Demography, Vol. 26, No. 4, November 1989

An Alternative Purging Method: Controlling the Composition-Dependent Interaction in an Analysis of Rates

Yu Xie

Population Studies Center, University of Michigan, 1225 South University Avenue, Ann Arbor, Michigan 48104

The purging method controlling for the composition-group interaction developed by Clogg and his associates has proven useful in demographic research. This article proposes an alternative method, partial *CD* purging, that controls the interaction between composition and the dependent variable. The purged rates from this new method are invariant to changes in the marginal distribution of composition, but those from the earlier purging method are not. Mathematical relationships between the proposed method and other techniques are also explored.

An important step has been taken by Clogg and his associates (Clogg, 1978; Clogg and Eliason, 1988; Clogg and Shockey, 1985; Clogg, Shockey, and Eliason, 1987) in a series of papers centered around the idea of purging the confounding effects of composition in crude rates. The purging method espoused in these papers provides a convenient way to bridge the gap between statistical models and descriptive rates (for a review, see Hoem, 1987). Communication between demographers versed in statistical modeling and other researchers is thus facilitated.

There are several variations on the purging method. Their essential feature is the elimination of either the partial or the marginal composition-group interaction, when the three-factor interaction is absent, in a cross-classification of counts by composition, group, and dependent variables. When the three-factor interaction is present, purging is still possible by concurrently adjusting for the three-factor interaction.

In this article I propose an alternative purging method that controls the partial interaction between composition and the dependent variable. The rates resulting from this alternative purging method differ in general from the purged rates of either partial or marginal composition-group purging. For some applications at least, the former have desirable properties.

What to Purge, *CG* or *CD*?

Following Clogg (1978), Clogg and Eliason (1988), and Clogg, Shockey, and Eliason (1987), let C, G, and D denote the composition, group, and dependent variables, in a cross-classified table of $C \times G \times D$.[1] The categories of C, G, and D are indexed as i ($i = 1, \ldots, I$), j ($j = 1, \ldots, J$), and k ($k = 1, \ldots, K$), respectively. The purging method proposed by Clogg and his associates eliminates the confounding effects of the partial or marginal CG interaction as well as of the three-factor CGD interaction when it is present. An alternative method is to purge the partial CD interaction and the three-factor CGD interaction. This alternative is reasonable because the relationship between G and D cannot be confounded by C once the CD and CGD interaction terms are purged.

It is well known that for the C variable to be confounding, both the *CG* and *CD* interaction terms must be present. Using linear regression as an analogy, bias due to an omitted variable is present only when the omitted variable both is correlated with the independent variable of interest *and* affects the dependent variable. For a three-way contingency table of $C \times G \times D$, the table is collapsible over C if either *CG* and *CGD* interactions or *CD* and *CGD* interactions are nil (Bishop, Fienberg, and Holland, 1975:39). When the table is collapsible in the dimension of C, information about C is irrelevant. Hence the *GD* association is not confounded by the C variable. Thus partial *CD* purging is an acceptable alternative method of purging, in an analogous way to partial *CG* purging. Rates resulting from the two methods, however, can be very different. As I will show, the partial *CD* purged rates are invariant to changes in the marginal distribution of C, but the partial *CG* purged rates are not.[2] This property makes partial *CD* purging preferable to *CG* purging at least for some applications.

Borrowing Clogg and Eliason's (1988) notation, a cell frequency F_{ijk} can be described by the following multiplicative model:

$$F_{ijk} = \tau\, \tau_i^C\, \tau_j^G\, \tau_k^D\, \tau_{ij}^{CG}\, \tau_{ik}^{CD}\, \tau_{jk}^{GD}\, \tau_{ijk}^{CGD}, \qquad (1)$$

where the τ parameters are subject to the usual normalizations (Clogg, 1978). For simplicity, I only consider the case with no three-factor interaction, where $\tau^{CGD} = 1$.[3] The partial *CG* purging method first obtains purged frequencies by dividing cell frequencies by τ^{CG} (F_{ijk}/τ^{CG}) and then calculates rates based on the purged frequencies. Similarly, the *CD* purging method divides observed frequencies by τ^{CD} to obtain purged frequencies (F_{ijk}/τ^{CD}) and then calculates rates from the purged frequencies.

In the following example, I will show that adjusted rates from the *CD* purging method are invariant to changes in the marginal distribution of composition, whereas those from the *CG* purging method are not. Invariance to the marginal distribution of composition can be desirable for a number of reasons. First, the invariance means that composition is statistically controlled as an exogenous variable. Everything else being equal, the *CD* purged rates, like structural parameters, do not vary as composition changes. This is especially useful in comparative (or trend) analysis (e.g., see Hauser and Grusky, 1988). Second, the invariance is already an old friend in the familiar framework of logit analysis. As I will show, there is a close relationship between logit analysis and partial *CD* purging. Third, for samples that are stratified on variable C, the marginal distribution of C in the sample is generally different from that in the population, depending on the particular sampling design. It is preferable to have measures that are not affected by sampling design.

