Songklanakarin J. Sci. Technol.
31 (2), 151-156, Mar. - Apr. 2009

Songklanakarin Journal
of Science and Technology
http://rdo.psu.ac.th/sjst

Original Article

Confidence intervals using contrasts for regression model

Phattrawan Tongkumchum[1] and Don McNeil[2]

[1] *Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand.*

[2] *Department of Statistics, School of Economic and Financial Studies, Macquarie University, Australia*

Received 3 June 2008; Accepted 13 October 2008

Abstract

A graph of confidence intervals can be used to report results from a regression model with explanatory variables as factors. In this paper we describe a method for computing and displaying confidence intervals using weighted sum contrasts to compare population means in unbalanced designs. We extend this method to models with covariates and logistic regression models.

Keywords: 95% confidence interval, weighted sum contrasts, unbalanced designs

1. Introduction

One of the most widely used graphs in statistical data analysis shows confidence intervals for population means of two or more groups. Conventional plots provided by statistical packages show separate confidence intervals containing the corresponding population mean for each group with specified probability, usually 95%. These intervals are referred to as “between-subject” confidence intervals (see, for example, Masson and Loftus, 2003). If the objective is to compare the means - in which case the null hypothesis is that the population means are all the same - this conventional plot can be misleading. When comparing two means, for example, individual 95% confidence intervals can overlap even though the difference is statistically significant at the 5% level. In this case it is more appropriate to graph a confidence interval for the difference between the means, possibly by treating one of the groups as a reference and plotting a confidence interval for the difference centred at the other mean, giving a “between-treatment” confidence interval. However, when comparing more than two groups, this method gives different graphs depending on which group is selected as the reference, and gives wider confidence intervals when the reference group has smaller sample size.

In this paper we suggest a way of constructing confidence intervals for comparing means that does not involve selecting a reference group and thus gives an informative confidence interval for comparing each mean with the overall mean. The method simply involves the application of appropriate contrasts in a regression model, and extends to generalised linear models including other factors.

2. Method

2.1 Contrast matrices

The method involves the choice of a particular contrast matrix from those described by Venables and Ripley (2002) as follows. Suppose that *f* is a factor with *k* classes used as an explanatory variable in a linear regression model being fitted to *n* observations. The equations expressing *k*-1 of the *k* contrasts in terms of the individual class means take the form $\boldsymbol{a}^* = \boldsymbol{D}_1\boldsymbol{a}$, where $\boldsymbol{a}$ is the column vector containing the *k* class means. Solving these equations gives $\boldsymbol{a} = \boldsymbol{C}_1\boldsymbol{a}^*$

*Corresponding author.
Email address: tphattra@bunga.pn.psu.ac.th

where $\boldsymbol{C}_1$ is the inverse of the matrix $\boldsymbol{D}_1$. We omit the first column of $\boldsymbol{C}_1$ to obtain the desired *contrast matrix* $\boldsymbol{C}$, which is then specified when fitting the regression model.

Let $\bar{y}_j$ denote the mean and n_j the sample size for class *j*, so that the overall sample mean is $\bar{y} = r_1\bar{y}_1 + r_2\bar{y}_2 + ... + r_k\bar{y}_k$ where $r_j = n_j/n$ is the proportion of cases in class *j*. Then the equations we use are as follows.

$$\alpha_1^* = r_1\bar{y}_1 + r_2\bar{y}_2 + ... + r_k\bar{y}_k \qquad (1)$$

$$\alpha_{j+1}^* = \bar{y}_j - \bar{y} \; (j = 1, 2, \ldots, k). \qquad (2)$$

The matrix $\boldsymbol{D}_1$ comprises equation (1) and any *k*-1 of the set (2). The matrix $\boldsymbol{C}$ then takes the form $\begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{r}^* \end{bmatrix}$ where $\boldsymbol{I}$ is the (*k*-1) × (*k*-1) identity matrix and $\boldsymbol{r}^*$ is the row vector having length *k*-1 with elements $-r_1/r_k, -r_2/r_k, \ldots, -r_{k-1}/r_k$. The standard errors that result when a regression model is fitted using $\boldsymbol{C}$ as the contrast matrix are used to obtain confidence intervals for the means used in the contrasts. Finally, we obtain the confidence interval for the omitted mean by repeating the procedure with this mean included and another omitted.

