

Confidence Intervals for Adjusted Proportions Using Logistic Regression

Noodchanath Kongchouy (Corresponding author)

Department of Mathematics, Faculty of Science and Technology, Prince of Songkla University

Hat Yai, Songkhla, 90112, Thailand

Tel: 66-74-288-8685 E-mail: nootchanath.k@psu.ac.th

Uraiwan Sampantarak

Pattani Inland Fisheries Research and Development Center, Department of Fisheries

Ministry of Agricultural and Cooperatives, Pattani 94160, Thailand

Tel: 66-84-653-9391 E-mail: uraiwans@fisheries.go.th; uraiwan111@hotmail.com

This study was partially funded by the Graduate School, Prince of Songkla University.

Abstract

This paper presents confidence intervals for adjusted proportions using logistic regression with weighted sum contrasts. The methods are applied to data from two studies, (1) imposex percentages among female gastropods at different locations in the Gulf of Thailand adjusted for different species, and (2) complication-based neonatal morbidity risk for births at a major hospital adjusted for demographic factors.

Keywords: Confidence interval, Proportion, Logistic regression, Sum contrasts

1. Introduction

Odds ratios are conventionally used for assessing associations between binary outcomes and categorical risk factors. They are often preferred in scientific studies because they give valid confidence intervals for these associations for case-control studies as well as for cohort studies and cross-sectional studies (Fernandez et al., 1999; Lim & Tongkumchum, 2009; Peters et al., 2000). A further advantage is that methods such as Mantel-Haenszel adjustment and logistic regression are available for adjusting odds ratios for confounding bias arising from covariates associated with both the binary outcome and the risk factor of interest. These issues are discussed in detail in the biostatistical literature (see, for example, McNeil, 1996; Woodward, 1999).

For cohort and cross-sectional studies where the proportions or percentages of an adverse outcome are of primary interest, it is also important to give confidence intervals for comparing these proportions. If there are no covariates of interest, these confidence intervals can be computed, either directly from the observed proportions with some adjustment such as an arcsine transformation to ensure that the confidence intervals are between 0 and 1 (see, for example, Armitage & Berry, 1994) using a logistic regression model. However, the situation is more complicated when adjustments for covariates are required, and methods for constructing such confidence intervals are not routinely provided in statistical packages.

In this paper we describe a general method for computing confidence intervals for comparing several proportions after adjusting for categorical covariates. The method is illustrated using two recently published applications of scientific interest: imposex among female gastropods in the Gulf of Thailand (Swennen, Sampantarak, & Rattanadaku, 2009), and complication-based neonatal morbidity risk among babies at a major hospital in southern Thailand (Rachatapantanakorn & Tongkumchum, 2009).

2. Methods

2.1 Logistic Regression

Logistic regression (Hosmer & Lemeshow, 2000; Kleinbaum & Klein, 2002) is a statistical method widely used to model the association between a binary outcome probability - the probability of a specific outcome - and a set of fixed determinants. When the determinants are categorical factors, these factors can be structured as a multi-way contingency table of counts and the data for analysis comprise the proportions of adverse outcomes in

the cells of this table. If, for example, the outcome variable Y takes values 0 and 1 (adverse outcome) and there are two factors with levels indexed by i and j , respectively, the model takes the form

$$\text{Prob}[Y_{ij} = 1] = 1 / (1 + e^{-(\alpha_i + \beta_j)}). \tag{1}$$

Note that if r and c are the numbers of factor levels specified in the model by the sets of parameters $\{\alpha\}$ and $\{\beta\}$, the number of independent parameters is $r + c - 1$, so it is necessary to put a constraint on these parameters when fitting the model. This constraint is conventionally achieved by replacing $\alpha_i + \beta_j$ in the model by $\mu + \alpha_i + \beta_j$ where $\alpha_1 = 0$ and $\beta_1 = 0$, in which case the model is written as

$$\text{Prob}[Y_{ij} = 1] = 1 / (1 + e^{-(\mu + \alpha_i + \beta_j)}). \tag{2}$$

This model provides estimates of odds ratios for comparing the outcome probabilities with respect to specified levels of each factor. Thus $\exp(\alpha_i)$ is the ratio of the odds of an adverse outcome for level i to that for level 1 for the first factor, whereas $\exp(\beta_j)$ is the ratio of the odds of an adverse outcome for level j to that for level 1 for the second factor. Thus each odds ratio in this model uses the first level of each factor as a baseline. To obtain odds ratios with respect to another baseline level, the data would need to be recoded so that the new baseline is constrained to take the value 0.

