Ophthalmic statistics note 11: logistic regression

John Stephenson,[1] Catey Bunce,[2,3] Caroline J Doré,[4] Nick Freemantle,[5] On behalf of the Ophthalmic Statistics Group

LOGISTIC REGRESSION

Previous notes in this series have been concerned with the common situation in ophthalmic and other clinical fields of describing relationships between one or more 'predictors' (explanatory variables) and, usually, one outcome measure (response variable). A classic method used in deriving relationships between outcomes and predictors is linear regression analysis. Linear regression is a member of a family of techniques known as general linear models, which also include analysis of variance and analysis of covariance; the latter of which was covered in a previous Ophthalmic Statistics Note.[1]

A key feature of all these models is that the outcome measure—for example, postoperative refractive prediction error or intraocular pressure—is continuous. While other notes in the series[2] warn of the dangers of unnecessary dichotomisation of variables, sometimes outcomes naturally fall into two categories.

Example 1: A study was conducted on 137 patients to identify risk factors for intraoperative retinal breaks caused by induction of a posterior hyaloid face separation during 23-gauge pars plana vitrectomy.[3] Putative risk factors for breaks were age at surgery, axial length of the operated eye and diagnosis, but the outcome variable here was whether or not the patient suffered a retinal break—a yes/no or dichotomous outcome.

Example 2: A study was conducted on 58 patients undergoing surgery for idiopathic macular hole identifying whether or not a patient develops an outer foveal defect (OFD).[4] Putative risk factors were age at surgery, characteristics of the macular hole such as base diameter and whether or not there was ocular comorbidity, but the outcome was whether or not the patient developed an OFD in their operated eye—a yes/no or dichotomous outcome.

In both examples, our objective is to examine relationships between a single outcome variable and several predictors. Typically, when faced with this challenge, we would use linear regression. Linear regression, however, requires a continuous outcome and thus if we were to use this method we would be violating a statistical assumption. In our last statistical note, we introduced the concept of transforming data in order to conduct valid statistical analyses. Focus in that note was on transformations of the explanatory or independent variables. It is, however, also possible to conduct transformations on outcomes so that while the outcome itself is not continuous, a transformation based upon that outcome is. We can then apply regression in the same manner we are accustomed to and identify associations between outcomes and risk factors, acknowledging that our associations actually relate to the transformation. As was the case in our previous note, the challenge, therefore, is in the interpretation of results after application of the transformation.

The transformation that we use to achieve this is called logistic regression. We assign our outcome variable numerical values of 1 and 0, representing yes and no, respectively. If we had 10 subjects and 5 had breaks and 5 did not, we would say intuitively that the probability of an event (p) was 5/10—the proportion of our group who had the event of interest. In logistic regression, our outcome of interest is based on this probability. However, probabilities are bounded by 0 and 1, where 0 indicates impossible and 1 indicates certainty. It, just like our original outcome, is not therefore normally distributed. A transformation of probability, known as the logit transformation, is not, however, constrained by bounds of 0 and 1 and logistic regression may then be used to explore associations between the covariates of interest and our logit transformation, where

$$\text{logit}\, p = \ln \frac{p}{1-p}$$

While this transformation may appear unintuitive, it should be noted that the quantity $\frac{p}{1-p}$ on the right-hand side of this equation is known as the *odds*. Odds will be familiar to those who attend horse racing—it is the probability that the event occurs divided by the probability that the event does not occur. This quantity will be familiar to gamblers who are used to seeing horses quoted as having, say, odds of 5 to 1 of winning a race. This does not mean that the probability of winning is 1 in 5, but rather that the horse has 1 'winning chance' and 5 'losing chances'; hence, a winning probability of 1 in 6.

Logistic regression was used in a study[5] to see whether macular hole inner opening was predictive of anatomical success of surgery to repair the hole. The regression equation for this model was

$$\text{logit}\, p = 10.89 - 0.016 \times \text{macular hole inner opening (in } \mu\text{m)} \quad (1)$$

The estimated probability of anatomical success can then be calculated, so that for a patient with a macular hole inner opening of 650 μm, the logit of p is given by

$$10.89 - 0.016 \times 650 = 0.49$$

Logits have no direct interpretation, and so to interpret this equation in a useful predictive sense, we need to 'undo' the logistic transformation. This can be achieved in two steps. First, the odds of the event are calculated by exponentiating or 'antilogging' the regression function:

$$\frac{p}{1-p} = \text{odds}(p) = \exp(0.49) = 1.63$$

Next, a bit of simple algebra is used to convert these odds to a probability:

$$p = \frac{\text{odds}(p)}{1 + \text{odds}(p)} = 0.62$$

So, preoperatively, our patient is predicted to have a 62% chance of anatomical success. This procedure (exponentiation and algebra) would not normally be the responsibility of the researcher: most statistical packages will routinely perform these transformations as part of their logistic regression function. In fact, unlike simple linear regression, in which parameters may be estimated using the least-squares method, it is not generally

