Journal of Official Statistics, Vol. 18, No. 1, 2002, pp. 45–60

Small Area Estimation via Generalized Linear Models

Alasdair Noble[1], Stephen Haslett[2], and Greg Arnold[1]

Marker (1999) proposed a general linear regression model framework for small area estimation. This framework included most methods that have been used for small area estimation except structure-preserving estimation (SPREE) which was not included because it was non-linear. Marker noted that SPREE can be expressed instead as a log-linear model. This article considers a generalized linear model in the sense of Nelder and Wedderburn (1972). All of the small area estimation methods discussed by Marker, as well as SPREE, are formulated in this more general setting, and a range of further extensions is considered.

Using an explicit log-linear model for SPREE allows an alternative approach to the estimation of the small area estimates. That is: model the census data with a log-linear model, fix the parameters for the main effects and interactions that are held constant and reestimate the other effects using the new margins from the survey data. This method is illustrated using New Zealand unemployment data for nine North Island regions by two sexes and three age groups.

The advantage of using generalized linear models is that the range of models can be extended beyond log-linear models fitted via SPREE. Models that contain any mix of discrete, interval, or continuous variables are possible, as is illustrated by an example.

Key words: SPREE; IGLS; log-linear models; multi level model.

1. Introduction

In recent years small area estimation has emerged as an important area of statistics as organisations try to extract the maximum information from sample survey data. The increasing costs of collecting sample survey data and the decreasing costs of computer power have combined with improved methodology to allow more sophisticated methods to be used to estimate statistics for small geographic areas or small domains of interest. The sample sizes for the survey data used in these small areas are rarely large enough to construct accurate direct estimators of any use. Good reviews of the various methods have been written by Ghosh and Rao (1994) and updating this by Rao (1999).

Marker (1999) reviewed nine methods and organised them into what he called a generalized linear regression framework which we shall refer to simply as linear regression. His linear regression framework includes models with normally distributed responses and various variance-covariance matrices including those of the form $\sigma^2 V$, where V could be a diagonal matrix with unequal entries, or a block diagonal matrix. Symptomatic

[1] Institute of Information Sciences and Technology, College of Sciences, Massey University, Private Bag 11222, Palmerston North, New Zealand. E-mail a.d.noble@massey.ac.nz
[2] Statistics Research and Consulting Centre, Massey University, Private Bag 11222, Palmerston North, New Zealand.
Acknowledgment: The authors would like to thank two referees and the associate editor for their helpful comments on an earlier version of this article.

accounting techniques which simply add or subtract variables as births and deaths to an old census estimate are not stochastic, and so do not strictly fit his framework except as linear models with no error term. The remaining eight techniques, vital rates, symptomatic regression, sample regression, components of variance regression, synthetic estimation, the base unit method, synthetic regression and structure-preserving estimation (SPREE), are discussed in Marker (1999) and the first seven are shown there to fit into the linear regression framework.

This article will extend the linear model to a generalized linear model as defined by Nelder and Wedderburn (1972) and McCullagh and Nelder (1983) by including a link function and allowing random response variables with distributions from the exponential family. As the linear regression models are a subset of this broader classification, the seven linear model methods above can also be classified within this wider framework. It will be shown that SPREE can be included in this extended classification, as a special case, and that the restriction in SPREE that only categorical variables can be used can be relaxed by using the generalized linear model. This allows an alternative approach to the problem of finding small area estimates given appropriate data.

2. Notation

Small area estimates are statistics estimated from sample survey data at a level of disaggregation too small for the survey alone to furnish estimates of adequate precision. Information from some other source is used to lend strength to the survey data. The ''small areas'' may be geographic entities. Alternatively they may be domains within a population, for example 20- to 25-year-old females of a particular ethnic group. The specific estimates required may be these individual domains or some larger aggregation of them. In this article we will use the term ''small area'' in a general sense but where it is clear that we are referring to a domain of interest, or to subgroups within an area, these will be explicitly stated.

Many of the models used below have the structure

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is a vector which is (some function of) a set of observations.

In some circumstances each element of $\mathbf{Y}$ corresponds to each small area or domain. In others there is one element for each subgroup in each small area, or even for each observation in each small area. For random or mixed parameter models, $\mathbf{Y}$ includes expected values of random parameters (see Schall 1991, for example). Generally $\mathbf{Y}$ will be defined to be $n \times 1$. In particular models, n will be defined explicitly.