Under appropriate conditions on the pattern of zeros in the data (see, for example, Section 7.2 of Venables & Ripley, 2002), this logistic regression model is fitted using maximum likelihood and the results include estimates of the parameters and their standard errors, from which confidence intervals can be plotted.

2.2 Contrasts

When the constraints $\alpha_1 = 0$ and $\beta_1 = 0$ are used in the logistic regression model (2) the confidence intervals apply to the *differences* between each of the sets of parameters and the first parameter specified in each factor. These differences are known as *treatment contrasts*. In practice, it is often preferable not to single out a specific level of a factor as a basis for comparison, but rather to treat all factor levels in the same way. For linear models of the form

$$Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}, \tag{3}$$

standard errors of these differences are obtained by using the standard sum contrasts available in commonly used software packages such as R program (R Development Core Team, 2007). However, as pointed out by Venables and Ripley (2002), these contrasts are not valid for unbalanced designs, which include logistic regression models. Thus it is necessary to construct specific contrasts for logistic regression, and this can be accomplished by using weighted sum contrasts rather than treatment contrasts (Tongkumchum & McNeil, 2009). These weighted sum contrasts provide standard errors for the differences between each factor level and their overall mean.

2.3 Adjustment for Covariates

The method is analogous to that commonly used by linear regression analysis for adjusting outcomes to reduce the effects of covariate factors, such as seasonal adjustment of unemployment rates. In this case, if $\{\beta\}$ represents the primary factor of interest and $\{\alpha\}$ the covariate factor, and ε_{ij} are independent errors with mean zero and common standard deviation, the model is given by Equation (3). The factors $\{\alpha\}$ and $\{\beta\}$ in this model are estimated as the row and column means of the data matrix y_{ij} , with a suitable constraint to ensure identifiability. To adjust for the covariate, the adjusted mean for level j of the primary factor is obtained by first removing the effect of the covariate from each observation by replacing y_{ij} by $y_{ij} - \hat{\alpha}_i$ and then adding a constant to ensure that the mean of the corrected observations remains the same as the mean of the original observations. As a result, the adjusted mean is $\bar{y}_j = \hat{\beta}_j + d$, where d is a constant chosen to ensure that the overall mean before and after the adjustment remains the same. It follows that $d = \bar{y} - \bar{\beta}$, where $\bar{\beta}$ is the mean of the estimated β parameters.

Similarly, the formula for adjusting the proportion for level j of a primary factor is

$$p_j^* = 1 / (1 + e^{-(\beta_j + d)}). \tag{4}$$

Note that this result follows from the fact that the estimate given by logistic regression for the adjusted odds ratio for level j compared to level 1 is $\exp(\hat{\beta}_j)$ and this must equate to $\left(\frac{p_j^*}{1 - p_j^*} \right) / \left(\frac{p_1^*}{1 - p_1^*} \right)$.

The constant d may be chosen to ensure that the overall proportion (or the total number N) of adverse outcomes

before and after the adjustment is the same, that is

$$\sum_{i=1}^r \sum_{j=1}^c n_{ij} p_j^* = \sum_{i=1}^r \sum_{j=1}^c n_{ij} p_{ij} = N. \tag{5}$$

Substituting Equation (4) into Equation (5), it follows that d must satisfy the equation

$$\sum_{j=1}^c \left(\sum_{i=1}^r n_{ij} \right) / \left(1 + e^{-(\beta_j + d)} \right) = N. \tag{6}$$

Equation (6) is non-linear and cannot be solved explicitly to give an expression for the constant d . However, it can be solved using the Newton-Raphson iterative procedure with Marquardt damping to ensure convergence. Note that this method extends straightforwardly to any number of covariate factors.

