Poisson regression

In statistics, **Poisson regression** is a form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable *Y* has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

Poisson regression models are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function.

Regression models

If $x \in \mathbb{R}^n$ is a vector of independent variables, then the model takes the form

$$\log(\mathrm{E}(Y|x)) = a'x + b,$$

where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Sometimes this is written more compactly as

$$\log(\mathrm{E}(Y|x)) = \theta'x,$$

where *x* is now an *n+1*-dimensional vector consisting of *n* independent variables concatenated to some constant, usually 1. Here *θ* is simply *a* concatenated to *b*.

Thus, when given a Poisson regression model *θ* and an input vector x , the predicted mean of the associated Poisson distribution is given by

$$\mathrm{E}(Y|x) = e^{\theta'x}.$$

If Y_i are independent observations with corresponding values x_i of the predictor variable, then *θ* can be estimated by maximum likelihood. The maximum-likelihood estimates lack a closed-form expression and must be found by numerical methods. The probability surface for maximum-likelihood Poisson regression is always convex, making Newton-Raphson or other gradient-based methods appropriate estimation techniques.

Maximum likelihood-based parameter estimation

Given a set of parameters *θ* and an input vector *x*, the mean of the predicted Poisson distribution, as stated above, is given by

$$\mathrm{E}(Y|x) = e^{\theta'x},$$

and thus, the Poisson distribution's probability mass function is given by

$$p(y|x;\theta) = \frac{e^{y(\theta'x)}e^{-e^{\theta'x}}}{y!}$$

Now suppose we are given a data set consisting of *m* vectors $x_i \in \mathbb{R}^{n+1}$, $i = 1, ..., m$, along with a set of *m* values $y_1, ..., y_m \in \mathbb{R}$. Then, for a given set of parameters *θ*, the probability of attaining this particular set of data is given by

$$p(y_1, ..., y_m|x_1, ..., x_m;\theta) = \prod_{i=1}^{m} \frac{e^{y_i(\theta'x_i)}e^{-e^{\theta'x_i}}}{y_i!}.$$

By the method of maximum likelihood, we wish to find the set of parameters *θ* that makes this probability as large as possible. To do this, the equation is first rewritten as a likelihood function in terms of θ:

$$L(\theta|X,Y) = \prod_{i=1}^{m} \frac{e^{y_i(\theta'x_i)}e^{-e^{\theta'x_i}}}{y_i!}.$$

Note that the expression on the right hand side has not actually changed. A formula in this form is typically difficult to work with; instead, one uses the *log-likelihood*:

$$\ell(\theta|X,Y) = \log L(\theta|X,Y) = \sum_{i=1}^{m} \left(y_i(\theta' x_i) - e^{\theta' x_i} - \log(y_i!)\right).$$

Notice that the parameters θ only appear in the first two terms of each term in the summation. Therefore, given that we are only interested in finding the best value for θ we may drop the $y_i!$ and simply write

$$\ell(\theta|X,Y) = \sum_{i=1}^{m} \left(y_i(\theta' x_i) - e^{\theta' x_i}\right).$$

To find a maximum, we need to solve an equation $\frac{\partial \ell(\theta|X,Y)}{\partial \theta} = 0$which has no closed-form solution. However, the negative log-likelihood, $-\ell(\theta|X,Y)$, is a convex function, and so standard convex optimization techniques such as gradient descent can be applied to find the optimal value of θ.

Poisson regression in practice

Poisson regression is appropriate when the dependent variable is a count, for instance of events such as the arrival of a telephone call at a call centre. The events must be independent in the sense that the arrival of one call will not make another more or less likely, but the probability per unit time of events is understood to be related to covariates such as time of day.

"Exposure" and offset

Poisson regression is also appropriate for rate data, where the rate is a count of events occurring to a particular unit of observation, divided by some measure of that unit's *exposure*. For example, biologists may count the number of tree species in a forest, and the rate would be the number of species per square kilometre. Demographers may model death rates in geographic areas as the count of deaths divided by person–years. More generally, event rates can be calculated as events per unit time, which allows the observation window to vary for each unit. In these examples, exposure is respectively unit area, person–years and unit time. In Poisson regression this is handled as an **offset**, where the exposure variable enters on the right-hand side of the equation, but with a parameter estimate (for log(exposure)) constrained to 1.

$$\log\left(\mathrm{E}(Y|x)\right) = \log\left(\text{exposure}\right) + \theta' x$$

which implies

$$\log\left(\mathrm{E}(Y|x)\right) - \log\left(\text{exposure}\right) = \log\left(\frac{\mathrm{E}(Y|x)}{\text{exposure}}\right) = \theta' x$$

Offset in the case of a GLM in R can be achieved using the offset() function:

```
glm.fit <- glm(y ~ offset(log(exposure)) + x, family=poisson(link=log) )
```

Overdispersion

A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as overdispersion and indicates that the model is not appropriate. A common reason is the omission of relevant explanatory variables, or dependent observations. Under some circumstances, the problem of overdispersion can be solved by using a negative binomial distribution instead.[1][2]

Another common problem with Poisson regression is excess zeros: if there are two processes at work, one determining whether there are zero events or any events, and a Poisson process determining how many events there are, there will be more zeros than a Poisson regression would predict. An example would be the distribution of

cigarettes smoked in an hour by members of a group where some individuals are non-smokers.

