Title

stata.com

expoisson — Exact Poisson regression

Description

`expoisson` fits an exact Poisson regression model, which produces more accurate inference in small samples than standard maximum-likelihood–based Poisson regression. For stratified data, `expoisson` conditions on the number of events in each stratum and is an alternative to fixed-effects Poisson regression.

Quick start

Exact Poisson regression of `y` on `x1`, `x2`, and `x3`

```
expoisson y x1 x2 x3
```

Add exposure variable `evar`

```
expoisson y x1 x2 x3, exposure(evar)
```

Same as above, but condition on values of `x3` to save time and memory

```
expoisson y x1 x2, exposure(evar) condvars(x3)
```

Same as above, and allow more memory for computing the conditional distribution of sufficient statistics

```
expoisson y x1 x2, exposure(evar) condvars(x3) memory(100m)
```

Report incidence-rate ratios rather than coefficients

```
expoisson y x1 x2 x3, irr
```

Report conditional scores tests

```
expoisson y x1 x2 x3, test(score)
```

Fit a model with strata identified by `svar`

```
expoisson y x1 x2 x3, group(svar)
```

Menu

Statistics > Exact statistics > Exact Poisson regression

Syntax

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

options	Description
Model	
`condvars(`*varlist*$_c$`)`	condition on variables in *varlist*$_c$
`group(`*varname*`)`	groups or strata are stratified by unique values of *varname*
`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
Options	
`memory(#[b|k|m|g])`	set limit on memory usage; default is `memory(25m)`
`saving(`*filename*`[, replace])`	save the joint conditional distribution to *filename*
Reporting	
`level(#)`	set confidence level; default is `level(95)`
`irr`	report incidence-rate ratios
`test(`*testopt*`)`	report p-value for observed sufficient statistic, conditional scores test, or conditional probabilities test
`mue(`*varlist*$_m$`)`	compute the median unbiased estimates for *varlist*$_m$
`midp`	use the mid-p-value rule
`[no]log`	display or suppress the enumeration log; default is to display
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
`coeflegend`	display legend instead of statistics

indepvars, *varlist*$_c$, and *varlist*$_m$ may contain factor variables; see **[U] 11.4.3 Factor variables**.

`by`, `collect`, and `statsby` are allowed; see **[U] 11.1.10 Prefix commands**.

`fweights` 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

`condvars(`*varlist*$_c$`)` specifies variables whose parameter estimates are not of interest to you. You can save substantial computer time and memory by moving such variables from *indepvars* to `condvars()`. Understand that you will get the same results for `x1` and `x3` whether you type

```
. expoisson y x1 x2 x3 x4
```

or

```
. expoisson y x1 x3, condvars(x2 x4)
```

`group(`*varname*`)` specifies the variable defining the strata, if any. A constant term is assumed for each stratum identified in *varname*, and the sufficient statistics for *indepvars* are conditioned on the observed number of successes within each group (as well as other variables in the model). The group variable must be integer valued.

exposure(*varname*$_e$), offset(*varname*$_o$); see [R] **Estimation options**.

Options

memory(#[b|k|m|g]) sets a limit on the amount of memory expoisson can use when computing the conditional distribution of the parameter sufficient statistics. The default is memory(25m), where m stands for megabyte, or 1,048,576 bytes. The following are also available: b stands for byte; k stands for kilobyte, which is equal to 1,024 bytes; and g stands for gigabyte, which is equal to 1,024 megabytes. The minimum setting allowed is 1m and the maximum is 2048m or 2g, but do not attempt to use more memory than is available on your computer. Also see the first technical note under example 3 on counting the conditional distribution.

saving(*filename* [, replace]) saves the joint conditional distribution for each independent variable specified in *indepvars*. There is one file for each variable, and it is named using the prefix *filename* with the variable name appended. For example, saving(mydata) with an independent variable named X would generate a data file named mydata_X.dta. Use replace to replace an existing file. Each file contains the conditional distribution for one of the independent variables specified in *indepvars* conditioned on all other *indepvars* and those variables specified in condvars(). There are two variables in each data file: the feasible sufficient statistics for the variable's parameter and their associated weights. The weights variable is named _w_.

