DISCUSSION OF PAPER PUBLISHED IN VOLUME LXXV

MINIMUM BIAS WITH GENERALIZED LINEAR MODELS

ROBERT L. BROWN

DISCUSSION BY GARY G. VENTER

1. INTRODUCTION

This paper is a welcome addition to CAS literature on cross-classification ratemaking. This review considers it in the context of other recent work outside the *PCAS*. Despite the title of the paper, the connection with general linear models does not seem to be the primary emphasis of the paper, and some skepticism about this aspect is voiced below.

In his paper, Robert Brown provides additional insight into minimum bias procedures as well as an introduction to generalized linear models. The cross-classification framework is that provided by Bailey [1]. For data with n rows and p columns, the cell in the ith row and jth column has n_{ij} exposure units, e.g., premium, which generate data, e.g., a loss ratio, of $r_{ij} = L_{ij}/n_{ij}$. This is modeled by n row parameters $x_1 \ldots x_n$ and p column parameters $y_1 \ldots y_p$.

Bailey models the ijth cell as an arithmetic function of x_i and y_j; for example, the multiplicative model uses the function $f(x_i,y_j) = x_iy_j$ to estimate future observations of r_{ij}. He then, in effect, applies the principle of balance; he requires that the row and column totals from the model balance to those from the data. In formulas, for each row i:

$$\sum_j n_{ij}r_{ij} = \sum_j n_{ij}f(x_i,y_j),$$

and for each column j:

$$\sum_i n_{ij}r_{ij} = \sum_i n_{ij}f(x_i,y_j).$$

There are $n + p$ such equations, which are enough to solve for all the x's and y's, and as Bailey notes, the solutions can be obtained iteratively. In fact, usually the equations are of the form

$$x_i = g_i(y_1, \ldots, y_p) \text{ and } y_j = h_j(x_1, \ldots, x_n).$$

By starting with reasonable initial values for the x's and y's, the g and h functions can be used to iteratively refine these values until stability is achieved. This is called fixed point iteration, and its convergence properties can be found in numerical analysis texts. Thus an estimation method is specified by giving its system of equations. Brown follows this convention, as does this review. A detail not usually mentioned is that only $n + p - 1$ independent equations are specified in such systems. The result of this is that one of the $n + p$ parameters can be set arbitrarily, e.g., to 1. In a multiplicative model, for example, multiplying the x's by a factor and dividing the y's by the same factor will not affect the cell estimates, so one less parameter is really needed.

As will be discussed more fully below, at least four types of alternatives to Bailey's method have been developed, mostly outside the CAS *Proceedings* or not recognized as relating to the minimum bias procedure. These are: (1) alternatives to the balance principle; (2) more general arithmetic functions; (3) using the arithmetic function as a base, but allowing individual cells to vary from that, based on their own data; and (4) estimating individual cells without postulating an arithmetic relationship between rows and columns. Brown's paper addresses primarily the first area. This review points out some remaining difficulties, and briefly recaps how they have been approached in other studies, using the above alternatives. The connection with general linear models is also discussed.

2. ALTERNATIVES TO BALANCE PRINCIPLE

Brown provides several alternatives to the principle of balance, although he does not give explicit reasons for abandoning it. One such reason may be that it assigns full credibility to each row and column in total, which may not be appropriate. A possible response, however, would be to credibility-adjust the row and column totals before applying the balance principle. Another response might be to find models that

automatically quantify the likely deviations from the cell estimates. However, this could probably be done without discarding the balance principle. Perhaps the basic motivation for abandoning balance is that the principle, while appealing, is not self evident, and thus more fundamental principles should be sought.

In any case, the first alternative Brown presents is to model the numerator of r_{ij}, i.e., L_{ij}, as a random draw from a distribution with mean $n_{ij}f(x_i,y_j)$. Given a distribution and an arithmetic function f, maximum likelihood estimation can be used to solve for all the parameters from the observations. Several distributions are illustrated, and for each a system of $n + p$ equations in $n + p$ unknowns is derived.