3. The Generalized Linear Model Approach

Marker (1999) has shown that the majority of the methods for small area estimation mentioned in the introduction can be classified in a linear regression framework.

The model for a linear regression is

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\mathbf{Y}$ is a vector of continuous variables

X variables are assumed known predictor variables

$\boldsymbol{\beta}$ is a vector of regression coefficients

$\boldsymbol{\epsilon}$ is the error vector with mean zero and variance covariance matrix Σ which would be $\sigma^2 I$ for a simple linear regression. In general Σ has to be a positive definite symmetric matrix but in practice it is most likely to be block diagonal.

Symptomatic accounting techniques fit the model by omitting the error term and using a known $\boldsymbol{\beta}$ vector. The values of $\boldsymbol{\beta}$ are simply 1 or -1 depending on whether the variable is added or subtracted. The other methods shown to fit Model (3.1) are

- Vital rates
- Symptomatic regression
- Sample regression
- Components of variance regression
- Synthetic estimation
- Composite estimators
- Empirical Bayes, superpopulation, and other Bayesian approaches using linear models.

3.1. Structure-preserving estimates

Structure-preserving estimates, via SPREE (Purcell and Kish 1980), update the cells in a contingency table formed from previous census data so that they sum to the margins found from new sample survey data. It uses the iterative proportional fitting (IPF) algorithm (Deming and Stephan 1940), which is essentially raking, to fit the new margins. This method does not fit the structure of the linear regression model, as noted by Marker (1999). The contingency table can be written as a log-linear model but it cannot be expressed directly as a linear regression model.

The concept which underpins SPREE is that of a log-linear model with all parameters estimated from the census data. The new margins will change some of those parameters. The assumption is that the remaining parameters (usually higher order interactions) stay the same, and thus a new model, with higher order effects than can be fitted using survey data alone, is found to estimate the small area counts. There are very clear data requirements which have to be fulfilled for this approach to work. First, census data are required for the variable under consideration, or a variable to which it is closely related. Second, sample survey data are needed for the variable of interest with values for the same explanatory variables and using the same categorisation as used in the census. In the traditional SPREE method the census information will be from a recent census, and the variable of interest will be the same as the one recorded in that census for a more recent period. Sample survey data need to be collected not only for the same variable of interest but also for the same categorical explanatory variables. An extension to this structure using related rather than the same variables is given in Green, Haslett, and Zingel (1998). The essential requirement is that the census and survey variables have the same cross-product ratios for all (interaction) terms not refitted using the survey data. Green et al. consider two differently defined measures of unemployment, and allow for measurement of survey errors in both variables.

3.2. The Generalized linear model

The wider class of models which includes the linear regression models discussed in the first part of this section is the generalized linear model (GLM). As the linear regression

models are a subset of this wider class, those methods discussed earlier can also be included in this generalized linear model framework. We will show that SPREE is implicitly a fitting procedure for a log-linear model, and so can also be included in this wider class of models. We also will show how the general concept underlying SPREE can be applied to other situations such as occur when explanatory variables are not categorical.

The generalized linear model as defined by McCullagh and Nelder (1983) is briefly introduced in this section. The model is

$$g(\mathbf{Y}_0) = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

where:

- $g(\,)$ is a possibly composite (Thompson and Baker 1981), monotonic differentiable, link function, which allows a wide variety of functions including the identity, log, logit, probit, powers, etc. (Nelder and Wedderburn 1972). It is applied to each element of $\mathbf{Y}_0$ and $g(\mathbf{Y}_0)$ like $\mathbf{Y}_0$ is $(n \times 1)$
- $\mathbf{Y}_0$ is a vector $(n \times 1)$ of values for the response variable, possibly supplemented by a set of random parameter equations. These will have a distribution which may be any member of the exponential family of distributions. (Note $\mathbf{Y}_0 = \mathbf{Y}$ for the linear regression model)
- X is a known $(n \times p)$ matrix of values each column of which is either a categorical or continuous explanatory variable
- $\boldsymbol{\beta}$ is a vector $(p \times 1)$ of parameters to be estimated
- $\boldsymbol{\epsilon}$ is a (transformed) vector $(n \times 1)$ of errors, or residuals.