Reporting

level(#); see [R] **Estimation options**. The level() option will not work on replay because confidence intervals are based on estimator-specific enumerations. To change the confidence level, you must refit the model.

irr reports estimated coefficients transformed to incidence-rate ratios, that is, $\exp(\beta)$ rather than β. 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.

test(sufficient|score|probability) reports the p-value associated with the observed sufficient statistics, the conditional scores tests, or the conditional probabilities tests, respectively. The default is test(sufficient). All the statistics are computed at estimation time regardless of which is specified. Each statistic may thus also be displayed when replaying results after estimation without having to refit the model; see [R] **expoisson postestimation**.

mue(*varlist*$_m$) specifies that median unbiased estimates (MUEs) be reported for the specified variables. By default, the conditional maximum likelihood estimates (CMLEs) are reported, except for those parameters for which the CMLEs are infinite. Specify mue(_all) if you want MUEs for all the *indepvars*.

midp instructs expoisson to use the mid-p-value rule when computing the MUEs, p-values, and confidence intervals. This adjustment is for the discreteness of the distribution and halves the value of the discrete probability of the observed statistic before adding it to the p-value. The mid-p-value rule cannot be applied to MUEs whose corresponding parameter CMLE is infinite.

log and nolog specify whether to display the enumeration log, which shows the progress of computing the conditional distribution of the sufficient statistics. The enumeration log is displayed by default unless you used set iterlog off to suppress it; see set iterlog in [R] ***set iter***.

display_options: noomitted, vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), fvwrapon(*style*), cformat(%*fmt*), pformat(%*fmt*), and sformat(%*fmt*); see [R] **Estimation options**.

Note that the maximum widths for `cformat()`, `pformat()`, and `sformat()` differ from those widths listed in [R] **Estimation options**. The maximum width for each format is 9 for `expoisson`.

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

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

Remarks and examples

stata.com

Exact Poisson regression estimates the model parameters by using the conditional distributions of the parameters' sufficient statistics, and the resulting parameter estimates are known as CMLEs. Exact Poisson regression is a small-sample alternative to the maximum-likelihood Poisson model. See [R] **poisson** and [XT] **xtpoisson** to obtain maximum likelihood estimates (MLEs) for the Poisson model and the fixed-effects Poisson model.

Let Y_i denote a Poisson random variable where we observe the outcome $Y_i = y_i$, $i = 1, \ldots, n$. Associated with each independent observation is a $1 \times p$ vector of covariates, $\mathbf{x}_i$. We will denote $\mu_i = E[Y_i \mid \mathbf{x}_i]$ and use the log-linear model to model the relationship between Y_i and $\mathbf{x}_i$,

$$\log(\mu_i) = \theta + \mathbf{x}_i\boldsymbol{\beta}$$

where the constant term, θ, and the $p \times 1$ vector of regression parameters, $\boldsymbol{\beta}$, are unknown. The probability of observing $Y_i = y_i$, $i = 1, \ldots, n$, is

$$\Pr(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. The MLEs for θ and $\boldsymbol{\beta}$ maximize the log of this function.

The sufficient statistics for θ and β_j, $j = 1, \ldots, p$, are $M = \sum_{i=1}^{n} Y_i$ and $T_j = \sum_{i=1}^{n} Y_i x_{ij}$, respectively, and we observe $M = m$ and $T_j = t_j$. `expoisson` tallies the conditional distribution for each T_j, given the other sufficient statistics $T_l = t_l$, $l \neq j$ and $M = m$. Denote one of these values to be $t_j^{(k)}$, $k = 1, \ldots, N$, with weight w_k that accounts for all the generated $\mathbf{Y}$ vectors that give rise to $t_j^{(k)}$. The conditional probability of observing $T_j = t_j$ has the form

$$\Pr(T_j = t_j \mid T_l = t_l, l \neq j, M = m) = \frac{w\ e^{t_j\beta_j}}{\sum_k w_k e^{t_j^{(k)}\beta_j}} \tag{1}$$

where the sum is over the subset of $\mathbf{T}$ vectors such that $(T_1^{(k)} = t_1, \ldots, T_j^{(k)} = t_j^{(k)}, \ldots, T_p^{(k)} = t_p)$ and w is the weight associated with the observed $\mathbf{t}$. The CMLE for β_j maximizes the log of this function.