For instance, assuming a normal distribution with a multiplicative model, i.e., that the L_{ij} are normally distributed with mean $n_{ij}x_iy_j$ and variance σ^2, gives the following equation for each x_i:

$$x_i \sum_j n_{ij}^2 y_j^2 = \sum_j n_{ij}^2 r_{ij} y_j,$$

and similarly for each y_j. Interestingly, these equations do not involve the σ^2 parameter of the normal distribution.

For the multiplicative model with a Poisson distribution assumption, Brown finds that the system of equations for Bailey's balanced multiplicative model is reproduced. This result was also shown by van Eeghen, Greup, and Nijssen [8]. While it shows that the Poisson distribution satisfies the principle of balance, it does not give much support for using a balanced model, in that the cell data is not usually Poisson distributed. In fact, this might be a reason for dropping the balance requirement, since most distributions will not reproduce it. Thus, the equivalence of the Poisson and Bailey models, rather than supporting their use, suggest that alternatives might be more appropriate.

For the exponential distribution, the following simple equations are produced:

$$x_i = \sum_{j=1}^{p} \frac{r_{ij}}{py_j}, \text{ and } y_j = \sum_{i=1}^{n} \frac{r_{ij}}{nx_i}.$$

This reviewer has found that the same equations hold for the gamma model, which adds a parameter to the exponential. Like the σ^2 of the Normal distribution, this parameter does not enter the equations for the x's and y's.

The most logical distribution for the multiplicative model would probably be the lognormal, because that results when the errors are also due to multiplicative effects. The estimating equations can be derived by using the additive normal model with the logs of the data.

The normal distribution models Brown uses are unusual in that each cell has the same variance σ^2 for L_{ij}. It is hard to see how this could occur from cells with different exposures. For instance, if each exposure unit has the same variance τ^2, then the cell variance would be $n_{ij}\tau^2$ due to the additivity of independent exposures, that is, it would not be constant but would be proportional to n_{ij}. Or, if there are additional gaps between the arithmetic function and the exposure unit means, which average to zero over all cells, i.e., $E(L_{ij}) = x_i y_j + g_{ij}$, with $E(g_{ij}) = 0$, then the variance of this gap, $\mathrm{Var}(g_{ij})$, would be added to $n_{ij}\tau^2$ to give the variance of L_{ij}. Only if the variance of the gap, i.e., error from the arithmetic function assumption, were large compared to the risk variance $n_{ij}\tau^2$ would the constant variance assumption be a reasonable approximation. However, in this case the use of that arithmetic function would be questionable.

It is not difficult to carry out the estimation assuming a variance of $n_{ij}\sigma^2$ rather than σ^2 for L_{ij}. For instance, for the multiplicative model, the x equations become:

$$x_i \sum_j n_{ij} y_j^2 = \sum_j n_{ij} r_{ij} y_j,$$

which are in fact the equations Brown derived for the least squares multiplicative model.

The latter is another alternative Brown presents, namely minimizing the weighted least squares difference between the data and the model. For instance, for the multiplicative model, minimize $\Sigma_{ij} n_{ij}(r_{ij} - x_i y_j)^2$.

This was also advocated by Sant [7], and, under the label "analysis of variance" approach, by Chamberlain [2] and others. The least squares approach has the advantage of not assuming a distributional form, al-

though it does still assume a particular arithmetic function of the parameters x_i and y_j. If the different cell means themselves come from a highly skewed distribution, e.g., display very large percentage differences among the cells, then minimizing the sum of squared errors could allow significant percentage errors for the low mean cells. Thus least squares works reasonably well only for certain types of distributions.

It is generally advisable when doing weighted least squares to use weights which are inversely proportional to the cell variances. The weights Brown uses are thus consistent with r_{ij} having variance inversely proportional to n_{ij}, which seems appropriate. However, the constant variance model for L_{ij} would lead to weights of n_{ij}^2, which, for the least squares model, would produce the system of equations Brown gave for the normal model.