When the classes are all of the same size, the contrasts resulting from the method described above are known as *deviation* contrasts (see, for example, Wendorf, 2004) or *sum* contrasts (Venables and Ripley, 2002) and are available using standard software including SPSS and R. However, when the classes are not all of the same size these standard contrasts do not give valid standard deviations for comparing the class means with the overall mean. For clarity we call these general contrasts *weighted sum* contrasts to distinguish them from the *unweighted sum* contrasts that are valid only for balanced designs.

3. Simple illustrations

To illustrate the various confidence intervals, consider some data from a study comparing blood lead levels among children attending primary schools along the Pattani River in Southern Thailand (Geater *et al.*, 2000). Figure 1 shows 95% confidence intervals for mean blood lead level by gender for 27 boys and 19 girls at Thamthalu Primary School, located in an area where the environment had been contaminated by smelting from tin mines. For these data (listed in the Appendix) the linear regression model gives the p-value 0.030 for testing the null hypothesis that the population mean is the same for each sex. The bottom right-hand panel shows individual ("within-treatment") confidence intervals. These confidence intervals are useful for showing that both the boys and the girls had average blood lead levels greater than 10 micrograms per decilitre, the safety threshold recommended by WHO, but given that the confidence intervals overlap, they do not enable the viewer to easily conclude that the means are statistically significantly different. The top left-hand panel shows a 95% confidence interval for the difference between the means, based on the standard error for this difference that results when the linear model is fitted using the standard treatment contrast, that is, when the model is fitted with an intercept and an indicator variable taking values 1 for girls and 0 for boys. The fact that this ("between-treatment") confidence interval is entirely below the line corresponding to the mean for the boys indicates that the means are statistically different at the 5% significance level.

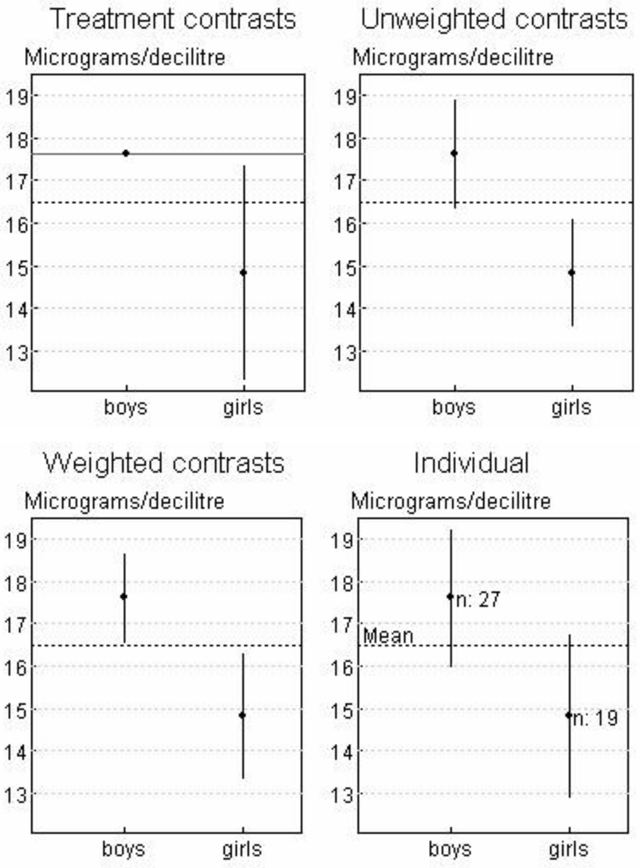

Figure 1. Various confidence intervals for blood lead levels by gender

The graphs in the top right and bottom left panels of Figure 1 show the confidence intervals based on the unweighted and weighted sum contrasts, respectively. The confidence intervals for the unweighted sum contrasts are necessarily of equal width because they take no account of the difference in the sample sizes, and thus give the confusing impression that the mean blood level for the boys is statistically no different from the overall mean whereas that for the girls is below the overall mean. In contrast, the confidence intervals for the weighted sum contrasts correctly show that the means are evidently different.

Figure 2 shows individual and comparative 95% confidence intervals for the blood lead levels of the same 46 children by four age groups.

