Democratic Confidence Intervals for Adjusted Means and Incidence Rates

Uraiwan Sampantarak
Pattani Inland Fisheries Research and Development Center
Inland Fisheries Research and Development Bureau
Department of Fisheries, Pattani 94160
Thailand

Noodchanath Kongchouy
Department of Mathematics, Faculty of Science
Prince of Songkla University, Songkhla 90112
Thailand

Metta Kuning
Department of Mathematics and Computer Science
Faculty of Science and Technology
Prince of Songkla University, Pattani 94000
Thailand

Abstract

This paper presents confidence intervals for adjusted means using linear regression with weighted sum contrasts and applies the method to a study comparing blood lead concentrations among Pattani school children. These intervals are "democratic" in the sense that they compare the mean for each factor level with the overall mean, rather than selecting a referent. Confidence intervals based on negative binomial GLM models with sum contrasts are also recommended for comparing incidence rates for levels of a study factor after adjusting for biases due to associations with covariate factors. This method is applied to the terrorism events in regions of Southern Thailand that occurred over the period from 2004 to 2009.

Keywords: Covariate adjustment, Democratic confidence intervals, Incidence rate, Linear regression

1. Introduction

Scientific studies commonly involve comparisons of means and rates with respect to a categorical study factor of interest. This study factor could be a demographic factor (age group and gender) or location or the period of time (year or month, say). For example, an environmental study may investigate the variation in an outcome of interest such as the mean abundance of a species of fish at specific locations along a river as pointed out previously (Nowell et al., 2009). Similarly, epidemiologists may compare mean blood lead concentrations among children in different age-gender groups exposed to contamination from mining activities near their area of residence. For climate scientists the outcome of research interest nowadays is global temperature increase with particular emphasis on the rates of increase at different latitudes in recent decades, and social scientists wish to know whether residents in different locations in a terrorist area have experienced different risks of attack, and how these incidence rates change with time.

For such studies, methods exist for measuring differences in these quantities, and for assessing the statistical significance of their differences. Analysis of variance (ANOVA) based on an appropriate statistical regression model provides relevant p-values for assessing the evidence that observed differences are real. Furthermore, this method can take into account distortions due to the effects of covariates that can mask or amplify the magnitudes of these differences. Such methods are well established and comprehensively explained in the statistical literature as pointed out previously (Fox, 1997; Venables & Ripley, 2002). Despite the many available statistical methods for assessing relations between study factors and outcomes of interest, these methods do not yet routinely provide appropriate confidence intervals for graphically displaying the results of the analysis. In this paper we show how such confidence intervals may be constructed for comparing means and incidence rates, and we give two illustrations of their application.

2. Methods

2.1 Confidence intervals for means using linear regression

Linear regression as pointed out by Cook & Weisberg (1999) is a statistical method widely used to model the association between a continuous outcome and a set of fixed determinants. The model expresses the outcome variable as an additive function of the determinants. For example, if there are two categorical determinants with levels indexed by subscripts *i* and *j*, the model takes the form

$$Y_{ij} = \mu + \alpha_i + \beta_j \,. \qquad (1)$$

In this case the number of parameters is $r + c - 1$ where r and c are the number of levels of the factors α and β, respectively, thus requiring two constraints, such as $\alpha_1 = 0$ and $\beta_1 = 0$. We also assume that the errors are independent and normally distributed with mean 0 and constant standard deviation. The model may be fitted to the observations y_{ij} by least squares, giving estimates and confidence intervals for the parameters. Equation (1) generalizes straightforwardly to any specified number of categorical determinants.

When the constraints $\alpha_1 = 0$ and $\beta_1 = 0$ are used 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. This can be achieved by using sum contrasts available in commonly used software packages such as R as pointed out by R Development Core Team (2009). However, as has been pointed out by Venables & Ripley (2002), Section 6.2, these contrasts are not valid for unbalanced designs, such as those for which the number of observations is not the same for each level of a factor. Thus it is necessary to construct specific contrasts for linear regression models where the sample sizes vary with the factor levels, and this can be accomplished by using *weighted sum contrasts* rather than treatment contrasts as pointed out previously (Tongkumchum & McNeil, 2009). These weighted sum contrasts provide standard errors for the differences between each factor level and their overall mean.

