BMJ

ENDGAMES

STATISTICAL QUESTION

Logistic regression

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, Tooting, London, UK

Researchers investigated whether reoperation after breast conserving surgery was associated with patients' characteristics. A cohort study design was used. Participants were 55 297 women who had primary breast conserving surgery in 156 English NHS trusts between 1 April 2005 and 31 March 2008. The primary outcome was at least one breast reoperation within three months of breast conserving surgery. During the three year study period, 11 032 (20.0%, 95% confidence interval 19.6% to 20.3%) women had at least one reoperation.[1]

Logistic regression analysis examined the association between breast reoperation and patients' characteristics (age at admission, tumour group, comorbidity, and socioeconomic deprivation) (table⇓). Tumour type was grouped by whether a carcinoma in situ component was recorded at the time of primary breast conserving surgery. The researchers commented that reoperation was nearly twice as likely when the tumour had a carcinoma in situ component recorded.

Which of the following statements, if any, are true?

a) The outcome variable for logistic regression is continuous

b) The odds ratio of breast reoperation for categorised age 50-59 years was 1.0

c) It can be concluded that the type of tumour was independently associated with breast reoperation

d) Conditional logistic regression was used to obtain the adjusted odds ratios

Answers

Statement *c* is true, whereas *a*, *b,* and *d* are false.

The aim of the study was to investigate whether the need for reoperation within three months of breast conserving surgery could be predicted from the patients' characteristics. A logistic regression model was used. Participants were 55 297 women who had breast conserving surgery during a three year period, of whom 11 032 (20.0%) had at least one reoperation.

Logistic regression is similar to other regression methods described in previous questions.[2 3] Referred to as multivariable analysis, logistic regression investigates the association between a dependent variable and one or more predictor variables simultaneously. The outcome variable is binary (*a* is false), in contrast to simple linear regression and multiple regression analyses where the outcome is continuous. In the example above, the outcome variable was reoperation (yes or no) within three months of breast conserving surgery. The predictor variables are sometimes referred to as explanatory variables and can be any mixture of continuous, binary, or categorical variables. In the example above, the explanatory variables were all categorical and included type of tumour (with and without in situ disease recorded), categorised age, socioeconomic deprivation index, and categorised comorbidities. It was assumed that the observations were independent of each other—that is, each woman had only one observation of the dependent and explanatory variables.

Logistic regression estimates the probability of the outcome occurring given values of the predictor values. Typically, the results of a logistic regression are presented as odds ratios. Odds ratios have been described in a previous question.[4] For the example above, the unadjusted and adjusted odds ratios describing the association between the outcome variable of breast reoperation and each category of the explanatory variables are presented (table). The unadjusted odds ratios, sometimes referred to as crude odds ratios, have not been adjusted for potential confounding by the other predictor variables. The adjusted odds ratios have been adjusted for potential confounding by the other explanatory variables in the analysis; in effect they represent the association between the outcome and explanatory variable when all other explanatory variables are constant.

The number 1 against age 50-59 years in the columns headed unadjusted and adjusted odds ratio does not represent an odds ratio (*b* is false). Each explanatory variable has a reference category, as indicated by the number 1 in the odds ratio columns. Sometimes the reference category is indicated by (1) instead. The other categories of the variable are compared against the reference category to derive the odds ratios. The odds ratio for a particular category is the odds of reoperation for that category divided by the odds of reoperation for the reference category. For example, the unadjusted odds ratio for the category less than 40 years for the explanatory variable age is 1.07. Therefore, the unadjusted odds of breast reoperation for women aged under

p.sedgwick@sgul.ac.uk

40 years were 1.07 times those of the reference category—that is, women aged 50-59 years.

Comparison of the unadjusted and adjusted odds ratios indicates the extent of confounding. After adjustment for confounding, a non-significant odds ratio become significant for the age category of less than 40 years and socioeconomic deprivation index 4, whereas a significant odds ratio becomes non-significant for socioeconomic deprivation index 5 and the one comorbidity category. Confounding was therefore suggested in the association between these categories of the explanatory variables and the outcome variable of breast reoperation. However, the difference in size between the unadjusted and adjusted odds was small, which suggests that the extent of confounding was minimal.

For each odds ratio the 95% confidence interval for the population odds ratio is presented, providing an interval estimate for the population parameter. A previous question explained that, if a 95% confidence interval for a population odds ratio excluded unity, the test of the statistical null hypotheses of no difference in odds between the categories of the explanatory variable will be rejected in favour of the alternative at the 5% level.[5]

If the association between an explanatory variable and the outcome is significant after adjusting for confounding, then the explanatory variable is said to be independently associated with the outcome. The association of type of tumour with breast reoperation was statistically significant after adjusting for the other explanatory variables, the 95% confidence interval for the population odds ratio not straddling unity. Therefore, type of tumour was considered to be independently associated with breast reoperation (*c* is true).

The above analysis was not a conditional logistic regression (*d* is false). A conditional logistic regression is one where the participants are matched, as in a matched case-control study. In such a study design, for each case (study participant who experienced the primary outcome), one or more controls (those that did not experience the primary outcome) matched on a series of variables such as age and sex are identified. The potential confounding effects of the matching variables can be controlled for more efficiently by matching cases and controls at the design stage of the study, rather than in the subsequent statistical analysis.[6]

Competing interests: None declared.

1 Jeevan R, Cromwell DA, Trivella M, Lawrence G, Kearins O, Pereira J, et al. Reoperation rates after breast conserving surgery for breast cancer among women in England: retrospective study of hospital episode statistics. *BMJ* 2012;345:e4505.
2 Sedgwick P. Simple linear regression. *BMJ* 2013;346:f2340.
3 Sedgwick P. Multiple regression. *BMJ* 2013;347:f4373.
4 Sedgwick P, Marston L. Odds ratios. *BMJ* 2010;341:c4414.
5 Sedgwick P. Confidence intervals and statistical significance: rules of thumb. *BMJ* 2012;345:e4960.
6 Sedgwick P. Why match in case-control studies. *BMJ* 2012;344:e691.

Cite this as: *BMJ* 2013;347:f4488

Table

Table 1| **Multivariable logistic regression analysis of patient characteristics on reoperation in women after breast conserving surgery**

Characteristic	Unadjusted odds ratio (95% CI)	Adjusted odds ratio (95% CI)
Type of tumour		
Without in situ disease recorded	1	1
With in situ disease recorded	1.91 (1.82 to 2.01)	1.91 (1.81 to 2.01)
Age (years)		
<40	1.07 (0.97 to 1.19)	1.15 (1.04 to 1.28)
40-49	1.24 (1.16 to 1.32)	1.30 (1.22 to 1.38)
50-59	1	1
60-69	0.87 (0.83 to 0.92)	0.88 (0.84 to 0.93)
≥70	0.63 (0.59 to 0.67)	0.67 (0.63 to 0.72)
Index of multiple deprivation		
1 (least deprived)	1	1
2	0.97 (0.91 to 1.03)	0.96 (0.90 to 1.02)
3	0.93 (0.88 to 0.99)	0.93 (0.88 to 0.99)
4	0.93 (0.87 to 1.00)	0.93 (0.87 to 0.99)
5 (most deprived)	0.93 (0.86 to 0.99)	0.94 (0.87 to 1.00)
Number of comorbidities		
0	1	1
1	0.88 (0.82 to 0.95)	0.96 (0.89 to 1.03)
≥2	0.61 (0.49 to 0.78)	0.73 (0.57 to 0.92)