Again the individual 95% confidence intervals in the right-hand panel enable the viewer to compare the mean for each age group with a standard value, and the confidence

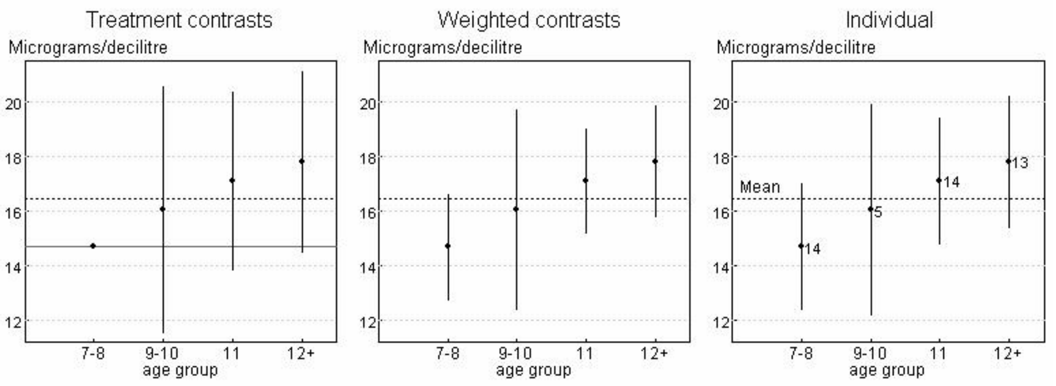

Figure 2. Confidence intervals for children's blood lead levels by age group.

intervals for the treatment contrasts shown in the left panel are appropriate when a natural reference group exists, but the weighted sum contrasts in the middle panel may still be the most appropriate way of graphing confidence intervals. Note that the confidence intervals based on the weighted sum contrasts are all smaller than the corresponding individual confidence intervals. It can be shown that this shrinkage factor for class *j* is $\sqrt{1-r_j}$.

4. Adjusting for other factors

When the model contains more than one explanatory factor its interpretation becomes more complex due to the possibility of confounding when these factors are correlated, and it is thus instructive to show appropriate confidence intervals in a graph. For example, the graph can show aligned confidence intervals for each factor, both before and after adjusting for the other factors. The unadjusted plots show how the sample means differ between the various classes for each factor, whereas the adjusted plots show the corresponding comparisons in the conceptual population, and comparing the two sets of plots shows the extent of distortions due to confounding.

Common sense dictates that the overall mean for a factor, obtained as the average of class means weighted by their sample sizes, must be the same before and after adjusting for other factors. For linear models this can be achieved simply by adding an appropriate constant to the coefficients obtained from the model.

Figure 3 compares 95% confidence intervals of blood lead levels by school and by age-gender class (comprising seven age groups in this case) for all five schools considered in the study by Geater *et al.* (2000). The unadjusted means appear in the left panels whereas those adjusted for the other factor appear on the right. For comparing the mean blood levels with respect to school, it is clear that adjusting for age-gender class makes very little difference to the result. However, when comparing the mean blood lead levels for different age-gender classes, adjusting for school makes a substantial difference to the pattern, where it can be seen that the confidence interval for 4-6 year-old boys increased by nearly 4 micrograms/decilitre when an adjustment was made for the school. In this case the confounding was due to the fact that the children attending Tachi School were older than those attending the other schools.

5. Logistic model

For a binary outcome, a graph of confidence intervals of population proportions is appropriate for comparing the difference of two or more groups. The proportions of adverse outcomes and their corresponding standard errors may be estimated by fitting a logistic regression model, and again it is appropriate to use weighted sum contrasts to obtain the standard errors underlying the confidence intervals for comparing these proportions.

If there are two categorical determinants, and p_{ij} denotes the probability of the adverse outcome in categories *i* and *j* of these determinants, respectively, the simplest such model takes the additive form $\ln(p_{ij}/(1-p_{ij}) = c+a_i+b_j$ and the prevalence itself is thus expressed as $p_{ij} = 1/(1+\exp(-c-a_i-b_j)$.

Logistic regression provides a straightforward method for adjusting a prevalence that varies with a determinant of interest for the effect of a covariate determinant. To calculate the adjusted prevalence for category *i* of the determinant of interest, the term b_j is replaced by a constant *b*, that is, $p_{ij}^* = 1/(1+\exp(-c-b-a_i)$.The value of *b* is chosen to ensure that sum of the expected number of adverse outcomes is equal to the sum of the observed number, that is, $\Sigma p_i^* n_i = \Sigma p_i n_{i,}$ where n_i is the sample size in category *i* of the determinant of interest. This method extends straightforwardly to additional covariates.