3. Applications

3.1 *Imposex among gastropods in the Gulf of Thailand*

Female gastropods were collected from 56 sampling sites grouped into 13 areas around the Gulf of Thailand in 2006 (Swennen et al., 2009) and tested for imposex. Since different species have different sensitivities to imposex and this variation could bias the estimation of imposex prevalence due to the fact that different species are found in different locations, it was necessary to take both factors (16 species groups and 13 areas) into account. The logistic regression model (Equation 2) was fitted to the data. Figure 1 shows plots of 95% confidence intervals for the proportions.

The overall imposex percentage was 25.2, indicated by the vertical lines in Figure 1. Although there were 208 (13 areas \times 16 species groups) cells in the data table, 124 of these cells contained no data because relatively few species groups were found in each area. Thus the number of degrees of freedom after fitting the two-factor model was reduced from 71 (84 – 13) to 56 (84 – 13 – 15).

Figure 1 indicates that the crude percentages overestimated the true values in the Pattaya area and underestimated them in the Tak Bai, Pattani and Namrin areas. Due to the fact that 15 additional parameters were needed to take the species factor into account, the adjusted percentages have wider confidence intervals than the crude percentages. This effect is particularly notable for the Tak Bai area because the imposex was only found to occur among one of the three species groups observed there.

3.2 *Complication-based neonatal morbidity risk*

Based on complications recorded for 19,268 singleton deliveries to mothers with no previous Caesarean-section birth at Pattani Hospital in Southern Thailand over a nine-year period from 1997 to 2005, Rachatapantanakorn and Tongkumchum (2009) classified babies as high or low risk, and used logistic regression to assess the effects of six demographic risk factors.

Figure 2 shows 95% confidence intervals of the crude and adjusted percentages for each determinant before and after adjusting for the other determinants. The main differences between the crude and adjusted risks occurred among three demographic factors: the number of pregnancies (gravid group), age-group and religion.

Mothers in gravid group 4 or more pregnancies had highest risk percentage before adjustment, but after adjusting for the other determinants the risk was highest for primigravid mothers. For the age-group effect, the crude percentage was lowest for mothers aged 25-29 years, whereas the adjusted percentage was lowest for mothers aged less than 20 years. It is also noteworthy that the quite large difference in risks between Muslim and non-Muslim mothers was reduced after adjusting for the other factors.

4. Discussion

In this paper we have described a simple method for adjusting proportions for categorical covariates based on a fitted logistic regression model that provides asymptotically valid confidence intervals for comparing proportions over different levels of a categorical risk factor. While this method is not entirely new (see, for example, related earlier work by Berthold et al., 2007; Graubard & Korn, 1999; Lane & Nelder, 1982), it is not widely used in scientific studies, particularly when comparing more than two proportions. A further advantage of the method is that by using appropriately weighted sum contrasts each proportion can be compared with the overall mean rather than with a specified reference group.

5. Acknowledgement

We are grateful to Professor Don McNeil for his helpful guidance.

References

- Armitage, P., & Berry, G. (1994). *Statistical methods in medical research*. (3rd ed.). UK: Blackwell-science, Oxford.
- Berthold, S. M., Wong, E. C., Schell, T. L., Marshall, G. N., Elliott, M. N., Takeuchi, D., & Hambarsoomians, K. (2007). U.S. Cambodian refugees' use of complementary and alternative medicine for mental health problems. *Psychiatric Services*, 58, 1212-1218.
- Fernandez, E., Schiaffino, A., Rajmil, L., Badia, X., & Segura, A. (1999). Gender inequalities in health and health care services use in Catalonia (Spain). *Journal of Epidemiology and Community Health*, 53(4), 218-222.
- Graubard, B. I., & Korn, E. L. (1999). Predictive margins with survey data. *Biometrics*, 55, 652-659.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. (2nd ed.). John Wiley and Sons.
- Kleinbaum, D. G., & Klein, M. (2002). *Logistic Regression: a Self-Learning Text*. (2nd ed.). New York: Springer-Verlag.
- Lane, P. W., & Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, 38(3), 613-621.
- Lim, A., & Tongkumchum, P. (2009). Methods for analyzing hospital length of stay with application to inpatients dying in Southern Thailand. *Global Journal of Health Science*, 1(1), 27-38.
- McNeil, D. (1996). *Epidemiological research methods*. Chichester: John Wiley & Sons.
- Peters, A., Liu, E., Verrier, R. L., Schwartz, J., Gold, D. R., Mittleman, M., Baliff, J., Oh, A. J., Allen, G., Monahan, K., & Dockery, D. W. (2000). Air pollution and incidence of cardiac arrhythmia. *Epidemiology*, 11, 11-17.
- R Development Core Team. (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>.
- Rachatapantanakorn, O., & Tongkumchum, P. (2009). Demographic determinants for caesarean delivery in Pattani hospital. *Southeast Asian Journal of Tropical Medicine and Public Health*, 40, 602-611.
- Swennen, C., Sampantarak, U., & Rattanadakul, N. (2009). TBT-pollution in the Gulf of Thailand: a re-inspection of imposex incidence after 10 years. *Marine Pollution Bulletin*, 58, 526-532.
- Tongkumchum, P., & McNeil, D. (2009). Confidence intervals using contrasts for regression model. *Songklanakarin Journal of Science and Technology*, 31, 151-156.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. (4th ed.). New York: Springer Science + Business Media, (Section 6.2).
- Woodward, M. (1999). *Epidemiology: study design and data analysis*. USA: Chapman & Hall.