The model as formulated above is linear in the parameters, $\boldsymbol{\beta}$, and the variance-covariance matrix of $\boldsymbol{\epsilon}$ requires iterative fitting but, as for linear regression, powers or other functions of the explanatory variables are allowed in the X matrix. Thus although the model (3.2) describes a non-linear relationship between $\mathbf{Y}_0$ and $\boldsymbol{\beta}$, the least squares or maximum likelihood solution to (3.2) can be achieved by iterated generalized least squares, where at each iteration a generalized least squares algorithm is applied to a linear (rather than non-linear) model. The underlying variance-covariance metric changes at each iteration and depends not only on $\mathrm{Var}(\mathbf{Y}_0)$ but also on the partial derivatives of $X\boldsymbol{\beta}$ with respect to $\boldsymbol{\mu} = E(\mathbf{Y}_0)$. The estimation procedure iterates until convergence between estimating the parameters $\boldsymbol{\beta}$ (and hence $\boldsymbol{\mu}$) and estimating Σ. It has been shown (del Pino 1989) that this results in maximum likelihood estimates of the parameters both for a generalized linear model and for the wider class of linearisable non-linear models and that the dependent variable in the generalized least squares algorithm both at each iteration and at convergence is an affine function of $\mathbf{Y}_0$. For ease of notation we can thus, in general, set $\mathbf{Y} = g(\mathbf{Y}_0)$ so that Equation (3.2) can be written as

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.3}$$

provided we understand that the solution needs to be iterative because the variance-covariance structure of the error process changes at each iteration.

The example which we will consider later is a nine by three by two contingency table of count data with a Poisson distribution for the individual cells. It has been shown (for

example in Bishop, Fienberg, and Holland (1975), Chapter 2) that such a table can be modeled by a log-linear model and hence will fit the GLM framework. The log-linear model is equivalent to a SPREE model

$$\log(\mathbf{m}) = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.4}$$

where $\mathbf{m}$ is a vector of counts with one count for each subgroup within each small area a, i.e., for each of the n cells in the table, so that $n = \text{AS}$ where A is the number of small areas and S is the number of subgroups
X is the $(n \times p)$ design matrix
$\boldsymbol{\beta}$ is the $(p \times 1)$ vector of parameters
$\boldsymbol{\epsilon}$ is a random error, as defined in Equations (3.2) or (3.3).

In SPREE as used in small area estimation we are generally fitting a saturated model, so there is no error term.

More generally a generalized linear model will be fitted to census data and then some of the lower order parameters in the model will be adjusted in line with sample survey data. We can partition the design matrix and the vector of parameters into two parts, the parameters that are estimated by the survey data and those that are not. Hence the two models can be written as:

$$\mathbf{Y}_c = X_1\boldsymbol{\beta}_{1c} + X_2\boldsymbol{\beta}_{2c} + \boldsymbol{\epsilon}_c \qquad \text{for the census data} \tag{3.5}$$

and

$$\mathbf{Y}_s = X_1\boldsymbol{\beta}_{1s} + X_2\boldsymbol{\beta}_{2s} + \boldsymbol{\epsilon}_s \qquad \text{for the survey data} \tag{3.6}$$

However, the survey data are not sufficiently detailed to estimate $X_2\boldsymbol{\beta}_{2s}$ so in SPREE we assume that

$$X_2\boldsymbol{\beta}_{2c} = X_2\boldsymbol{\beta}_{2s} \tag{3.7}$$

Finally the cell estimates and hence small area estimates are predicted from the new model. An explicit example is given in Section 5.

Green, Haslett, and Zingel (1998) note that the data from the census and survey do not have to be for the same variable, and they suggest high correlation can be a useful but not sufficient indicator of equal higher order interactions. The same comment applies even if census and survey data are for the same variable from different time points.