As an illustration of the method, we consider the adverse outcome to be discontinuation for students who were admitted to study 4-year bachelor degrees at Pattani Campus of Prince of Songkla University from 1999 to 2002 (Sitichai *et al.*, 2008). The explanatory variables are year of admission, faculty, and gender-religion group. These data are listed in the Appendix. Figure 4 shows 95% confidence interval graphs of the discontinuation rates for each factor based on an additive logistic model using both weighted sum contrasts

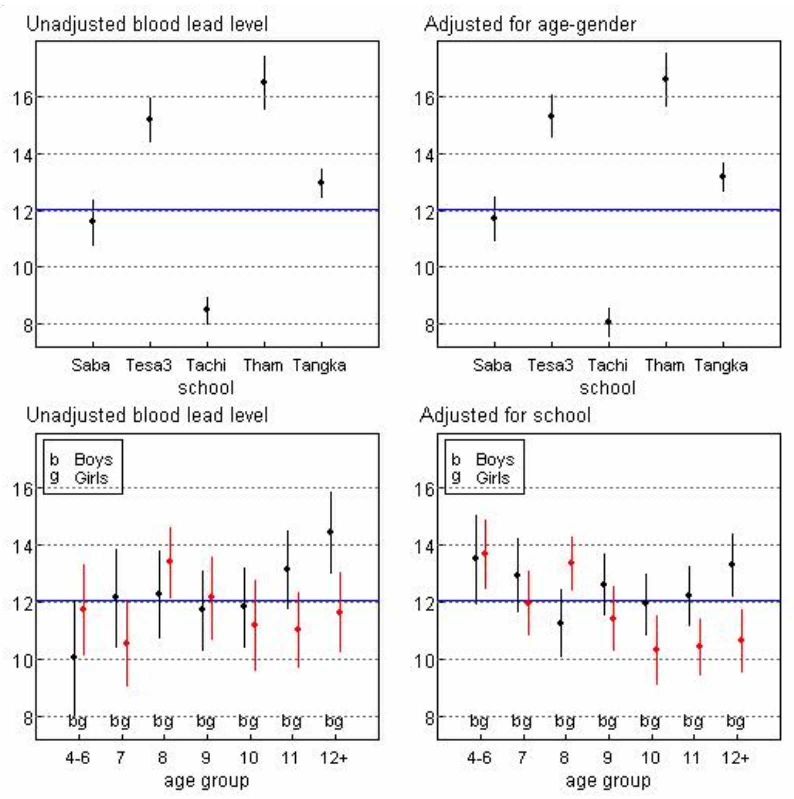

Figure 3. Confidence intervals for children's blood lead levels by school and by gender-age class

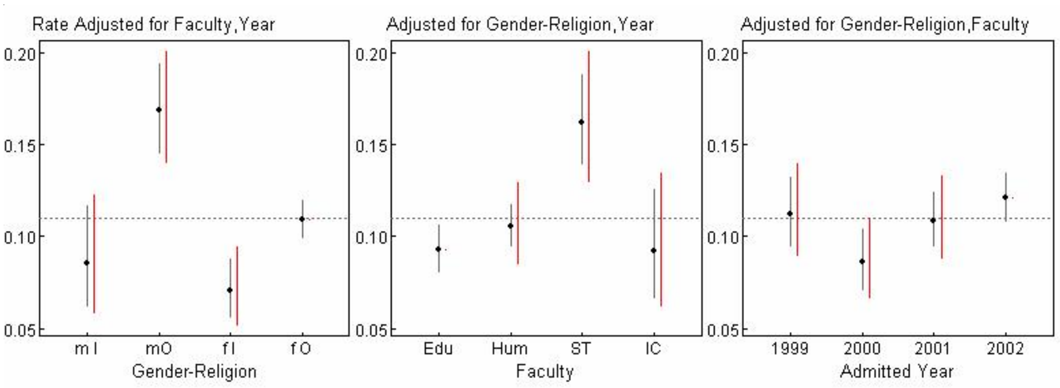

Figure 4. Confidence intervals for discontinuation rates by gender-religion, faculty and admitted year

and treatment contrasts.