Specifying nuisance variables in `condvars()` prevents `expoisson` from estimating their associated regression coefficients. These variables are still conditional variables when tallying the conditional distribution for the variables in *indepvars*.

Inferences from MLEs rely on asymptotics, and if your sample size is small, these inferences may not be valid. On the other hand, inferences from the CMLEs are exact in that they use the conditional distribution of the sufficient statistics outlined above.

For small datasets, the dependent variable can be completely determined by the data. Here the MLEs and the CMLEs are unbounded. When this occurs, `expoisson` will compute the MUE, the regression estimate that places the observed sufficient statistic at the median of the conditional distribution.

See [R] **exlogistic** for a more thorough discussion of exact estimation and related statistics.

▷ Example 1

Armitage, Berry, and Matthews (2002, 499–501) fit a log-linear model to data containing the number of cerebrovascular accidents experienced by 41 men during a fixed period, each of whom had recovered from a previous cerebrovascular accident and was hypertensive. Sixteen men received treatment, and in the original data, there are three age groups (40–49, 50–59, ≥60), but we pool the first two age groups to simplify the example. Armitage, Berry, and Matthews point out that this was not a controlled trial, but the data are useful to inquire whether there is evidence of fewer accidents for the treatment group and if age may be an important factor. The dependent variable `count` contains the number of accidents, variable `treat` is an indicator for the treatment group (1 = treatment, 0 = control), and variable `age` is an indicator for the age group (0 = 40−59; 1 = ≥60).

First, we load the dataset, list it, and tabulate the cerebrovascular accident counts by treatment and age group.

```
. use https://www.stata-press.com/data/r18/cerebacc
(Cerebrovascular accidents in hypotensive-treated and control groups)
. list
```

	treat	count	age
1.	Control	0	40/59
2.	Control	0	>=60
3.	Control	1	40/59
4.	Control	1	>=60
5.	Control	2	40/59
6.	Control	2	>=60
7.	Control	3	40/59
	(output omitted)		
35.	Treatment	0	40/59
36.	Treatment	0	40/59
37.	Treatment	0	40/59
38.	Treatment	0	40/59
39.	Treatment	1	40/59
40.	Treatment	1	40/59
41.	Treatment	1	40/59

```
. tabulate treat age [fw=count]
```

Hypotensive drug treatment	Age group 40/59	>=60	Total
Control	15	10	25
Treatment	4	0	4
Total	19	10	29

Next, we estimate the CMLE with `expoisson` and, for comparison, the MLE with `poisson`.

```
. expoisson count i.treat i.age

Estimating: 1.treat
Enumerating sample-space combinations:
Observation 1:  Enumerations =         11
Observation 2:  Enumerations =         11
Observation 3:  Enumerations =         11
  (output omitted)
Observation 39: Enumerations =        410
Observation 40: Enumerations =        410
Observation 41: Enumerations =         30

Estimating: 1.age
Enumerating sample-space combinations:
Observation 1:  Enumerations =          5
Observation 2:  Enumerations =         15
Observation 3:  Enumerations =         15
  (output omitted)
Observation 39: Enumerations =        455
Observation 40: Enumerations =        455
Observation 41: Enumerations =         30

Exact Poisson regression                                Number of obs = 41
```

count	Coefficient	Suff.	2*Pr(Suff.)	[95% conf.	interval]
treat					
Treatment	-1.594306	4	0.0026	-3.005089	-.4701708
age					
>=60	-.5112067	10	0.2794	-1.416179	.3429232

```
. poisson count i.treat i.age, nolog

Poisson regression                                      Number of obs =     41
                                                        LR chi2(2)    =  10.64
                                                        Prob > chi2   = 0.0049
Log likelihood = -38.97981                              Pseudo R2     = 0.1201
```

count	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
treat						
Treatment	-1.594306	.5573614	-2.86	0.004	-2.686714	-.5018975
age						
>=60	-.5112067	.4043525	-1.26	0.206	-1.303723	.2813096
_cons	.233344	.2556594	0.91	0.361	-.2677391	.7344271