The method for adjusting means to reduce the effects of covariate factors involves 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. This method is widely used in practice. For example, economists seasonally adjust time series such as unemployment rates in this way. This method also applies to data that need to be transformed to satisfy the normality assumption, by first applying the method to the transformed data and then rescaling the result to ensure that the means of the untransformed data are the same before and after adjustment. It also extends straightforwardly to any number of covariate factors.

2.2 Confidence intervals for incidence rates using the negative binomial model

The Poisson generalized linear model is widely used for modeling event counts in incidence rates as pointed out by Crawley (2005). For two additive factors as in the linear model given by equation (1), if P_{ij} is the population denominator, the expected value of the cell count N_{ij} is expressed as

$$E[N_{ij}] = P_{ij} \exp(\mu + \alpha_i + \beta_j) \,. \qquad (2)$$

However, the Poisson model often does not fit incidence data in practice because it assumes that the variance is equal to the mean, and in many situations the variance is substantially greater than the mean as pointed out previously (Jansakul & Hinde, 2004; Kongchouy et al., 2010; Mohai & Sara, 2007). The standard negative binomial GLM is a generalization of the Poisson model with the same mean λ, but the variance is $\lambda(1 + \lambda/\theta)$ where $\theta > 0$ as pointed out previously (see, for example, Chapter 7, Venables & Ripley, 2002). This over-dispersion is often the result of clustering as pointed out by Demidenko (2007). Since *deviances* rather than sums of squares are appropriate for assessing the contributions from the factors in generalized linear models, the ANOVA table is replaced by an analysis of deviance table for these models, where θ is kept fixed as pointed out previously (see, Section 7.4, Venables & Ripley, 2002).

By analogy with the method used for means based on the linear regression model, it is reasonable to define the adjusted incidence rate for level *j* of factor β as $\exp(\hat{\beta}_j + k)$, where the constant *c* is chosen to ensure that the total number of adverse events based on the fitted model matches the number observed, that is,

$$\sum n_{ij} = \sum P_{ij} \exp(\hat{\beta}_j + k). \quad (3)$$

Thus the constant is

$$k = \log\left(\sum n_{ij} / \left(P_{ij} \exp(\hat{\beta}_j)\right)\right). \quad (4)$$

3. Illustrations

3.1 Blood lead concentration in Pattani school children

A study involving 433 boys and girls aged 4-13 attending five schools in three different locations along the Pattani River in Southern Thailand as pointed out previously (Geater et al., 2000) compared geometric means of blood lead concentration in micrograms/deciliter at the five schools. The locations were (a) a village halfway up the river (school 1), (b) two villages about 10 km downstream from the river source near recently operating tin mines (schools 2 and 3), and (c) the river mouth near a ship repair facility (schools 4 and 5). The left panel of Figure 1 shows a histogram of the blood lead concentrations, with skewness coefficient 0.77. A logarithmic transformation reduces this skewness to –0.10, as shown by the histogram in the right panel.

Figure 2 shows 95% confidence intervals comparing the means of the log-transformed blood lead concentrations with respect to the two factors (1) age-group and gender combined, and (2) school, before and after adjusting for the other factor. Note that the adjustment has very little effect on the comparison with respect to schools. The children at the school halfway up the river had much lower blood lead levels, probably because those in the schools at the river mouth were exposed to contamination from the lead in ship paint, whereas those in the schools near the river source were exposed to contamination from previous mining activity.

However, the adjustment has a substantial effect on the comparison with respect to the combined age-group and gender factor. There is no clear pattern in the crude means, but after adjusting for the difference between schools, these means show a tendency to decrease with age for both sexes, except for the boys aged 11-13. This effect could be due to the fact that both the boys and the girls swim in the river when young, but only the boys continue after they reach puberty. This example is instructive because it shows how bias can occur in unbalanced study designs. Table 1 shows the numbers of children in the study sample classified by gender, age group and school. Note that 15 of the 17 boys aged 4-6 were at School 1 and this school had the lowest average blood lead levels. If the sample were perfectly balanced there would be only 5 boys aged 4-6 at this school with the remaining 12 at the other schools, and as a result the average blood lead level for these young boys would be much higher. The adjusted averages thus reflect what would be expected in a balanced sample, which would give a more accurate picture of the study population.