The shorter (left-most) intervals containing the points are based on the weighted sum contrasts. To minimise the widths of confidence intervals based on the treatment contrasts, each referent group is taken as the class with the largest sample size. However, no such choice is necessary when plotting the confidence intervals based on the weighted sum contrasts.

6. Conclusions

We have described the use of weighted sum contrasts for graphing confidence intervals with the objective to compare population means in unbalanced designs. The method extends the widely used sum contrasts by setting an appropriate contrast matrix for a factor in the regression model. This method can also be used in fitting a model with covariates and a logistic regression model. In a future paper we plan to extend this method to the situation when there is an ordinal explanatory variable in the model.

References

Geater, A., Duerawee, M., Chompikul, J., Chairatanamanokorn, S., Pongsuwan, N., Chongsuvivatwong, V. and McNeil, D. 2000. Blood lead levels among school children living in the Pattani river basin: two contamination scenarios, Journal of Environmental Medicine 2(1), 11-16.

Masson, MJ. and Loftus, GR. 2003. Using confidence intervals for graphically based data interpretation, Canadian Journal of Experimental Psychology, 57(3), 203-220.

Sitichai, R., Tongkumchum, P. and McNeil, N. 2008. Discontinuation among university students in Pattani, Songklanarin Journal of Social Sciences and Humanities, 14(3), 399-408.

Venables, W.N. and Ripley, B.D. 2002. Modern Applied Statistics with S. NewYork, Springger-Verlag.

Wendorf, CA. 2004. Primer on Multiple Regression Coding: Common Forms and the Additional Case of Repeated Contrasts, Understanding Statistics, 3(1), 47-57.

Appendix

Blood lead level data

ID	bloodLead	age	gender	ID	bloodLead	age	gender
1	28.1	12+	boy	24	15.1	12+	boy
2	15.5	12+	boy	25	24.8	12+	boy
3	14.4	11	boy	26	23.2	11	boy
4	10.6	12+	girl	27	15.3	11	girl
5	20.6	12+	boy	28	13.7	11	boy
6	21.4	12+	boy	29	12.7	11	boy
7	12.0	11	girl	30	16.6	11	boy
8	15.9	11	girl	31	11.0	12+	girl
9	19.0	11	boy	32	21.0	11	girl
10	18.3	11	boy	33	24.0	11	boy
11	12.1	9-10	girl	34	17.3	9-10	boy
12	12.5	7-8	boy	35	19.4	9-10	girl
13	18.8	7-8	boy	36	13.3	7-8	boy
14	15.7	7-8	girl	37	18.6	7-8	boy
15	17.3	7-8	girl	38	8.1	7-8	boy
16	15.1	9-10	girl	39	16.7	7-8	boy
17	21.3	12+	boy	40	13.2	7-8	boy
18	14.9	12+	girl	41	9.9	7-8	girl
19	20.4	12+	boy	42	21.1	7-8	boy
20	13.5	12+	girl	43	14.9	7-8	girl
21	12.3	11	girl	44	15.1	7-8	girl
22	21.3	11	girl	45	10.5	7-8	boy
23	14.5	12+	girl	46	16.4	9-10	boy

Student discontinuation data

Faculty	*Year admitted*	MI		FI		MO		FO	
		Total	*Disc.*	*Total*	*Disc.*	*Total*	*Disc.*	*Total*	*Disc.*
Edu	1999	10	0	41	0	47	5	148	13
	2000	16	2	68	2	48	8	137	7
	2001	12	0	60	6	61	7	227	24
	2002	28	3	94	5	83	15	352	36
Hum	1999	15	0	32	2	98	17	267	21
	2000	20	1	57	3	62	9	208	18
	2001	20	1	61	2	93	13	360	42
	2002	56	3	128	9	118	14	399	60
ST	1999	11	4	15	7	70	29	79	9
	2000	14	3	7	0	64	17	86	7
	2001	14	2	18	2	85	16	106	14
	2002	22	2	26	4	82	14	136	20
IC	1999	22	0	63	3	0	0	0	0
	2000	34	0	86	2	0	0	0	0
	2001	71	7	124	8	1	0	3	0
	2002	78	8	181	12	4	1	1	1