`expoisson` generates an enumeration log for each independent variable in *indepvars*. The conditional distribution of the parameter sufficient statistic is tallied for each independent variable. The conditional distribution for `treat`, for example, has 30 records containing the weights, w_k, and feasible sufficient statistics, $t^{(k)}_{\text{treat}}$. In essence, the set of points $(w_k, t^{(k)}_{\text{treat}})$, $k = 1, \ldots, 30$, tallied by `expoisson` now become the data to estimate the regression coefficient for `treat`, using (1) as the likelihood. Remember that 1 of the 30 $(w_k, t^{(k)}_{\text{treat}})$ must contain the observed sufficient statistic, $t_{\text{treat}} = \sum_{i=1}^{41} \texttt{treat}_i \times \texttt{count}_i = 4$, and its relative position in the sorted set of points (sorted by $t^{(k)}_{\text{treat}}$) is how the sufficient-statistic p-value is computed. This algorithm is repeated for the `age` variable.

The regression coefficients for `treat` and `age` are numerically identical for both Poisson models. Both models provide evidence that the treatment reduces the rate of cerebrovascular accidents, $\approx e^{-1.59} \approx 0.204$, or a reduction of about 80%. There is no evidence that age plays a role in the rate of accidents.

The results based on the sufficient statistic provide stronger evidence that treatment reduces the rate of cerebrovascular accidents than the corresponding asymptotic statistics. However, the exact confidence intervals are wider than their asymptotic counterparts.

◁

▷ Example 2

Agresti (2013, 129) used the data from Laird and Olivier (1981) to demonstrate the Poisson model for modeling rates. The data consist of patient survival after heart valve replacement operations. The sample consists of 109 patients that are classified by type of heart valve (aortic, mitral) and by age ($<$55, $\geq$55). Follow-up observations cover lengths from 3 to 97 months, and the time at risk, or exposure, is stored in the variable `TAR`. The response is whether the subject died. First, we take a look at the data and then estimate the incidence rates (IRs) with `expoisson` and `poisson`.

```
. use https://www.stata-press.com/data/r18/heartvalve
(Heart valve replacement data)
. list
```

	age	valve	deaths	TAR
1.	<55	Aortic	4	1259
2.	<55	Mitral	1	2082
3.	>=55	Aortic	7	1417
4.	>=55	Mitral	9	1647

The `age` variable is coded 0 for age $<$55 and 1 for age $\geq$55, and the `valve` variable is coded 0 for the aortic valve and 1 for the mitral valve. The total number of deaths, $M = 21$, is small enough that enumerating the conditional distributions for age and valve type is feasible and asymptotic inferences associated with standard maximum-likelihood Poisson regression may be questionable.

```
. expoisson deaths i.age i.valve, exposure(TAR) irr
Estimating: 1.age
Enumerating sample-space combinations:
Observation 1:  Enumerations =         11
Observation 2:  Enumerations =         11
Observation 3:  Enumerations =        132
Observation 4:  Enumerations =         22
Estimating: 1.valve
Enumerating sample-space combinations:
Observation 1:  Enumerations =         17
Observation 2:  Enumerations =         17
Observation 3:  Enumerations =        102
Observation 4:  Enumerations =         22
Exact Poisson regression                                  Number of obs = 4
```

deaths	IRR	Suff.	2*Pr(Suff.)	[95% conf.	interval]
age					
>=55	3.390401	16	0.0194	1.182297	11.86935
valve					
Mitral	.7190197	10	0.5889	.2729881	1.870068
ln(TAR)	1 (exposure)				

```
. poisson deaths i.age i.valve, exposure(TAR) irr nolog
Poisson regression                                Number of obs =      4
                                                  LR chi2(2)    =   7.62
                                                  Prob > chi2   = 0.0222
Log likelihood = -8.1747285                       Pseudo R2     = 0.3178
```

deaths	IRR	Std. err.	z	P>|z|	[95% conf.	interval]
age						
>=55	3.390401	1.741967	2.38	0.017	1.238537	9.280965
valve						
Mitral	.7190197	.3150492	-0.75	0.452	.3046311	1.6971
_cons	.0018142	.0009191	-12.46	0.000	.0006722	.0048968
ln(TAR)	1	(exposure)				

```
Note: _cons estimates baseline incidence rate.
```

The CMLE and the MLE are numerically identical. We have strong evidence that the death rate for the older age group is higher than the younger age group, specifically 3.4 times higher ($p = 0.017$). This means that for every death in the younger group each month, we would expect about three deaths in the older group. The IR estimate for valve type is approximately 0.72, but we do not have enough evidence to claim that it is different from one. The exact Poisson confidence intervals are a bit wider than the asymptotic confidence intervals.