An Example With Hypothetical Data

I demonstrate the invariance property of the *CD* purging method by using the hypothetical data shown in Table 1. Case 1 is a simple $2 \times 2 \times 2$ three-way table with no three-factor interaction. In case 2, I alter the table by doubling the frequencies of the cells that are related to C_1, thus changing the marginal distribution of variable C. The changes in the parameters are displayed in Table 2.

As shown in Table 2, cases 1 and 2 are identical except for the main effect, τ, and the marginal effect, τ^C. The other two marginal parameters and all of the two-way association parameters are identical. Case 3 was constructed in such a way that τ^C is the same as in case 1 while τ^G and τ^{CG} are different. I apply both the partial *CG* purging and the partial *CD* purging methods to each of the three cases and obtain the purged rates reported in Table 3.

As expected, the crude rates differ across cases 1, 2, and 3. What is especially interesting is that the *CG* purged rates in case 2 differ from those in case 1, whereas the *CD* purged rates are identical in the two cases. Recall that the only difference between the two cases is

Table 1. Hypothetical Data

	C_1		C_2	
Case	D_1	D_2	D_1	D_2
1				
G_1	1	2	2	8
G_2	2	8	8	64
2				
G_1	2	4	2	8
G_2	4	16	8	64
3				
G_1	2	4	2	8
G_2	1	4	8	64

that the marginal distribution of C is altered. It is thus illustrated in the comparison of case 1 and case 2 that the *CD* purged rates are invariant to changes in the marginal distribution of the C variable, but the *CG* purged rates are not. Note, however, that both the *CG* and *CD* purged rates are invariant to changes in the *CG* association. This is shown by contrasting case 1 and case 3.

Proof of Invariance

The invariance property of partial *CD* purged rates can be proved analytically. Without loss of generality, let us assume that there is a dichotomous dependent variable ($D = D_1$, D_2) and that we want to know the proportion of occurrences of D_1. The summary rate for the jth group is

$$R_j = F_{+j1}/(F_{+j1} + F_{+j2}), \tag{2}$$

where F_{+jk} is the sum of the cell frequencies over variable C. Substituting equation (1) into equation (2) and simplifying gives

$$R_j = \frac{\tau_1^D \tau_{j1}^{GD} \sum_{i=1}^{I} \tau_i^C \tau_{ij}^{CG} \tau_{i1}^{CD} \tau_{ij1}^{CGD}}{\tau_1^D \tau_{j1}^{GD} \sum_{i=1}^{I} \tau_i^C \tau_{ij}^{CG} \tau_{i1}^{CD} \tau_{ij1}^{CGD} + \tau_2^D \tau_{j2}^{GD} \sum_{i=1}^{I} \tau_i^C \tau_{ij}^{CG} \tau_{i2}^{CD} \tau_{ij2}^{CGD}}. \tag{3}$$

Equation (3) states that in general the τ and τ^G parameters are irrelevant in calculations of any kind of summary rates (also see Clogg, 1978). By assumption (or by simultaneously purging), $\tau^{CGD} = 1$. For the *CD* purging method, $\tau^{CD} = 1$. Then expression (3) can be reduced to

$$R_j^* = \tau_1^D \tau_{j1}^{GD}/(\tau_1^D \tau_{j1}^{GD} + \tau_2^D \tau_{j2}^{GD}). \tag{4}$$

Table 2. Parameters of the Three Hypothetical Cases

Case	τ	τ^C	τ^G	τ^D	τ^{CG}	τ^{CD}	τ^{GD}
1	4.757	0.500	0.500	0.500	1.189	1.189	1.189
2	6.727	0.707	0.500	0.500	1.189	1.189	1.189
3	4.757	0.500	0.707	0.500	1.682	1.189	1.189

Table 3. Comparison of Summary Rates (for $D = D_1$) From Partial *CG* and *CD* Purging and Schoen's Index

Case	Crude	*CG* purging	*CD* purging	Schoen's index
1				
G_1	0.231	0.223	0.261	0.258
G_2	0.122	0.126	0.150	0.149
2				
G_1	0.250	0.240	0.261	0.258
G_2	0.130	0.136	0.150	0.149
3				
G_1	0.250	0.223	0.261	0.258
G_2	0.117	0.126	0.150	0.149