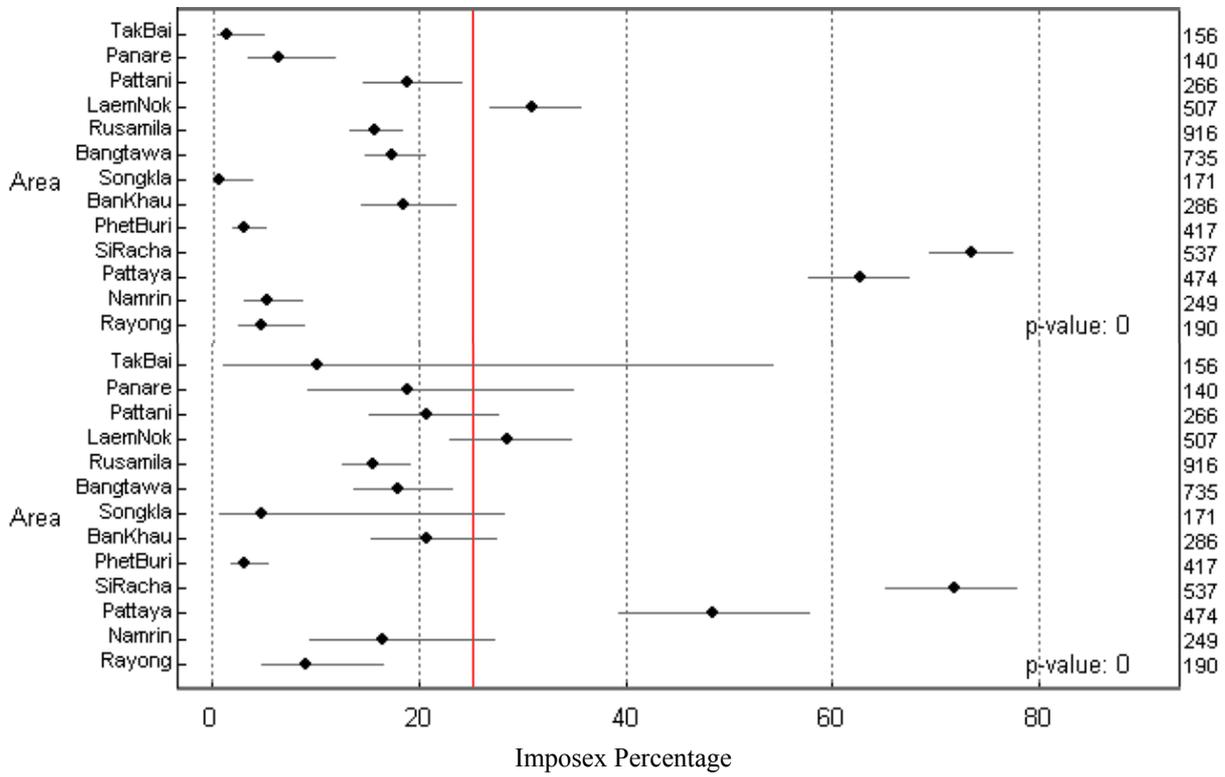


Figure 1. The 95% confidence intervals for percentages of female gastropods with imposex disease at various locations around the Gulf of Thailand; the upper and lower panels show crude and species-adjusted percentages, and the sample sizes are given on the right.

