy}.
\texttt{fweights}, \texttt{iweights}, and \texttt{pweights} are allowed; see \texttt{[U] 11.1.6 weight}.
collinear and \texttt{coeflegend} do not appear in the dialog box.
See \texttt{[U] 20 Estimation and postestimation commands} for more capabilities of estimation commands.

\textbf{Options}

\textbf{Model}

\texttt{noconstant}; see \texttt{[R] Estimation options}.
\texttt{ll(# | varname)} specifies the truncation point, which is a nonnegative integer. The default is zero truncation, \texttt{ll(0)}. 
dispersion(mean | constant) specifies the parameterization of the model. dispersion(mean), the default, yields a model with dispersion equal to \(1 + \alpha \exp(x_j \beta + \text{offset}_j)\); that is, the dispersion is a function of the expected mean: \(\exp(x_j \beta + \text{offset}_j)\). dispersion(constant) has dispersion equal to \(1 + \delta\); that is, it is a constant for all observations.

exposure(varname), offset(varname), constraints(constraints); see [R] Estimation options.

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.

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, \(e^{\beta_i}\) rather than \(\beta_i\). Standard errors and confidence intervals are similarly transformed. This option affects how results are displayed, not how they are estimated or stored. irr may be specified at estimation or when replaying previously estimated results.

cnoconsreport; see [R] Estimation options.

display_options: noci, nopvalues, noomitted, vsquish, noemptycells, baselevels, allbaselevels, nolabels, fvwrap(#), cformat(%fmt), pformat(%fmt), sformat(%fmt), and nolstretch; see [R] Estimation options.

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 options are available with tnbreg but are not shown in the dialog box: collinear, 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 \texttt{[R] nbreg}.

\section*{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 (\textit{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: \textit{died}, \textit{hmo}, \textit{type1}, \textit{type2}, and \textit{type3}.

The \textit{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 \textit{hmo} variable was set to 1 if the patient belonged to a health maintenance organization (HMO).

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

```
use https://www.stata-press.com/data/r17/medpar
(Arizona ventilator data)
.tnbreg los died hmo type2-type3, vce(cluster provnum) nolog
```

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>std. err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>died</td>
<td>-.2521884</td>
<td>.061533</td>
<td>-4.10</td>
<td>0.000</td>
<td>-.3727908 -.1315859</td>
</tr>
<tr>
<td>hmo</td>
<td>-.0754173</td>
<td>.0533132</td>
<td>-1.41</td>
<td>0.157</td>
<td>-.1799091 .0290746</td>
</tr>
<tr>
<td>type2</td>
<td>.2685095</td>
<td>.0666474</td>
<td>4.03</td>
<td>0.000</td>
<td>.137883 .3991359</td>
</tr>
<tr>
<td>type3</td>
<td>.7668101</td>
<td>.2183505</td>
<td>3.51</td>
<td>0.000</td>
<td>.338851 1.194769</td>
</tr>
<tr>
<td>_cons</td>
<td>2.224028</td>
<td>.034727</td>
<td>64.04</td>
<td>0.000</td>
<td>2.155964 2.292091</td>
</tr>
</tbody>
</table>

Because observations within the same hospital (\textit{provnum}) are likely to be correlated, we specified the \texttt{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.

\section*{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 \textit{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 \( \text{los} \), was related to the binary variables: \( \text{died} \), 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.

\[
\begin{align*}
\text{. use https://www.stata-press.com/data/r17/medproviders} \\
\text{. tnbreg los died hmo type2-type3, ll(minstay) vce(cluster hospital) nolog}
\end{align*}
\]

Truncated negative binomial regression

<table>
<thead>
<tr>
<th></th>
<th>Number of obs = 2,144</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truncation points:</td>
<td>Wald chi2(4) = 15.22</td>
</tr>
<tr>
<td>minstay</td>
<td>Prob &gt; chi2 = 0.0043</td>
</tr>
<tr>
<td>Dispersion: mean</td>
<td>Log pseudolikelihood = -7864.0928</td>
</tr>
<tr>
<td></td>
<td>Pseudo R2 = 0.0007</td>
</tr>
</tbody>
</table>