Other generalized linear models such as the negative binomial model may function better in these cases.

Use in survival analysis

Poisson regression creates proportional hazards models, one class of survival analysis: see proportional hazards models for descriptions of Cox models.

Tests of over dispersion

One method for testing for over dispersion in the data is to regress a variable (z_i) against the predicted values of *t* estimated from the Poisson regression.[3] This test has three steps.

1. Estimate a poisson regression of y_i on x_i and generate the predicted values (t_i)

2. Calculate the z_i variable

$$z_i = \frac{(y_i - t_i)^2 - y_i}{t_i\sqrt{2}}$$

3. Regress z_i against t_i with ordinary least squares. In symbols

$$z_i = at_i + e_i$$

where *a* is a constant and e_i is a random variable with an expectation of zero.

The null hypothesis being tested here is that the data are Poission distributed: in this case $a = 0$.

Extensions

Regularized Poisson Regression

When estimating the parameters for Poisson regression, one typically tries to find values for θ that maximize the likelihood of an expression of the form

$$\sum_{i=1}^{m} \log(p(y_i; e^{\theta' x})),$$

where *m* is the number of examples in the data set, and $p(y_i; e^{\theta' x})$ is the probability mass function of the Poisson distribution with the mean set to $e^{\theta' x}$. Regularization can be added to this optimization problem by instead maximizing

$$\sum_{i=1}^{m} \log(p(y_i; e^{\theta' x})) - \lambda \|\theta\|_2^2,$$

for some positive constant λ . This technique, similar to ridge regression, can reduce overfitting.

Implementations

Some statistics packages include implementations of Poisson regression.

- MATLAB Statistics Toolbox: Poisson regression can be performed using the "glmfit" and "glmval" functions.[4]
- Microsoft Excel: Excel is not capable of doing Poisson regression by default. One of the Excel Add-ins for Poisson regression is XPost [5]
- R: The function for fitting a generalized linear model in R is glm(), and can be used for Poisson Regression
- SAS: Poisson regression in SAS is done by using GENMOD
- SPSS: In SPSS, Poisson regression is done by using the GENLIN command
- Stata: Stata has a procedure for Poisson regression named "poisson"
- mPlus: mPlus allows for Poisson regression using the command COUNT IS when specifying the data

References

- Cameron, A.C. and P.K. Trivedi (1998). *Regression analysis of count data,* Cambridge University Press. ISBN 0-521-63201-3
- Christensen, Ronald (1997). *Log-linear models and logistic regression*. Springer Texts in Statistics (Second ed.). New York: Springer-Verlag. pp. xvi+483. ISBN 0-387-98247-7. MR1633357.
- Hilbe, J.M. (2007). *Negative Binomial Regression*, Cambridge University Press. ISBN 978-0-521-85772-7

[1] Paternoster R, Brame R (1997). "Multiple routes to delinquency? A test of developmental and general theories of crime". *Criminology* **35**: 45–84.

[2] Berk R, MacDonald J (2008). "Overdispersion and Poisson regression" (http://www.crim.upenn.edu/faculty/papers/berk/regression.pdf). *Journal of Quantitative Criminology* **24**: 269–284. .

[3] https://files.nyu.edu/mrg217/public/count.pdf

[4] http://www.mathworks.com/help/toolbox/stats/glmfit.html

[5] http://www.indiana.edu/~jslsoc/files_research/xpost/xpost.pdf

Article Sources and Contributors

Poisson regression *Source*: http://en.wikipedia.org/w/index.php?oldid=517970130 *Contributors*: Aetheling, Awaterl, Baccyak4H, Benwing, BlueScreenD, BrendanH, Brusegadi, Den fjättrade ankan, DrMicro, Faridani, Farmanesh, Fernagom, Free Software Knight, Gak, Giftlite, Gpeilon, Headbomb, Jitse Niesen, Kelonii, Kiefer.Wolfowitz, Kodiologist, MarkSweep, Melcombe, Michael Hardy, Pstevens, Qwfp, Rjwilmsi, Skittleys, Talgalili, طاها, 26 anonymous edits

License