In this case, the adjusted rates are functions of only the τ^D and τ^{GD} parameters, for the τ^C and τ^{CG} parameters drop out of the equation. To be more explicit, the purged rates are invariant to changes in both the marginal distribution of composition and the CG association. In contrast, the CG purged rates still involve τ^C and τ^{CD}. This can be shown by substituting $\tau^{CG} = 1$ and $\tau^{CGD} = 1$ into equation (3):

$$R_j^{**} = \frac{\tau_1^D \tau_{j1}^{GD} \sum_{i=1}^{I} \tau_i^C \tau_{i1}^{CD}}{\tau_1^D \tau_{j1}^{GD} \sum_{i=1}^{I} \tau_i^C \tau_{i1}^{CD} + \tau_2^D \tau_{j2}^{GD} \sum_{i=1}^{I} \tau_i^C \tau_{i2}^{CD}}. \quad (5)$$

Equation (5) reveals that the CG purged rates are invariant to changes in the CG association but not to changes in the marginal distribution of C. This property was illustrated in the numerical example (case 1 vs. case 3 and case 1 vs. case 2).

Another way to interpret the difference between the CG and the CD purged rates is to consider the properties of collapsible three-way tables. From equation (2), summary rates can be seen as measures obtained from the two-way $G \times D$ table collapsed over C. In general, a three-way table is not collapsible in the dimension of C because the C variable has confounding effects on the two-way $G \times D$ table. After either the CG or CD interactions are eliminated in calculating the purged frequencies, the three-way table is then collapsible over C. In this case, the summary rates computed directly from the collapsed table are "adjusted" rates. After collapsing, we have a two-way table that can be modeled as[4]

$$F_{+jk}^{*} = \delta \delta_j^G \delta_k^D \delta_{jk}^{GD}. \quad (6)$$

According to Bishop, Fienberg, and Holland (1975:41), when the $C \times G \times D$ table is collapsible over variable C because τ^{CD} is 1, parameters τ^D and τ^{GD} are preserved ($\delta^D = \tau^D$, $\delta^{GD} = \tau^{GD}$), but τ and τ^G are not ($\delta \neq \tau$, $\delta^G \neq \tau^G$). Conversely, τ and τ^D are not preserved after collapsing if the collapsibility is derived from $\tau^{CG} = 1$. It is thus evident that at the level of odds ratios, there is a complete equivalence among partial CG purging, partial CD purging, and logit analysis because after collapsing, we have $\delta^{GD} = \tau^{GD}$ (also see Santi, 1989). At the level of rates (or odds), however, adjusted rates from the two purging methods differ because the δ^D terms in the two collapsed $G \times D$ tables are not the same. For the partial CD purging method, the δ^D parameter, which determines the summary rates after δ^{GD} is fixed, is identical to the original τ^D, whereas for the partial CG purging method, δ^D is something different.

Comments on Applications of *CD* Purging and Other Related Methods

The preceding discussion indicates that the marginal distribution of composition is implicitly controlled in partial *CD* purging. As a result, partial *CD* purging is invariant to changes in the marginal distribution of composition. Given this invariance, one can arbitrarily change the marginal distribution of composition while retaining the group-dependent association within each of the compositional categories. This should not affect the partial *CD* purged rates. In particular, one can make the composition evenly distributed. This convenient form makes partial *CD* purging equivalent to Teachman's (1977) method and logit analysis.

Following Schoen's (1970) mortality index, which is the geometric mean of rates across compositional categories, Teachman (1977) suggested a method that computes an adjusted rate from the geometric mean of odds: $R_j^T = w_j/(1 - w_j)$, where w_j is the geometric mean of odds F_{ij1}/F_{ij2} across all compositional categories. Teachman's method is equivalent to the partial *CD* purging method proposed in this article. This equivalence can be proved by noting that the τ^D and τ^{GD} terms in equation (4) can be viewed as geometric means of $\tau^{(C)D}$ and $\tau^{(C)GD}$ from the following equation for each of the categories of C (Bishop, Fienberg, and Holland, 1975:32):

$$F_{(i)jk} = \tau_{(i)}\, \tau_{(i)j}^{(C)G}\, \tau_{(i)k}^{(C)D}\, \tau_{(i)jk}^{(C)GD}. \tag{7}$$

As Clogg (1978) showed, Teachman's method yields numerically different rates from Schoen's index. What is common to the two measures is that they both standardize for composition by assigning equal weights to compositional categories. Thus Schoen's index has the same invariance property as the *CD* purged rates. For my numerical example, Schoen's index is displayed in the last column of Table 3.