You can use `ir` (see [R] **Epitab**) to estimate IRs and exact confidence intervals for one covariate, and we compare these confidence intervals with those from `expoisson`, where we estimate the IR by using age only.

```
. ir deaths age TAR
Incidence-rate comparison
```

	Age of patient Exposed	Unexposed	Total		
Number of deaths	16	5	21		
Time at risk	3064	3341	6405		
Incidence rate	.0052219	.0014966	.0032787		
	Point estimate		[95% conf.	interval]	
Inc. rate diff.	.0037254		.00085	.0066007	
Inc. rate ratio	3.489295		1.221441	12.17875	(exact)
Attr. frac. ex.	.7134092		.1812948	.9178898	(exact)
Attr. frac. pop	.5435498				

```
Mid-p-values for tests of incidence-rate difference:
  Adj Pr(Exposed Number of deaths <= 16) = 0.9951 (lower one-sided)
  Adj Pr(Exposed Number of deaths >= 16) = 0.0049 (upper one-sided)
                    Two-sided p-value = 0.0099
. expoisson deaths age, exposure(TAR) irr midp nolog
Exact Poisson regression
                                                Number of obs = 4
```

deaths	IRR	Suff.	2*Pr(Suff.)	[95% conf.	interval]
age	3.489295	16	0.0099	1.324926	10.64922
ln(TAR)	1	(exposure)			

```
Note: Mid-p-value computed for the probabilities and CIs.
```

Both `ir` and `expoisson` give identical IRs and p-values. Both report the two-sided exact p-value by using the mid-p-value rule that accounts for the discreteness in the distribution by subtracting $p_{1/2} = \Pr(T = t)/2$ from $p_l = \Pr(T \leq t)$ and $p_g = \Pr(T \geq t)$, computing $2 \times \min(p_l - p_{1/2}, p_g - p_{1/2})$. By default, `expoisson` will not use the mid-p-value rule (when you exclude the `midp` option), and here the two-sided exact p-value would be $2 \times \min(p_l, p_g) = 0.0158$. The confidence intervals differ because `expoisson` uses the mid-p-value rule when computing the confidence intervals, yet `ir` does not. You can verify this by executing `expoisson` without the `midp` option for this example; you will get the same confidence intervals as `ir`.

You can replay `expoisson` to view the conditional scores test or the conditional probabilities test by using the `test()` option.

```
. expoisson, test(score) irr
Exact Poisson regression
                                                Number of obs = 4
```

deaths	IRR	Score	Pr>=Score	[95% conf.	interval]
age	3.489295	6.76528	0.0113	1.324926	10.64922
ln(TAR)	1	(exposure)			

```
Note: Mid-p-value computed for the probabilities and CIs.
```

All the statistics for `expoisson` are defined in *Methods and formulas* of [R] **exlogistic**. Apart from enumerating the conditional distributions for the logistic and Poisson sufficient statistics, computationally, the primary difference between `exlogistic` and `expoisson` is the weighting values in the likelihood for the parameter sufficient statistics. ◁