(Std. err. adjusted for 100 clusters in hospital)

| los    | Coefficient std. err. | z      | P>|z|  | [95% conf. interval] |
|--------|------------------------|--------|-------|----------------------|
| died   | 0.078104               | 0.0303598 | 2.57  | 0.010               | 0.0185998 | 0.1376081 |
| hmo    | -0.0731132             | 0.0368897 | -1.98 | 0.047               | -0.1454158 | -0.0008107 |
| type2  | 0.0294132              | 0.0390166 | 0.75  | 0.451               | -0.047058 | 0.1058845 |
| type3  | 0.0626348              | 0.0540123 | 1.16  | 0.246               | -0.0432273 | 0.168497 |
| _cons  | 3.014964               | 0.0290895 | 103.64 | 0.000               | 2.95795   | 3.071978  |
| /lnalpha | -0.996512            | 0.0828691 | -1.158932 | -0.8340916 |
| alpha  | 0.3691649              | 0.0305923 |       |                     | 0.313821 | 0.4342688 |

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 \( (p > 0.05) \).
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-$R^2$
- e(ll) log likelihood
- e(ll0) log likelihood, constant-only model
- e(llc) log likelihood, comparison model
- e(alpha) value of alpha
- e(delta) value of delta
- e(N_clust) number of clusters
- e(chi2) $\chi^2$
- e(chi2_c) $\chi^2$ for comparison test
- e(p) p-value for model test
- 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 $\chi^2$ test
- e(chi2_ct) Wald or LR: type of model $\chi^2$ 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
In addition to the above, the following is stored in \( r() \):

- \( \text{r(table)} \) matrix containing the coefficients with their standard errors, test statistics, \( p \)-values, and confidence intervals

Note that results stored in \( r() \) are updated when the command is replayed and will be replaced when any \( r \)-class command is run after the estimation command.

### 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

\[
f(y_j | \nu_j) = \frac{(\nu_j \mu_j)^{y_j} e^{-\nu_j \mu_j}}{\Gamma(y_j + 1)}
\]

where \( \mu_j = \exp(x_j \beta + \text{offset}_j) \) and \( \nu_j \) 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 \( \alpha \), where \( \alpha \) is our ancillary parameter.

The unconditional likelihood for the \( j \)th observation is therefore

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

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

\[
\Pr(Y = y_j \mid y_j > \tau_j, x_j) = \frac{f(y_j)}{\Pr(Y > \tau_j \mid x_j)}
\]

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

\[
\ln L = \sum_{j=1}^n w_j \left[ \ln \{\Gamma(m + y_j)\} - \ln \{\Gamma(y_j + 1)\} - \ln \{\Pr(Y > \tau_j \mid p_j, m)\} \right]
\]

\[
- \ln \{\Gamma(m)\} + m \ln(p_j) + y_j \ln(1 - p_j) - \Pr(Y > \tau_j \mid p_j, m)
\]
Constant-dispersion model

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

$$m_j = \frac{\mu_j}{\delta} \quad p = \frac{1}{1 + \delta}$$

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

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 VCEs with survey data, see [SVY] Variance estimation.

Acknowledgment

We gratefully acknowledge the previous work by Joseph Hilbe (1944–2017) (1999), a former editor of the Stata Technical Bulletin and coauthor of the Stata Press book Generalized Linear Models and Extensions.

References


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


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

Also see

[R] \texttt{tnbreg postestimation} — Postestimation tools for tnbreg

[R] \texttt{nbreg} — Negative binomial regression

[R] \texttt{poisson} — Poisson regression

[R] \texttt{tpoisson} — Truncated Poisson regression

[R] \texttt{znb} — Zero-inflated negative binomial regression

[R] \texttt{zip} — Zero-inflated Poisson regression

[BAYES] \texttt{bayes: tnbreg} — Bayesian truncated negative binomial regression

[SVY] \texttt{svy estimation} — Estimation commands for survey data

[XT] \texttt{xtnbreg} — Fixed-effects, random-effects, & population-averaged negative binomial models

[U] 20 Estimation and postestimation commands