3. GENERALIZATIONS OF ARITHMETIC FUNCTIONS

Although not mentioned in the paper, both of Brown's alternatives, as well as Bailey's original method, can be generalized to use other arithmetic functions of the row and column parameters. For example, the function $x_i y_j + z_j$ has sometimes been used to good effect in class-by-territory ratemaking. This is a combination of additive and multiplicative effects that uses $n + 2p$ parameters. Maximum likelihood estimation with the constant variance normal distribution, for instance, provides a set of $n + 2p$ equations which have the forms:

$$x_i \sum_j n_{ij}^2 y_j^2 = \sum_j n_{ij}^2 (r_{ij} - z_j) y_j,$$

$$y_j \sum_i n_{ij}^2 x_i^2 = \sum_i n_{ij}^2 (r_{ij} - z_j) x_i, \text{ and}$$

$$z_j \sum_i n_{ij}^2 = \sum_i n_{ij}^2 (r_{ij} - x_i y_j).$$

The squares on the exposures would be dropped under the assumption of the variance of L_{ij} proportional to $n_{ij}\sigma^2$. The combined additive-multiplicative function is sometimes appropriate when the high rated

classes in the high rated territories, for example, get too much charge from a multiplicative model and not enough from an additive one. Other arithmetic functions are possible, also, such as $x_i^{.8}y_j^{1.2}$, etc., although the term "arithmetic" might be a misnomer for such functions. There is a wide variety of possibilities of this type which have been largely unexplored. An important exception is Harrington [4], who applies an additive model after applying the Box-Cox transformation to the data. This transformation is $r_{ij}^c = (r_{ij}^c - 1)/c$. This is really a common generalization of both the additive ($c = 1$) and multiplicative models, in that the limit of r_{ij}^c as c goes to zero is $\ln(r_{ij})$, giving an additive log model. By searching for the best fitting c parameter, improved fits can be produced.

4. GLIM DISCUSSION

The GLIM section of Brown's paper is somewhat difficult to follow, but he does recommend background material. Even so, it will not be clear to those without experience with linear models how GLIM as defined might apply to the cross-classification problem. The following example illustrates how this can be done for the multiplicative model with Normal constant variance errors.

If L_{ij} denotes the numerator $n_{ij}r_{ij}$ of r_{ij}, and μ_{ij} its expected value, the Normal density can be put in the GLIM form:

$$f(L_{ij}) = \exp\left[\frac{\mu_{ij}L_{ij} - .5\mu_{ij}^2}{\sigma^2} - \frac{L_{ij}^2}{2\sigma^2} - .5\ln(2\pi\sigma^2)\right].$$

Since the GLIM definition uses variables x and y, let the row and column effects formerly denoted by x and y now be denoted by w and z instead. The observed vector Y to be modeled is the set of L_{ij} all strung out in a single vector, i.e., if $k = (i - 1)p + j$, then $y_k = L_{ij}$. There are $m = np$ of these y_k's. The coefficients β_h to be estimated will be interpreted as the $\ln(w_i)$'s and $\ln(z_j)$'s listed as a single vector (z's after all the w's), followed by a constant term which should turn out to be 1. Thus, there are $q = n + p + 1$ of these β_h's. The explanatory vector, x_h, for $h < q$ is a list of m elements x_{kh} that are all 0's except for 1's which occur when y_k comes from either a row or a column corresponding to β_h. That is for $k = (i - 1)p + j$, $x_{kh} = 1$ only for $h = i$ and $h = n + j$. The last vector x_q, consists of the logs of all the exposures n_{ij}.

With these definitions, let $n_k = \Sigma_{h=1}^{q} x_{kh}\beta_h$. If we defined the x's right, then $n_k = \ln(w_i) + \ln(z_j) + \ln(n_{ij}) = \ln(\mu_{ij}) = \ln(\mu_k)$. Therefore the link function g is the log function. From the form of the density function, it can be seen, in Brown's notation, that the dispersion parameter ϕ is σ^2, $a(\phi) = \phi$, and $c(y,\phi) = -.5[(y^2/\phi) + \ln(2\pi\phi)]$. Also, $\theta_k = \mu_k$, and $b(\theta) = .5\theta^2$.