From Equations (3.5), (3.6), and (3.7) it can be seen that while high correlation may be a reasonable criterion to use the actual requirement is Equation (3.7), and this could be true even if $\mathbf{Y}_c$ and $\mathbf{Y}_s$ are not highly correlated. If $\mathbf{Y}_c$ and $\mathbf{Y}_s$ are correlated then it is still possible for Equation (3.7) to not hold if correlation between $X_1\boldsymbol{\beta}_{1c}$ and $X_1\boldsymbol{\beta}_{1s}$ provides the only basis of the correlation between $\mathbf{Y}_c$ and $\mathbf{Y}_s$. On the other hand if the correlation of $X_1\boldsymbol{\beta}_{1c}$ and $X_1\boldsymbol{\beta}_{1s}$ is lower than of $\mathbf{Y}_c$ and $\mathbf{Y}_s$ then there must also be correlation between $X_2\boldsymbol{\beta}_{2c}$ and $X_2\boldsymbol{\beta}_{2s}$ and the algorithm should produce useful estimates. Essentially the same issue arises in all other methods of fitting small area estimates. In SPREE it is hidden because the log-linear model is not explicit and the fitting is carried out using iterative proportional fitting rather than by fitting an explicit log-linear model.

One benefit, then, of the model using Equation (3.2) is that assumptions and model are considerably more explicit.

4. An Algorithm for SPREE Using the GLM

Traditionally SPREE models have been fitted using the IPF algorithm to fit a new margin, or margins, to a census table. The census table data have generally been for the survey variable but at an earlier time and the new margins have typically come from a sample survey that is more recent than the census. All of those census-based main effects and interactions that are not changed by the new sample margins must be sufficiently accurate (unless replicate census data are available) for them to be assumed error free.

An extension discussed earlier is that the census table is not from an earlier census for the same variable, but may be a previous or even more current census of a variable strongly enough related to the variable of interest. The actual requirement is that it can be assumed that in the generalized linear model (that is equivalent to the SPREE model), the census and sample have the same non-main effects and interactions for all effects that are not going to be re-estimated using the survey data. A further extension involves using balanced repeated replicates for the sample data, and a range of census estimates under a superpopulation model as projections, to provide sound variance estimates both for parameters and small area estimates. The joint superpopulation-design variance for individual cell estimates (or sums of cell estimates) can be decomposed into two conditional variance components. The first of these can be estimated from the variation with the margins fixed at the survey estimates and initial cell entries determined from a sequence of census values for the cells. The second can be estimated by fixing the cell values at the appropriate census figures and varying the margins based on balanced repeated replicates. A more complete exposition can be found in Green et al. (1998). As an alternative to balanced repeated replicates, jacknife estimates could be used for estimating variance due to sampling error in the survey margins.

We now describe heuristically a simple alternative method for fitting the traditional SPREE model.

A – Fit a log-linear model to the census data.
B – Decide on which effects/interactions will stay the same.
C – Constrain those parameters to be the same as for the model for the census data by estimating the offset, $X_2\boldsymbol{\beta}_{2c}$, from Equation (3.6) for each cell in the table and subtracting that from the data values.
D – Refit the model using the new margins from the survey. These may need to be expressed in a table of the same order as the census data with only the new effects and interactions present.
E – Predict the new cell values from the new model.

Details of how to apply this method in various statistical packages are given in the Appendix.

5. Example

In New Zealand, Statistics New Zealand carries out a Household Labour Force Survey (HLFS) quarterly and the Department of Work and Income (which was previously called the Department of Labour, DoL, and Work and Income New Zealand, WINZ) collects monthly data on registered unemployed. The New Zealand population was 3,618,300 at 5 March 1996 and the sample size for the HLFS is approximately 19,000 per quarter.

The definitions of unemployed used by the two organisations differ. The two unemployment variables are highly correlated and this allows the association structure generated by the Department of Work and Income data to be used for a SPREE approach to estimating unemployment rates under the Statistics New Zealand definition (Green et al. 1998). Statistics New Zealand uses the International Labour Organisation definition of unemployment. The Department of Work and Income has a more restricted administrative definition of unemployment. Statistics New Zealand is interested in producing estimates of ILO unemployment at Regional (16 nationally) and even Territorial Authority (74) level. The sample sizes in the HLFS are too small for a number of Territorial Authorities for Statistics New Zealand to publish direct estimates. Even some Regional Authorities have larger errors than Statistics New Zealand would like. Thus the census data are provided by the Department of Work and Income and the sample survey data are the HLFS from Statistics New Zealand.

For the example below, the data detailed here are restricted to the North Island of New Zealand which includes nine of the sixteen national Regional Authorities. The unemployment counts are divided into the three age groups 15 to 24, 25 to 49 and 50 or over, and the two sexes male and female.

