Poisson regression

Pamela Warner

Reader in Medical Statistics, Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

Correspondence to
Dr Pamela Warner, Centre for Population Health Sciences, University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK;
p.warner@ed.ac.uk

Received 13 May 2015
Accepted 13 May 2015

BACKGROUND

The Rashed *et al*.[1] article in this issue uses Poisson regression as a statistical tool. These notes are intended to provide some supplementary explanation of this method (see Box 1 for a glossary of terms used in this article).

WHAT IS POISSON REGRESSION?

When the response variable is a count of the number of occurrences of an event (such as a births, or prescriptions for contraception), Poisson regression is commonly used to model the association between that count of events and any potential explanatory variables.[2] Poisson regression is also suitable for analysis of rate data, where the rate is the count of events occurring per some unit of observation – for example, a specified time period (see also Warner[3]). In the Rashed *et al*. article, what is modelled is rate of prescriptions for contraception *within calendar years*.[1]

Box 1 Glossary of statistical terms used in this article*

Confidence interval (95%)	This defines a range of values within which we are 95% confident the true population value lies (e.g. for an estimated population mean, rate, difference in proportions).
Cross-classified/ Cross-tabulation	Subdividing individuals studied in terms of two (or more) categorical variables, jointly. If two variables then an R×C table of counts is produced, where R=number of rows and C=number of columns. Each cell of the table is termed a frequency count of individuals with that combination of row and column values.
Estimation (inference)	This involves use of a summary value calculated from the sample data (e.g. mean, rate) to make an inference about the true value for that entity (parameter) in the background population from which the sample was drawn. The single parameter is a 'point estimate', but it is good practice to provide also an interval estimate. See *Confidence interval.*
Explanatory variable	A feature potentially associated with outcome (e.g. calendar year).
Frequency count	See *Cross-classified.*
Hypothesis test/ Significance test	The process of testing aims to enable a binary decision to be made about the null hypothesis (NH) – reject NH or not. This decision is based on the significance probability (*p* value) obtained via the test.
Independent (events)	The occurrence of any one event does not influence the occurrence of other events.
Multivariate modelling	A method of analysis that allows assessment of the association between an outcome and some explanatory variable(s) of interest (or here, event rate). The analysis is comparing rates between risk factor subgroups, and these comparisons are expressed as rate ratios.
Null hypothesis (NH)	A statement, prior to testing, of no effect (e.g. 'no association' between row and column classifications). See also *Significance probability.*
Odds (of an outcome)	The number of occurrences of that outcome, divided by non-occurrences (i.e. cases÷non-cases).
Poisson distribution	A discrete probability distribution for the number of events that occur.
p value	See *Significance probability.*
Rate	A rate summarises a quantity in relation to another unit of measurement, usually *but not always* in relation to time. For example, births per year is a rate per time unit, whereas perinatal mortality expresses neonatal deaths per live births.
Response variable	The outcome variable to be modelled/tested.
Significance probability (*p* value)	The probability, if the NH is true, of obtaining the observed data or something more 'extreme' (i.e. further from the NH). The smaller the *p* value is, then the less likely this data would be under the NH, and so the greater our doubts that the NH is indeed true.
Valid method of analysis	Used here loosely to mean a method that is suited to the research question and data variable(s) to be analysed, the latter including that the data to be analysed satisfy the data assumptions required for the method, so that results will not be misleading, on that account.

*Definitions of other terms used can be found in previous Noteworthy Statistics articles.[3–5]

To cite: Warner P. *J Fam Plann Reprod Health Care* 2015;**41**:223–224.

WHEN/WHY IS IT USEFUL?

This multivariate method has a range of uses, but the main one is comparison of rates between circumstances (e.g. historical time, i.e. years 2002 to 2011)[1] or between specific subgroup characteristics (e.g. younger and older teenagers).[1] It can also be used to control for confounding, and to provide population estimates [with confidence intervals (CIs)].

There are many similarities in approach between logistic regression[4] and Poisson regression. The key difference between the methods is that logistic regression focuses on a binary outcome (e.g. being a 'case' or not, such as positive test result for chlamydia), and models this *odds* of being a case, within subgroups cross-classified in terms of potential explanatory variables (e.g. use of condoms, etc.), whereas Poisson regression focuses on how many cases there are (per cross-classification, and perhaps per unit of measurement, e.g. calendar year). Poisson regression is particularly useful for fairly rare events.

WHAT PRECAUTIONS ARE NEEDED?

In order to be valid, Poisson regression requires that the count variable has a Poisson distribution, and that the occurrence of events is independent. However, many research variables exhibit more dispersion than they should, if truly following a Poisson distribution. The impact of this on the results of analysis would be (slightly) misleading CIs, narrower than they truly should be to provide the stated 'confidence' in the estimate. Dispersion therefore needs to be checked as a preliminary to Poisson regression analysis. Other pitfalls with Poisson regression are the same as those encountered with other methods of multivariate modelling, including logistic regression.[4 5]

EXAMPLE

Figure 1 in the Rashed *et al*. article shows the estimates (with CIs) that were obtained for rates of prescription for contraception by calendar year for the two teenage subgroups.[1] The model has also been used to compare prescription rates between the two teenage subgroups, combined across calendar years. Not surprisingly, given the graph, the null hypothesis of no difference between prescription rates for older and younger teenagers can be rejected ($p<0.001$, Results paragraph 1). In Figure 1 the increase in prescribing across years appears to have been modest,[1] and no p value is reported, so presumably there was no statistically significant difference across calendar years.

OVERVIEW

Poisson regression is a useful method for analysis of event rates and comparison across levels of explanatory variables.

Competing interests None declared.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1 Rashed AN, Hsia Y, Wilton L, *et al*. Trends and patterns of hormonal contraception prescribing for adolescents in primary care in the UK. *J Fam Plann Reprod Health Care* 2015:41:216–222.
2 Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. Oxford, UK: Blackwell Science, 2003.
3 Warner P. Rate and rate ratio. *J Fam Plann Reprod Health Care* 2009;35:111–113.
4 Warner P. Testing and quantifying association in binary data. *J Fam Plann Reprod Health Care* 2009;35:26–27.
5 Warner P. Modelling with multiple explanatory variables. *J Fam Plann Reprod Health Care* 2011;37:32–34.