Thus, this GLIM model is just the original Normal model with constant variance, assuming that maximum likelihood is used to estimate the GLIM parameters. For some reason, the constant variance assumption seems to be inherent in the GLIM models, although it is not necessary when using regular maximum likelihood methods outside of GLIM. For this application, then, GLIM seems to require a fair amount of work to properly arrange the data, with benefits that are unclear.

From the deviances shown in the paper for 12 models, as well as their apparent reliance on density functions, it would appear that deviances cannot be compared across distributions to determine the best fitting model. They probably can be compared to evaluate link functions for one distribution.

5. ALTERNATIVES TO ARITHMETIC FUNCTIONS

Another criticism of minimum bias methods has been the strict reliance on the arithmetic function. Just because data is organized in rows and columns does not imply that there is such an arithmetic relationship. For instance, if loggers have 20% more injuries than cab drivers nationwide, can we expect this will hold true in New York? If office workers have a 90% lower work related accident frequency than workers in general, will this be the case in lower Manhattan? The multiplicative models assume such relationships will hold, and the additive models are based on similar assumptions. In some lines of insurance, it is felt that any arithmetic function of row and column averages can adequately model individual cell results.

At least two methods have been developed in response to this criticism: allowing individual cells to vary from the arithmetic function, or estimating individual cells without using an arithmetic function, e.g., by credibility methods.

The first method was used in the 1981 Massachusetts auto rate hearings, where the calculated relativity was credibility weighted with the cell data r_{ij}. Thus, cells with enough credibility could be based largely upon their own experience. As described in DuMouchel [3], the arithmetic function f was the combined additive-multiplicative function, and the credibility for cell ij was given by:

$$Z_{ij} = \frac{n_{ij}}{n_{ij} + K_j}.$$

Here K_j is the ratio of two variance components s_j^2/t^2, where s_j^2 is the within-cell variance scalar over time, and t^2 is the average variance of true cell means from their calculated relativities. More precisely, for time period t, r_{ijt} has mean μ_{ij} and variance s_j^2/n_{ij}, and μ_{ij} has mean $f(x_i,y_j,z_j)$ and variance t^2. If there are c time periods in the data, s_j^2 is estimated by:

$$\hat{s}_j^2 = \sum_{i,t} n_{it}(r_{ijt} - r_{ij})^2/n(c - 1).$$

DuMouchel gives a somewhat intricate method of estimating t^2. A Bühlmann-Straub type estimation would also be possible. For this, let

$$W = \sum_{i,j,t} n_{ijt}(r_{ijt} - f(x_i,y_j,z_j))^2.$$

Then it can be shown that

$$E(W) = nc \sum_j s_j^2 + t^2 \sum_{i,j,t} n_{ijt}.$$

This means that W is an unbiased estimator of the right hand side, and can thus be used to estimate t^2. That is,

$$t^2 = \left[W - nc \sum_j \hat{s}_j^2\right] \div \sum_{i,j,t} n_{ijt}.$$

If the estimate is negative, it should be set to zero, which would give full credibility to the model and none to the cell data. In the Massachusetts case, DuMouchel found that the combined additive-multiplicative model fit the data very well, so that the credibilities given individual cell data were low.

Other approaches to giving credibility to individual cell variation from the arithmetic function can be used. An example is found in Weisberg, Tomberlin, and Chatterjee [10], who use similar model assumptions to those of DuMouchel, but just with pure additive or multiplicative functions f. They use a different, possibly more general, statisical method to estimate the credibilities.

Another alternative is to incorporate so-called interaction effects, which are essentially additional parameters for specific cells. This was suggested by Chamberlain [2], who showed how to measure the significance of such terms. Jee [6], who summarizes and tests many of the above methods, added all individual cell variables that improved the F statistic at a 15% significance level, and found that this improved the predictive accuracy of the additive, multiplicative, and Box-Cox models.