The design matrix for the model is written with a constant term; sex is +1 for males and -1 for females; for ages 15 to 24 Age 1 is +1 and Age 2 is 0, for ages 25 to 49 Age 1 is 0 and Age 2 is +1, and for ages 50 and over both Age 1 and Age 2 are -1. Similarly, for the nine regions, so that the parameter estimate for the last (and most southern) region, Wellington, is the negative of the sum of the parameter estimates of the other eight regions. The interaction terms in the design matrix, which are all categorical, are simply products of the appropriate main effects.

Letting the subscript '1' denote sex, '2' and '3' denote age groups and 'a' $= 1, 2, \ldots A$ denote area, the model fitted to the census data is:

$$\log(m_{123a}) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_1^{(a)}\beta_1^{(a)} + \ldots x_{(A-1)}^{(a)}\beta_{(A-1)}^{(a)} + \text{interaction terms}$$

where β_0 is an intercept term. In this example the parameters denoted as $\beta_1^{(a)}$ to $\beta_{(A-1)}^{(a)}$ and those denoted as interaction terms are not re-estimated from the survey data. Only β_0, the intercept, and β_1, β_2, and β_3, which correspond to the margins for sex and the two age groups, are to be re-estimated.

The data from the Department of Work and Income, in Table 5.1, are used to form the census association structure for a saturated log-linear model and the fit for that model is shown below. The data are for the quarter ending December 1996.

The HLFS survey data are contained in the marginal entries in Table 5.2.

Applying the algorithm described earlier:

Step A Estimate the log-linear model for the census data. The estimated parameters are shown in Table 5.3.

Step B Select those parameters which remain the same, and those which will be changed. These parameters have been selected because the new margins only give information about the two age effects and the sex effect as well as a different population size. Hence these effects and the constant term will change in the refitted model. There are four independent new pieces of information and so four parameters will change in the model. (Note that the survey data might also have been used to refit

Table 5.1. Department of Work and Income census data: Counts of unemployed for the quarter ending December 1996

Region		Age groups 15 to 24	25 to 49	50 and over
Northland	Male	1,794.4	4,386.6	632.2
	Female	1,060.3	2,555.0	465.0
Auckland	Male	7,233.1	16,542.7	2,324.7
	Female	5,040.4	8,104.5	1,796.9
Waikato	Male	2,924.9	4,955.0	792.0
	Female	2,047.8	2,849.2	654.0
Bay of Plenty	Male	2,552.5	5,169.5	755.0
	Female	1,862.4	1,681.2	601.0
Gisborne	Male	680.0	1,511.0	203.0
	Female	438.0	1,011.0	120.0
Hawkes Bay	Male	1,819.9	3,470.9	550.0
	Female	1,137.0	1,109.1	377.0
Taranaki	Male	1,295.9	1,982.1	332.2
	Female	896.0	1,530.7	277.1
Manawatu-Wanganui	Male	2,552.7	4,287.6	752.0
	Female	1,693.3	2,220.2	538.5
Wellington	Male	4,073.4	6,911.3	1,145.7
	Female	2,555.7	2,395.0	753.3

the Sex*Age 1 and Sex*Age 2 effects, if sufficiently accurate survey estimates were available for these sub-populations.)

Step C Constrain parameters that stay constant to be the same as they were under the census-based log-linear model.

Step D Refit the model using the new sample-based marginals from the HLFS. To do this we need to build a new table from marginal data using only the effects that will change, i.e., in this case construct the independent 3×2 table for the new age and sex margins as in Table 5.2.

The elements in Table 5.2 are then divided by nine to give preliminary estimates for the nine regions. Each of these values is then repeated nine times to generate the new data. This new data includes information from the survey to estimate the constant term and the sex and two age effects. The region effect and all interaction terms are estimated from the census data.

Refit the model to the new data with the necessary parameters constrained. The parameter estimates are in Table 5.3. The $\boldsymbol{\beta}_1$, from Equation (3.5), are given in the top section. The values from the census data are discarded, and replaced by the new estimates from the survey data. The next section of the table are the coefficients which remain constant, $\boldsymbol{\beta}_2$ from Equation (3.5), from Step A.