[1]School of Human and Health Sciences, University of Huddersfield, Queensgate, Huddersfield, UK; [2]NHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK; [3]Reader in Medical Statistics, Department of Primary Care & Public Health Sciences, King's College London, 4th Floor, Addison House, Guy's Campus, London, SE1 1UL; [4]Comprehensive Clinical Trials Unit, University College London, London, UK; [5]Medical School, University College London, London, UK

Correspondence to Dr John Stephenson, School of Human and Health Sciences, University of Huddersfield, Queensgate, Huddersfield GB-HD1 3DH, UK; J.Stephenson@hud.ac.uk

Table 1 Computer output from macular hole study (edited)

	B	SE	Wald	df	p Value	OR	95% CI for OR Lower	95% CI for OR Upper
Macular hole inner opening	−1.637	0.539	9.214	1	0.002	0.195	0.068	0.560
Constant	10.890	3.293	10.938	1	0.001	53647.735		

practical to conduct logistic regression, in which parameters are generally estimated using other means, by hand: computer software is usually required.

Assessing the effect of a covariate also requires us to undo the logistic transformation. The computer output (slightly edited) summarising the model above (table 1) includes the ORs associated with the model parameters (some software will label these columns as 'Exp(B)': the exponent of the parameter estimate in eq. (1) above). These represent the ratio of two odds: the odds of the baseline event and the odds of the event associated with a unit increase in the predictor variable (defined to be a 100 µm increase in macular hole inner opening in this case). If the ratio is significantly different from 1 (ie, if the associated CI does not include 1), then the variable is associated with the outcome: either positively if the OR is greater than 1 or negatively if the OR is less than 1. As such, the OR is a generally more meaningful quantity than the parameter estimate (typically labelled B as in this table) from which it was derived. We do not need the columns in the table headed 'SE' , 'Wald' or 'df' (degrees of freedom) to interpret the OR.

The OR for a particular parameter is not the same as the risk ratio (relative risk), although for rare events it is a reasonable approximation. Although it is not as intuitive as the risk ratio, it possesses certain advantages; for instance, it is not constrained by large baseline risks. The relationship between odds and risk ratios, and other quantities such as prevalence and exposure rates, may be found in many standard texts, for example.[6]

The estimation of the OR may be considered to be the back-transformation of the results into the original data units. In this example, we see that an increase of 100 µm in macular hole inner opening leads to a significant reduction (p=0.002) in odds of anatomical success of 80.5% (calculated by multiplying 1–0.195 by 100). The associated CI for the OR (0.068 to 0.560) confirms that this reduction is statistically significant as it excludes the value 1.00, which corresponds to no effect. We can ignore the line of the output for the constant: these statistics have little practical value.

Lessons learnt

- Mathematical functions (transformations) may be applied to outcome (explanatory) variables.
- Studies exploring relationships between one or several predictor variables and a dichotomous outcome typically make use of one such transformation the logit in a technique known as logistic regression.
- Logistic regression typically yields ORs with 95% CIs. An OR of 1 corresponds to no association with the predictor variable and so a CI excluding 1 is evidence of association.

Contributors JS drafted the paper. CB, CJD and NF critically reviewed and revised the paper. JS and CB redrafted the paper after review. JS, CB and CJD critically reviewed the redraft.

Funding CB is partly funded by the National Institute of Health Research (NIHR) Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

To cite Stephenson J, Bunce C, Doré CJ, *et al*. *Br J Ophthalmol* 2016;**100**:1594–1595.

Published Online First 3 November 2016

▶ http://dx.doi.org/10.1136/bjophthalmol-2016-308824

Br J Ophthalmol 2016;**100**:1594–1595. doi:10.1136/bjophthalmol-2016-309223

REFERENCES

1 Nash R, Bunce C, Freemantle N, *et al*. Ophthalmic Statistics Note 4: analysing data from randomised controlled trials with baseline and follow-up measurements. *Br J Ophthalmol* 2014;98:1467–9.
2 Cumberland PM, Czanner G, Bunce C, *et al*. Ophthalmic Statistics Note 3: the perils of dichotomising continuous variables. *Br J Ophthalmol* 2014;98:841–3.
3 Rahman R, Murray CD, Stephenson J. Risk factors for iatrogenic retinal breaks induced by separation of posterior hyaloid face during 23-gauge pars plana vitrectomy. *Eye* 2013;27:652–6.
4 Rahman R, Oxley L, Stephenson J. Persistent outer retinal fluid following non-posturing surgery for idiopathic macular hole. *Br J Ophthalmol* 2013;97:1451–4.
5 Wakely L, Rahman R, Stephenson J. A comparison of several methods of macular hole measurement using optical coherence tomography, and their value in predicting anatomical and visual outcomes. *Br J Ophthalmol* 2012;96:1003–7.
6 Kirkwood BR, Sterne JAC. *Essential medical statistics*. 2nd edn. Oxford: Blackwell Science, 2003.