The credibility only method, not using any arithmetic function, is illustrated by the national relativity approach often used in workers compensation, as described by Harwayne [5]. The indicated percentage change in non-serious pure premiums for the ith class in industry group 1 in state j, for example, is calculated by a variant of the following. Let x_i be the indicated change for class i countrywide, and let y_j be the indicated change for industry group 1 in state j, with r_{ij} the indicated change based on the cell data alone. If the expected number of claims for the ijth cell is at least 300, r_{ij} receives full credibility. Otherwise, the credibility it receives, z_{ij}, is the ratio of expected claims to 300, raised to the two-thirds power. The credibility given to x_i is calculated by essentially the same rule, but it is limited to $(1 - z_{ij}) \div 2$. The balance of the credibility goes to y_j. In formulas, the estimate for the ij cell is:

$$\hat{r}_{ij} = z_{ij}r_{ij} + z_ix_i + z_jy_j,$$

where z_{ij} and z_i are calculated by the rule $(\text{expected claims}/300)^{2/3}$, where the expected claims are for the class in the state or the class countrywide, as appropriate. Although the estimate uses the row and column averages, there is no mathematical relationship postulated between the cell and the totals for the row and column it is in. The x's and y's in the previous models were parameters to be estimated from the data, presumably with some estimation error, while here they are statistics calculated exactly.

The credibilities above may work well in practice, but they could be criticized as being ad hoc. A least squares credibility type approach is given in Venter [9]. The estimate for the gth row and hth column for a future time period is a linear sum of the observations for all the cells available, i.e.,

$$\hat{r}_{gh} = q + \sum_{i,j} z_{ij} r_{ij},$$

where the z's are the weights in the linear function, and q is the constant term. These are found by minimizing the expected squared error $E(\hat{r}_{gh} - r_{gh0})^2$, where r_{gh0} is a future observation of the cell. Thus the credibility estimator is the linear function of all the cell data that minimizes the expected squared error between the estimate and a future observation. This is the standard least squares credibility, applied to the cross-classification problem. As is often the case with credibility, it will probably work better with indicated changes than with pure premium itself.

To express the resulting weights z_{ij} more compactly, introduce the notation $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. The weights are derived as functions of four variance components: u^2 is the variance between row means, v^2 is the variance between column means, w^2 is the variance of a cell mean from row-column additivity, and s^2 is the average relative variance of the cells from their means over time. Also, m is the overall mean of all cells. More precisely, the assumption is:

$$\mathrm{Cov}(r_{ijt} r_{abk}) = u^2 \delta_{ia} + v^2 \delta_{jb} + w^2 \delta_{ia} \delta_{jb} + s^2 \delta_{ia} \delta_{jb} \delta_{tk} \div n_{ijt}.$$

This holds for both additive and multiplicative models, and many others as well. The weights z are expressed in terms of ratios of the variance components. K, J, and L are the ratios of s^2 to u^2, v^2, and w^2, respectively. Using a dot in a subscript to denote summation over that subscript, the weights are:

$q = m(1 - z_{..})$, and

$$\frac{z_{ij}}{w_{ij}} = \frac{\delta_{gi}}{K} + \frac{\delta_{hj}}{J} + \frac{\delta_{gi}\delta_{hj}}{L} - \frac{z_{i.}}{K} - \frac{z_{.j}}{J}, \text{ where } \frac{1}{w_{ij}} = \frac{1}{n_{ij}} + \frac{1}{L}.$$

This requires the summed row and column weights $z_{i.}$ and $z_{.j}$, which can be found from the system of $n + p$ linear equations below, one for each row a and column b:

$$z_{a.}\left[1 + \frac{w_{a.}}{K}\right] = \frac{1}{J}\left[w_{ah} - \sum_j w_{aj}z_{.j}\right] + \delta_{ga}\left[\frac{w_{a.}}{K} + \frac{w_{ah}}{L}\right], \text{ and}$$

$$z_{.b}\left[1 + \frac{w_{.b}}{J}\right] = \frac{1}{K}\left[w_{gb} - \sum_i w_{ib}z_{i.}\right] + \delta_{hb}\left[\frac{w_{.b}}{J} + \frac{w_{gb}}{L}\right].$$

As these equations are linear, they can be solved by matrix methods, although iteration may also work well. The resulting weights differ from credibilities in that they are not necessarily between zero and one, although they are derived in the same manner as credibility weights in the single dimension case.