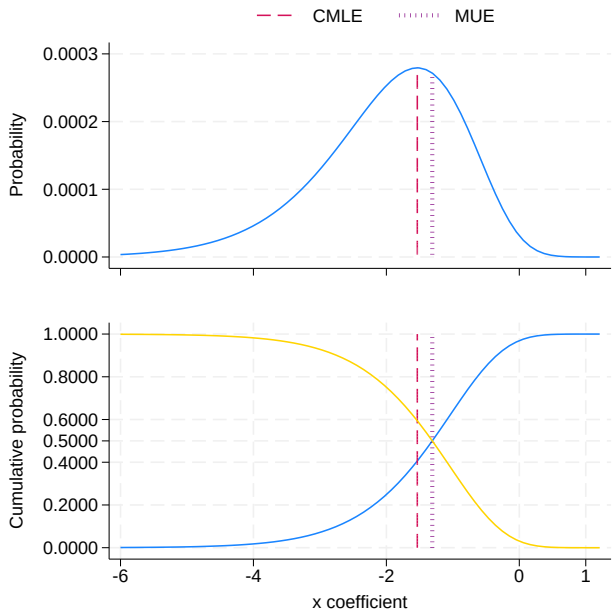
▷ Example 3

In this example, we fabricate data that will demonstrate the difference between the CMLE and the MUE when the CMLE is not infinite. A difference in these estimates will be more pronounced when the probability of the coefficient sufficient statistic is skewed when plotted as a function of the regression coefficient.

```
. clear
. input y x
            y          x
  1. 0 2
  2. 1 1
  3. 1 0
  4. 0 0
  5. 0 .5
  6. 1 .5
  7. 2 .01
  8. 3 .001
  9. 4 .0001
 10. end
. expoisson y x, test(score)
Enumerating sample-space combinations:
Observation 1:  Enumerations =         13
Observation 2:  Enumerations =         91
Observation 3:  Enumerations =        169
Observation 4:  Enumerations =        169
Observation 5:  Enumerations =        313
Observation 6:  Enumerations =        313
Observation 7:  Enumerations =       1469
Observation 8:  Enumerations =       5525
Observation 9:  Enumerations =       5479
Exact Poisson regression                                Number of obs = 9
------------------------------------------------------------------------------
           y | Coefficient      Score    Pr>=Score     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |  -1.534468   2.955316       0.0810    -3.761718    .0485548
------------------------------------------------------------------------------
. expoisson y x, test(score) mue(x) nolog
Exact Poisson regression                                Number of obs = 9
------------------------------------------------------------------------------
           y | Coefficient      Score    Pr>=Score     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |  -1.309268*  2.955316       0.0810    -3.761718    .0485548
------------------------------------------------------------------------------
(*) median unbiased estimates (MUE)
```

We observe (x_i, y_i), $i = 1, \ldots, 9$. If we condition on $m = \sum_{i=1}^{9} y_i = 12$, the conditional distribution of $T_x = \sum_i Y_i x_i$ has a size of 5,479 elements. For each entry in this enumeration, a realization of $Y_i = y_i^{(k)}$, $k = 1, \ldots, 5{,}479$, is generated such that $\sum_i y_i^{(k)} = 12$. One of these realizations produces the observed $t_x = \sum_i y_i x_i \approx 1.5234$.

Below is a graphical display comparing the CMLE with the MUE. We plot $\Pr(T_x = t_x \mid M = 12, \beta_x)$ versus β_x, $-6 \leq \beta_x \leq 1$, in the upper panel and the cumulative probabilities, $\Pr(T_x \leq t_x \mid M = 12, \beta_x)$ and $\Pr(T_x \geq t_x \mid M = 12, \beta_x)$, in the lower panel.



The location of the CMLE, indicated by the dashed line, is at the mode of the probability profile, and the MUE, indicated by the dotted line, is to the right of the mode. If we solve for the $\beta_x^{(u)}$ and $\beta_x^{(l)}$ such that $\Pr(T_x \leq t_x \mid M = 12, \beta_x^{(u)}) = 1/2$ and $\Pr(T_x \geq t_x \mid M = 12, \beta_x^{(l)}) = 1/2$, the MUE is $(\beta_x^{(u)} + \beta_x^{(l)})/2$. As you can see in the lower panel, the MUE cuts through the intersection of these cumulative probability profiles.

◁

❑ Technical note

The memory() option limits the amount of memory that expoisson will consume when computing the conditional distribution of the parameter sufficient statistics. memory() is independent of the data maximum memory setting (see set max_memory in [D] **memory**), and it is possible for expoisson to exceed the memory limit specified in set max_memory without terminating. By default, a log is provided that displays the number of enumerations (the size of the conditional distribution) after processing each observation. Typically, you will see the number of enumerations increase, and then at some point they will decrease as the multivariate shift algorithm (Hirji, Mehta, and Patel 1987) determines that some of the enumerations cannot achieve the observed sufficient statistics of the conditioning variables. When the algorithm is complete, however, it is necessary to store the conditional distribution of the parameter sufficient statistics as a dataset. It is possible, therefore, to get a memory error when the algorithm has completed if there is not enough memory to store the conditional distribution.