The *CD* purging method is also closely related to logit analysis. In a logit analysis of the $C \times G \times D$ table, the log odds ratio of the dependent variable, that is, $\log[\Pr(D = D_1)/\Pr(D = D_2)]$, would be regressed on both *C* and *G*. Predicted rates can be obtained by first varying *G* (while holding *C* constant) and then converting log odds ratios to probabilities (rates). For all three cases in my example with the hypothetical data in Table 1, the same logit regression is obtained and shown as follows:

$$\log[\Pr(D = D_1)/\Pr(D = D_2)] = 0.6931 + 0.6931X_C + 0.6931X_G, \tag{8}$$

where $X_C = 1$ if $C = C_1$ and $X_C = 0$ otherwise, and $X_G = 1$ if $G = G_1$ and $X_G = 0$ otherwise. In other words, variations in the marginal distributions of *C* and *G* and in the *CG* interaction do not affect the parameters of the logit regression. This invariance property accounts for the common use of the word "structural" in regression analysis. As was shown before, the *CD* purged rates have the same desirable property of invariance as logit analysis. In fact, the *CD* purged rates can be obtained by giving equal weights (i.e., $X_C = 0.5$) to the *C* variable in equation (8) and then calculating the predicted rates conditional on X_G. From this point of view, *CD* purging can be seen as a special application of logit analysis.

When adjusted rates are affected by the marginal distribution of composition, the rate adjustment requires a set of values for the composition variable. They can be given either explicitly, by having a "standard group" as in conventional direct standardization, or implicitly, by estimation as in the partial and marginal *CG* purging (Clogg, 1978; Clogg and Eliason, 1988; Clogg, Shockey, and Eliason, 1987). In some sense, the *CG* purged rates have a "scale" that is interpretable with reference to a particular composition criterion. When the researcher is concerned with such a "scale" of purged rates, as exemplified in Clogg and Shockey (1985), the method of partial and marginal *CG* purging (including direct standardization as a special case) appears to be the natural one to use. In many other instances, the researcher may be concerned with "intrinsic" rates that are invariant to the changes in

the distribution of composition. When this is the case, the method of partial *CD* purging is superior. Which method is more appropriate in practice depends largely on the specific research problem.

If the three-factor interaction is present, the *CD* and *CGD* interactions can be purged simultaneously. As in the case of *CG* purging, the jackknife technique can also be applied for statistical inference (Clogg and Eliason, 1988; Clogg, Shockey, and Eliason, 1987). In particular, the jackknife can be used to estimate the standard errors of the adjusted rates and to test for group differences in the adjusted rates.[5]

Notes

[1] In general, each of these variables can represent a combination of several variables (Clogg and Eliason, 1988:272).

[2] It can also be shown that the marginal *CG* purged rates are not invariant to changes in the marginal distribution of composition. Since the partial *CD* method resembles the partial *CG* purging method in a symmetrical way, this article focuses on the comparison of partial *CG* and partial *CD* purging.

[3] If the three-factor interaction is not nil, one can concurrently purge τ^{CGD} (Clogg and Eliason, 1988; Clogg, Shockey, and Eliason, 1987). All of the conclusions to be drawn in the article are then applicable. No other changes are necessary.

[4] F^* instead of F is used here to refer to purged frequencies. For partial *CG* purging, $F^* = F/\tau^{CG}$; for partial *CD* purging, $F^* = F/\tau^{CD}$.

[5] One can easily modify Clogg and Eliason's computer program PURGE to obtain *CD* purged rates and *CD*-and-*CGD* purged rates and to perform jackknife estimation of the standard errors of the adjusted rates and of group differences in the adjusted rates. For this purpose, a modified version of PURGE in FORTRAN code is available from the author on request.

Acknowledgment

I am grateful to Clifford C. Clogg, Robert M. Hauser, Jan M. Hoem, Arthur Sakamoto, Lawrence Santi, and two anonymous reviewers for their helpful advice and comments.

References

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass.: MIT Press.

Clogg, C. C. 1978. Adjustment of rates using multiplicative models. *Demography* 15:523–539.

Clogg, C. C., and S. R. Eliason. 1988. A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons. *American Sociological Review* 53:267–283.

Clogg, C. C., and J. W. Shockey. 1985. The effect of changing demographic composition on recent trends in underemployment. *Demography* 22:395–414.

Clogg, C. C., J. W. Shockey, and S. R. Eliason. 1987. *A General Statistical Framework for Adjustment of Rates.* Unpublished manuscript, Pennsylvania State University, Population Issues Research Center.

Hauser, R. M., and D. Grusky. 1988. Cross national variation in occupational distributions, relative mobility changes, and intergenerational shifts in occupational distributions. *American Journal of Sociology* 53:723–741.

Hoem, J. M. 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *International Statistical Review* 55:119–152.

Santi, L. L. 1989. Partialling and purging: Equivalencies among log-linear analysis, logit analysis, and partial *CG* purging. *Sociological Methods and Research* 17:376–397.

Schoen, R. 1970. The geometric mean of the age-specific death rates as a summary index of mortality. *Demography* 7:317–324.

Teachman, J. D. 1977. The relationship between Schoen's *del* and a log-linear measure. *Demography* 14:239–241.