3.2 Victims of violence in Southern Thailand

For our second illustration we consider incidence rates per 100,000 population of non-Muslim victims of terrorism events classified by gender, age group (< 25, 25-44, and 45 or more), province of residence (Pattani, Yala and Narathiwat) and year (the six calendar years from 2004 to 2009 inclusive). These data were retrieved from a database maintained by the Deep South Coordination Centre (DSCC), Faculty of Science and Technology, Prince of Songkla University, Pattani Campus. The population denominators were obtained from the 2000 Population and Housing Census in Thailand.

To allow for interactions between pairs of these factors, we first fitted a negative binomial model containing all such interactions, for which the analysis of deviance is listed in the top panel of Table 2. Even though the age-group by year interaction is highly significant in model A, model B (for which the estimated value of θ (12.1) has standard error 2.4) was preferred to model A because it is difficult to interpret an interaction between one factor and two other factors at the same time, and the interaction between year and province is statistically much stronger than that between age-group and year. Moreover, an interaction between age-group and year is likely to be largely influenced by demographic changes in the population rather than by changes in the nature of the risk itself.

Figure 3 shows 95% confidence intervals for adjusted incidence rates based on model B. To simplify interpretation of the interaction between province and year, these factors are replaced by a single factor with 18 levels. The highest risk is seen for males aged 25-44 and the lowest risk is seen in females aged less than 25. For all three provinces the risk increased from 2004 to 2007 and then decreased with different rates, but these patterns differ by regions; Pattani increased more slowly at first but failed to decrease as much after 2007, and the risks in 2008 and 2009 stabilized to similar values in all three provinces. The hollow red circles show the crude (unadjusted) incidence rates, which are not substantially different from the adjusted rates.

4. Discussion

In this paper we have described a simple method for adjusting means and incidence rates for categorical covariates and providing corresponding confidence intervals based on a fitted linear regression model. This problem is not new and the solution we have given for means is well known. However, the fact that this solution can be applied in principle to more general statistical models is not well known, although the problem was considered as pointed out by Lane & Nelder (1982) who gave a slightly different solution. A further advantage of the method is that by using appropriately weighted sum contrasts the mean or incidence rate for each level of the study factor can be compared with the overall mean or incidence rate rather than with that for a specified reference group.

Acknowledgements

The authors express appreciation to the Deep South Co-ordination Center (DSCC), Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand for generously providing the data, the Institute of Research and Development for Health of Southern, Thailand and the European Union (EU) for financial support. Gratefully acknowledge to Emeritus Prof. Dr. Don McNeil for his guidance and advice on the statistical analysis and to the anonymous referees for invaluable suggestions.

References

Cook, R. D., & Weisberg, S. (1999). Applied regression including computing and graphics. Hoboken NJ: John Wiley.

Crawley, M. J. (2005). Statistics: an introduction using R. Colchester: John Wiley & Sons.

Demidenko, E. (2007). Poisson regression for clustered data. International statistical review, 75, 96-113.

Fox, J. (1997). Applied regression analysis, linear models, and related methods. ThousandOaks, CA: Sage.

Geater, A., Duerawee, M., Chompikul, J., Chairatanamanokorn, S., Pongsuwan, N., Chongsuvivatwong, V., & McNeil, D. (2000). Blood lead levels among school children living in the Pattani river basin: two contamination scenarios. Journal of Environmental Medicine, 2, 11-16.

Jansakul, N., & Hinde, J. P. (2004). Linear mean-variance negative binomial models for analysis of orange tissue-culture data. Songklanakarin Journal of Science and Technology, 26, 683-696.

Kongchouy, N., Choonpradub, C., & Kuning, M. (2010). Methods for modeling incidence rates with application to pneumonia among children in Surat Thani province, Thailand. Chiang Mai Journal of Science, 37, 29-38.

Lane, P. W., & Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. Biometrics, 38, 613-621.

Mohai, P., & Saha, R. (2007). Racial inequality in the distribution of hazardous waste: a national-level reassessment. Social Problems, 54, 343-370.

Nowell, L. K., Crawford, C. G., Gilliom, R. J., Nakagaki, N., Stone, W. W., Thelin, G. P., & Wolock, D. M. (2009). Regression models for explaining and predicting concentrations of organochlorine pesticides in fish from streams in the United States. Environmental Toxicology and Chemistry, 28, 1346-1358.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] Available: http://www.R-project.org.