❑

❑ Technical note

Computing the conditional distributions and reported statistics requires data sorting and numerical comparisons. If there is at least one single-precision variable specified in the model, expoisson will make comparisons with a relative precision of 2^{-5}. Otherwise, a relative precision of 2^{-11} is used. Be careful if you use recast to promote a single-precision variable to double precision (see [D] **recast**). You might try listing the data in full precision (maybe %20.15g; see [D] **format**) to make sure that this is really what you want. See [D] **Data types** for information on precision of numeric storage types.

❑

Stored results

expoisson stores the following in e():

Scalars

e(N)	number of observations
e(k_groups)	number of groups
e(relative_weight)	relative weight for the observed e(sufficient) and e(condvars)
e(sum_y)	sum of *depvar*
e(k_indvars)	number of independent variables
e(k_condvars)	number of conditioning variables
e(midp)	mid-p-value rule indicator
e(eps)	relative difference tolerance

Macros

e(cmd)	expoisson
e(cmdline)	command as typed
e(title)	title in estimation output
e(depvar)	name of dependent variable
e(indvars)	independent variables
e(condvars)	conditional variables
e(groupvar)	group variable
e(exposure)	exposure variable
e(offset)	linear offset variable
e(level)	confidence level
e(wtype)	weight type
e(wexp)	weight expression
e(datasignature)	the checksum
e(datasignaturevars)	variables used in calculation of checksum
e(properties)	b
e(estat_cmd)	program used to implement estat
e(marginsnotok)	predictions disallowed by margins

Matrices

`e(b)`	coefficient vector
`e(mue_indicators)`	indicator for elements of `e(b)` estimated using MUE instead of CMLE
`e(se)`	`e(b)` standard errors (CMLEs only)
`e(ci)`	matrix of `e(level)` confidence intervals for `e(b)`
`e(sum_y_groups)`	sum of `e(depvar)` for each group
`e(N_g)`	number of observations in each group
`e(sufficient)`	sufficient statistics for `e(b)`
`e(p_sufficient)`	p-value for `e(sufficient)`
`e(scoretest)`	conditional scores tests for *indepvars*
`e(p_scoretest)`	p-values for `e(scoretest)`
`e(probtest)`	conditional probabilities tests for *indepvars*
`e(p_probtest)`	p-value for `e(probtest)`

Functions

`e(sample)`	marks estimation sample

Methods and formulas

Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a set of n independent Poisson random variables. For each $i = 1, \ldots, n$, we observe $Y_i = y_i \geq 0$, and associated with each observation is the covariate row vector of length p, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. Denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ to be the column vector of regression parameters and θ to be the constant. The sufficient statistic for β_j is $T_j = \sum_{i=1}^n Y_i x_{ij}$, $j = 1, \ldots, p$, and for θ is $M = \sum_{i=1}^n Y_i$. We observe $T_j = t_j$, $t_j = \sum_{i=1}^n y_i x_{ij}$, and $M = m$, $m = \sum_{i=1}^n y_i$. Let κ_i be the exposure for the ith observation. Then the probability of observing $(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$ is

$$\Pr(Y_1 = y_1, \ldots, Y_n = y_n \mid \boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\kappa}) = \frac{\exp(m\theta + \mathbf{t}\boldsymbol{\beta})}{\exp\{\sum_{i=1}^n \kappa_i \exp(\theta + \mathbf{x}_i\boldsymbol{\beta})\}} \prod_{i=1}^n \frac{\kappa_i^{y_i}}{y_i!}$$

where $\mathbf{t} = (t_1, \ldots, t_p)$, $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)^T$.