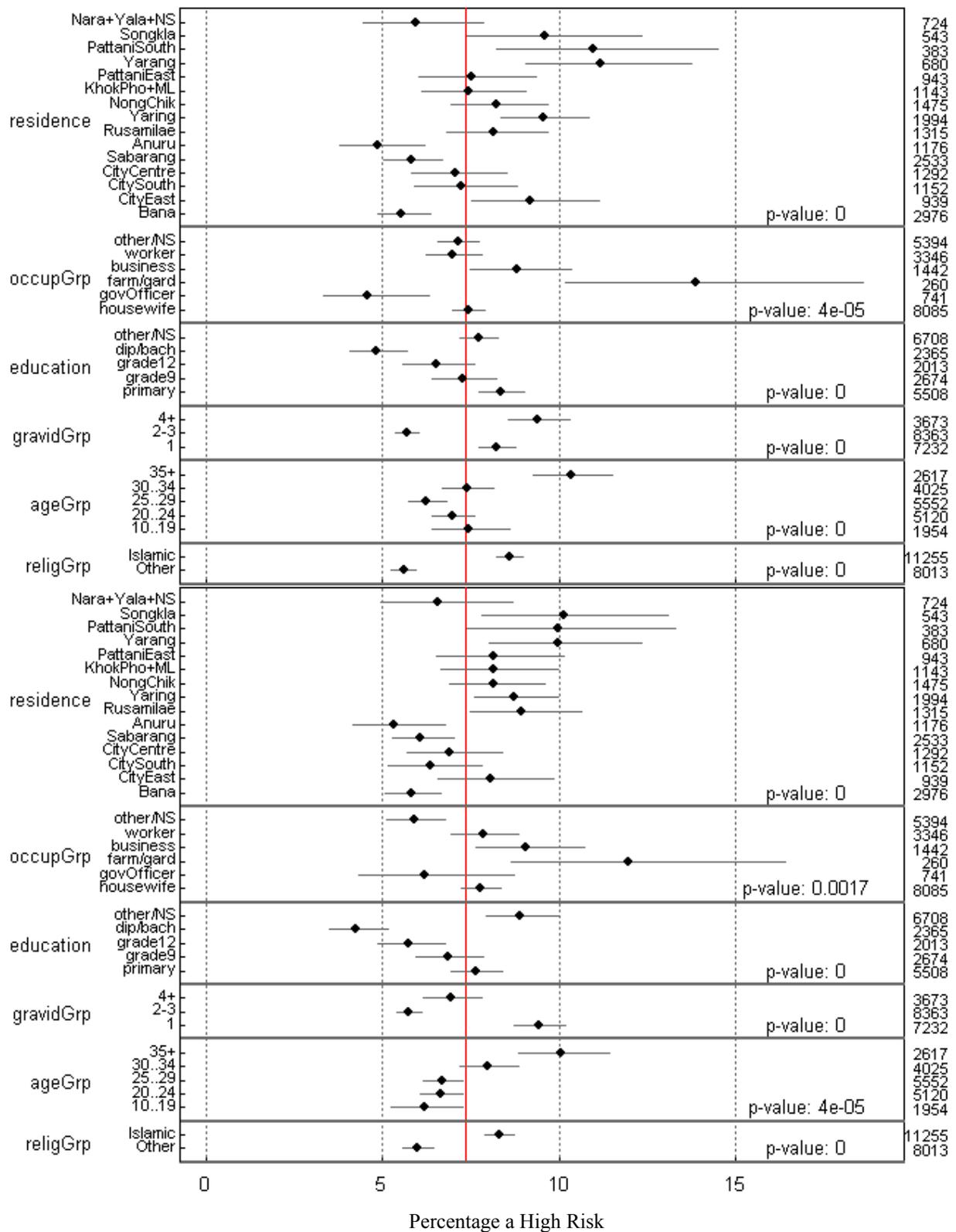


Figure 2. The 95% confidence intervals for complication-based risks among singleton deliveries to mothers with no previous Caesarean-section birth at Pattani Hospital, with respect to each of six demographic factors before (upper panel) and after (lower panel) adjustment for other factors.