Table 5.2. Independent 3×2 table of unemployed constructed from new margins from the HLFS survey data

	Age groups 15 to 24	25 to 49	50 and over	New margin for sex
Female	19,078	22,503	5,534	47,116
Male	25,156	29,671	7,296	62,125
New margin for age groups	44,235	52,175	12,831	109,241

Table 5.3. Table of coefficients for the full model with categorical variables for region, sex, and two age groups

β_1 Effects		Constant	Sex	Age 1	Age 2		
Estimated from census data		7.300	0.240	0.223	0.723		
Estimated from survey data		7.226	0.105	0.374	0.505		
β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.							
Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
−0.086	1.262	0.238	−0.308	1.142	0.077	−0.580	0.097
Two-way interactions		Sex*Age 1	−0.034	Sex*Age 2	0.118		
Two-way interactions, Regions with Sex							
−0.012	−0.018	−0.057	0.037	−0.013	0.091	−0.106	−0.007
Two-way interactions, Regions with Age 1							
−0.207	−0.079	0.042	0.068	−0.078	0.057	0.040	0.040
Two-way interactions, Regions with Age 2							
0.179	0.072	−0.029	−0.130	0.239	−0.133	0.020	−0.065
Three-way interactions, Regions with Sex and Age 1							
0.068	−0.007	0.029	−0.086	0.026	−0.062	0.083	0.006
Three-way interactions, Regions with Sex and Age 2							
−0.077	0.017	−0.025	0.166	−0.145	0.121	−0.123	−0.023

Table 5.4. New estimates of counts for unemployment in each sex by age category for the nine regions

Region		Age groups		
		15 to 24	25 to 49	Over 50
Northland	Male	1,693.7	2,861.9	547.8
	Female	1,312.3	2,185.7	528.4
Auckland	Male	6,827.2	10,792.7	2,014.4
	Female	6,238.3	6,933.1	2,041.7
Waikato	Male	2,760.8	3,232.7	686.3
	Female	2,534.5	2,437.4	743.1
Bay of Plenty	Male	2,409.2	3,372.6	654.2
	Female	2,305.0	1,438.2	682.9
Gisborne	Male	641.8	985.8	175.9
	Female	542.1	864.8	136.4
Hawkes Bay	Male	1,717.8	2,264.4	476.6
	Female	1,407.2	948.8	428.3
Taranaki	Male	1,223.2	1,293.1	287.9
	Female	1,108.9	1,309.4	314.9
Manawatu-Wanganui	Male	2,409.4	2,797.3	651.6
	Female	2,095.7	1,899.3	611.9
Wellington	Male	3,844.8	4,509.0	992.8
	Female	3,163.1	2,048.8	855.9

Step E Predict the new values for the table from the revised model parameters from Step D (Table 5.4).

An iterative proportional fit for the same original table and the new margins yields the same result but without estimating the log-linear model parameters.

The advantage of the explicit use of a generalized linear model is that the algorithm is not then restricted to categorical data. Continuous variables may be incorporated and, via a careful respecification of the model detailed for example in Goldstein (1995), random effects as well as fixed effects may be fitted. The next section will demonstrate this greater flexibility.

6. A Quadratic Model for Age

The most detailed data that Statistics New Zealand collects is in five-yearly categories, which makes 11 age categories for each region by sex. With the new algorithm we are able to model this in a more parsimonious way using a quadratic term for the age groups. In general the counts follow a curve which may be approximated by a quadratic. Additional polynomial or other terms could be included but the purpose of this article is not to find the best model, but simply to demonstrate that the new algorithm opens up a much wider range of possible small area models than available in regression and SPREE.

Using the same algorithm as in Section 4 and the data in Tables 6.1 and 6.2 we estimate the parameters in the model, in Table 6.3, and hence the predicted values in Table 6.4 below.