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.

Table 1. Sample sizes of schoolchildren classified by gender, age group, and school

	boys age group							girls age group						
school	4-6	7	8	9	10	11	12+	4-6	7	8	9	10	11	12+
1	15	10	7	14	10	8	8	16	14	13	3	2	9	6
2	0	2	7	2	0	8	8	0	0	5	1	2	6	5
3	1	6	6	8	15	6	10	5	5	12	12	13	10	19
4	0	2	6	3	6	7	7	2	3	6	6	3	7	4
5	1	3	3	7	2	6	0	4	8	6	9	7	6	0

Table 2. Analysis of deviance for model containing interactions (top panel) and reduced models for fitting victim violence incidence rates in Southern Thailand

Source of variance	df	deviance	deviance/df	p-value
Full model (θ = 136.5)				
gender	1	2171.0	2171.0	< 0.00001
age-group	2	957.0	478.5	< 0.00001
year	5	587.0	117.2	< 0.00001
province	2	79.5	39.7	< 0.00001
gender × age group	2	10.2	5.1	0.12772
gender × year	5	26.7	5.3	0.05467
gender × province	2	20.1	10.0	0.01710
age-group × year	10	103.0	10.3	< 0.00001
age-group × province	4	19.9	5.0	0.08885
year × province	10	151.0	15.1	< 0.00001
residuals	64	155.6		
Reduced model A (θ = 23.8)				
gender	1	958.8	958.8	< 0.00001
age-group	2	380.6	190.3	< 0.00001
year	5	290.6	58.1	< 0.00001
province	2	30.6	15.3	< 0.00001
age-group × year	10	47.7	4.8	0.00109
year × province	10	73.7	7.4	< 0.00001
residuals	77	132.5		
Reduced model B (θ = 12.1)				
gender	1	583.0	583.0	< 0.00001
age-group	2	232.0	116.0	< 0.00001
year	5	185.0	36.9	< 0.00001
province	2	26.2	13.1	0.00084
year × province	10	50.3	5.0	< 0.00001
residuals	87			

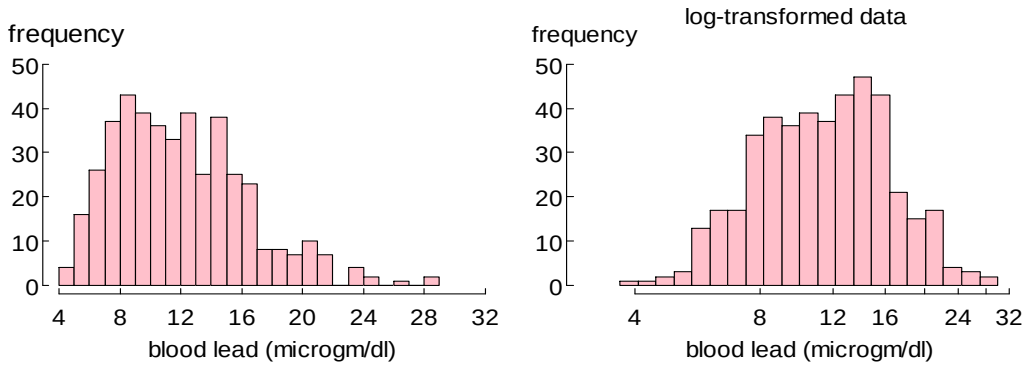

Figure 1. Histograms of blood lead concentrations in micrograms per deciliter before (left panel) and after transformation using natural logarithms (right panel)