A method for estimating the required variance components is to compute the four sums of squared differences below:

$$D1 = \sum_{i,j,t} n_{ijt}(r_{ijt} - r_{ij})^2 ,$$

$$D2 = \sum_{i,j,t} n_{ijt}(r_{ijt} - x_i)^2 ,$$

$$D3 = \sum_{i,j,t} n_{ijt}(r_{ijt} - y_j)^2 , \text{ and}$$

$$D4 = \sum_{i,j,t} n_{ijt}(r_{ijt} - \hat{m})^2 .$$

Using their expected values below, these can be used to estimate s^2, J, K, and L.

$$\mathrm{E}(D1) = s^2 np(c - 1),$$

$$\mathrm{E}(D2) = s^2 \left[n(pc - 1) + \frac{1}{J + L} \left(n_{\cdot\cdot} - \sum_{ij} n_{ij}^2 \div n_{i\cdot} \right) \right],$$

$$\mathrm{E}(D3) = s^2 \left[p(nc - 1) + \frac{1}{K + L} \left(n_{\cdot\cdot} - \sum_{ij} n_{ij}^2 \div n_{\cdot j} \right) \right], \text{ and}$$

$$\mathrm{E}(D4) = s^2 \left[npc - 1 + \frac{1}{K} \left(n_{\cdot\cdot} - \sum_{i} n_{i\cdot}^2 \div n_{\cdot\cdot} \right) + \frac{1}{J} \left(n_{\cdot\cdot} - \sum_{j} n_{\cdot j}^2 \div n_{\cdot\cdot} \right) + \frac{1}{L} \left(n_{\cdot\cdot} - \sum_{i,j} n_{ij}^2 \div n_{\cdot\cdot} \right) \right].$$

Brown's paper is a valuable addition to the *Proceedings*, particularly the least squares and maximum likelihood methods. Further empirical studies on how well all of the above models work would be a good area for future research. Both the goodness of fit and accuracy of prediction should be tested, and any distributional assumptions should be reviewed through an analysis of the residuals.

REFERENCES

[1] Bailey, R.A., "Insurance Rates with Minimum Bias," *PCAS* L, 1963.

[2] Chamberlain, C., "Relativity Pricing through Analysis of Variance," 1980 Discussion Paper Program, Casualty Actuarial Society.

[3] DuMouchel, W.H., "The Massachusetts Automobile Insurance Classification Scheme," *The Statistician,* 32, 1983.

[4] Harrington, S.E., "Estimation and Testing for Functional Form in Pure Premium Regression Models," *ASTIN Bulletin,* Vol. 16S, 1986.

[5] Harwayne, F., "Use of National Experience Indications in Workers' Compensation," *PCAS* LXIV, 1977.

[6] Jee, B., "A Comparative Analysis of Alternative Pure Premium Models in the Automobile Risk Classification System," *Journal of Risk and Insurance,* Vol. LVI:3, 1989.

[7] Sant, D.T., "Estimating Expected Losses in Auto Insurance," *Journal of Risk and Insurance,* Vol. XLVII:1, 1980.

[8] van Eeghen, J., Greup, E.K., and Nijssen, J.A., "Rate Making," *Surveys of Actuarial Studies,* Vol. 2, 1983, National-Nederlanden N.V., Rotterdam.

[9] Venter, G.G., "Structured Credibility in Application—Hierarchical, Multi-dimensional and Multivariate Models," *ARCH*:2, 1985.

[10] Weisberg, H.I., Tomberlin, T.J., and Chatterjee, S., "Predicting Insurance Losses under Cross-Classification: A Comparison of Alternative Approaches," *Journal of Business and Economic Statistics,* Vol. 2, 1984.