The joint distribution of the sufficient statistics $(\mathbf{T}, M)$ is obtained by summing over all possible sequences $Y_1 \geq 0, \ldots, Y_n \geq 0$ such that $\mathbf{T} = \mathbf{t}$ and $M = m$. This probability function is

$$\Pr(T_1 = t_1, \ldots, T_p = t_p, M = m \mid \boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\kappa}) = \frac{\exp(m\theta + \mathbf{t}\boldsymbol{\beta})}{\exp\{\sum_{i=1}^n \kappa_i \exp(\theta + \mathbf{x}_i\boldsymbol{\beta})\}} \left(\sum_{\mathbf{u}} \prod_{i=1}^n \frac{\kappa_i^{u_i}}{u_i!} \right)$$

where the sum $\sum_{\mathbf{u}}$ is over all nonnegative vectors $\mathbf{u}$ of length n such that $\sum_{i=1}^n u_i = m$ and $\sum_{i=1}^n u_i \mathbf{x}_i = \mathbf{t}$.

Conditional distribution

Without loss of generality, we will restrict our discussion to the conditional distribution of the sufficient statistic for β_1, T_1. If we condition on observing $M = m$ and $T_2 = t_2, \ldots, T_p = t_p$, the probability function of $(T_1 \mid \beta_1, T_2 = t_2, \ldots, T_p = t_p, M = m)$ is

$$\Pr(T_1 = t_1 \mid \beta_1, T_2 = t_2, \ldots, T_p = t_p, M = m) = \frac{\left(\sum_{\mathbf{u}} \prod_{i=1}^n \frac{\kappa_i^{u_i}}{u_i!} \right) e^{t_1\beta_1}}{\sum_{\mathbf{v}} \left(\prod_{i=1}^n \frac{\kappa_i^{v_i}}{v_i!} \right) e^{\beta_1 \sum_i v_i x_{i1}}} \tag{2}$$

where the sum $\sum_{\mathbf{u}}$ is over all nonnegative vectors $\mathbf{u}$ of length n such that $\sum_{i=1}^{n} u_i = m$ and $\sum_{i=1}^{n} u_i \mathbf{x}_i = \mathbf{t}$, and the sum $\sum_{\mathbf{v}}$ is over all nonnegative vectors $\mathbf{v}$ of length n such that $\sum_{i=1}^{n} v_i = m$, $\sum_{i=1}^{n} v_i x_{i2} = t_2, \ldots, \sum_{i=1}^{n} v_i x_{ip} = t_p$. The CMLE for β_1 is the value that maximizes the log of (1). This optimization task is carried out by `ml` (see [R] **ml**), using the conditional distribution of $(T_1 \mid T_2 = t_2, \ldots, T_p = t_p, M = m)$ as a dataset. This dataset consists of the feasible values and weights for T_1,

$$\left\{ \left(s_1, \prod_{i=1}^{n} \frac{\kappa_i^{v_i}}{v_i!} \right) : \sum_{i=1}^{n} v_i = m, \sum_{i=1}^{n} v_i x_{i1} = s_1, \sum_{i=1}^{n} v_i x_{i2} = t_2, \ldots, \sum_{i=1}^{n} v_i x_{ip} = t_p \right\}$$

Computing the CMLE, MUE, confidence intervals, conditional hypothesis tests, and sufficient statistic p-values is discussed in *Methods and formulas* of [R] **exlogistic**. The only difference between the two techniques is the use of the weights; that is, the weights for exact logistic are the combinatorial coefficients, $c(\mathbf{t}, m)$, in (1) of *Methods and formulas* in [R] **exlogistic**. `expoisson` and `exlogistic` use the same `ml` likelihood evaluator to compute the CMLEs as well as the same ado-programs and Mata functions to compute the MUEs and estimate statistics.

References

Agresti, A. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley.

Armitage, P., G. Berry, and J. N. S. Matthews. 2002. *Statistical Methods in Medical Research*. 4th ed. Oxford: Blackwell.

Cox, D. R., and E. J. Snell. 1989. *Analysis of Binary Data*. 2nd ed. London: Chapman and Hall.

Hirji, K. F., C. R. Mehta, and N. R. Patel. 1987. Computing distributions for exact logistic regression. *Journal of the American Statistical Association* 82: 1110–1117. https://doi.org/10.2307/2289388.

Laird, N. M., and D. Olivier. 1981. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* 76: 231–240. https://doi.org/10.2307/2287816.

Also see

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

[R] **poisson** — Poisson regression

[XT] **xtpoisson** — Fixed-effects, random-effects, and population-averaged Poisson models

[U] **20 Estimation and postestimation commands**