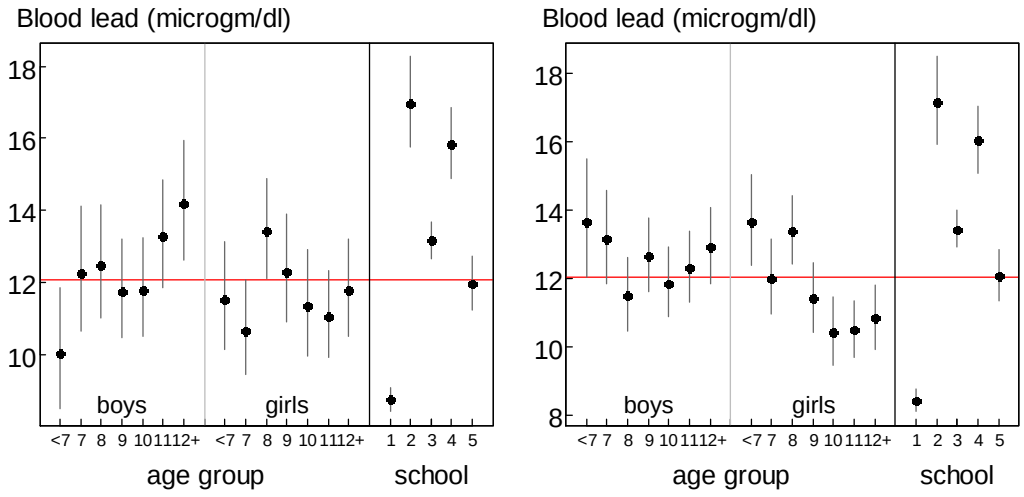

Figure 2. The 95% confidence intervals of average blood lead levels before (left panel) and after (right panel) adjusting for the other factor

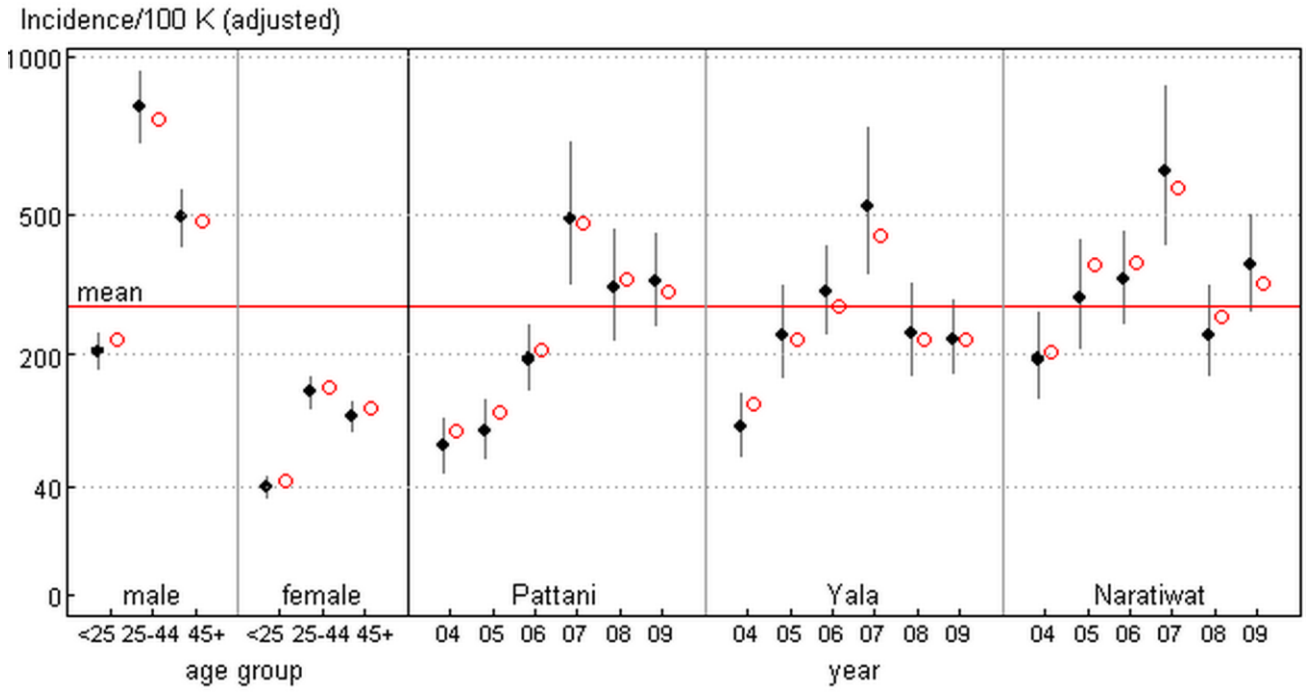

Figure 3. Annual incidence rates of victim terrorism violence per 100,000 populations in the three southernmost provinces of Thailand classified by gender, age group, and years 2004-2007 (the hollow circles indicate corresponding unadjusted incidence rates)