Table 6.1. The new unemployment marginals for the five-yearly age groups. The magin for sex stays as before

Age	17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5
Margin	24,760	19,475	14,254	11,325	9,965	9,067	7,564	5,214	3,654	3,269	694

Table 6.2. Counts for unemployment from Department of Work and Income data in five-yearly intervals Census data in five-yearly age groups. Column headings are the centre of the age range for that group

Region		17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5
Northland	M	882	912	968	954	918	869	678	402	180	50	0
	F	525	535	540	535	511	485	484	235	178	52	0
Auckland	M	3603	3630	3690	3666	3476	3032	2679	1520	631	174	0
	F	2572	2468	2148	1863	1598	1386	1110	789	523	396	89
Waikato	M	1523	1402	1298	1156	1008	893	600	459	263	68	2
	F	1096	952	774	683	551	476	365	307	215	132	0
Bay of Plenty	M	1302	1251	1219	1176	1077	927	771	451	231	73	0
	F	953	909	534	382	270	261	234	220	186	143	52
Gisborne	M	347	333	328	327	312	298	246	121	69	13	0
	F	221	217	220	224	218	187	162	64	47	9	0
Hawkes Bay	M	924	896	843	791	703	601	533	306	195	49	0
	F	675	462	376	249	213	143	128	123	106	83	65
Taranaki	M	690	606	554	485	397	326	220	176	116	40	0
	F	470	426	406	355	308	256	206	142	93	39	3
Manawatu-	M	1357	1196	1073	986	861	704	664	362	225	165	0
Wanganui	F	911	782	683	555	427	316	239	201	158	100	80
Wellington	M	2263	1810	1664	1513	1380	1268	1086	732	321	93	0
	F	1413	1143	796	546	405	365	283	251	204	189	109

Table 6.3. Table of coefficients for the full model with categorical variables for region and sex, and linear and quadratic terms for Age

β_1 Effects		Constant	Sex	Age	Age squared		
Estimated from census data		6.20135	−0.41529	0.05591	−0.00131		
Estimated from survey data		7.50247	−0.54423	−0.02632	−0.00029		
β_2 Effects and interactions that are estimated from the census data and used in the final model. Main effects for Regions.							
Northland	Auckland	Waikato	Bay of Plenty	Gisborne	Hawkes Bay	Taranaki	Manawatu-Wanganui
−1.54162	0.72169	0.41673	0.69281	−2.42406	0.53471	−0.53682	0.68335
Two-way interactions		Sex by Age	0.04182	Sex by Age squared	−0.00055		
Two-way interactions, Regions with Sex							
0.53136	−0.16852	0.17345	−0.76242	0.69340	−0.72056	0.69263	0.04976
Two-way interactions, Regions with Age							
0.07943	0.02992	−0.00755	−0.03641	0.07922	−0.05184	0.00277	−0.03603
Two-way interactions, Regions with Age squared							
−0.00091	−0.00035	0.00006	0.00046	−0.00100	0.00066	−0.00009	0.00045
Three-way interactions, Regions with Sex and Age							
−0.03027	0.00867	−0.01273	0.04885	−0.04633	0.04922	−0.04676	−0.00479
Three-way interactions, Regions with Sex and Age squared							
0.00035	−0.00011	0.00015	−0.00062	0.00061	−0.00062	0.00057	0.00008

Table 6.4. Predictions for unemployment in five-yearly intervals
Predictions in five-yearly age groups. Column headings are the centre of the age range for that group

Region		17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5
Northland	M	775	810	789	718	608	481	355	244	157	94	52
	F	598	602	576	524	454	374	293	218	154	108	66
Auckland	M	3177	3219	3057	2723	2274	1781	1308	900	581	352	200
	F	3368	2675	2126	1691	1347	1074	857	684	547	437	350
Waikato	M	1439	1239	1033	835	654	496	365	260	179	120	78
	F	1393	1052	800	614	475	371	292	232	185	150	122
Bay of Plenty	M	1182	1115	1001	855	696	539	397	278	186	118	71
	F	1371	830	537	371	274	216	182	164	158	162	179
Gisborne	M	299	299	281	248	206	161	119	82	54	33	19
	F	149	254	241	214	177	138	100	68	43	25	14
Hawkes Bay	M	858	781	683	575	464	361	269	193	133	88	56
	F	906	528	332	225	165	131	111	103	102	110	127
Taranaki	M	669	540	429	334	255	192	142	103	73	51	35
	F	583	490	404	326	258	200	152	114	83	59	42
Manawatu-	M	1281	1062	868	698	553	432	332	251	187	138	100
Wanganui	F	1256	867	618	455	346	271	219	183	158	141	129
Wellington	M	1976	1694	1415	1151	912	704	529	388	277	192	130
	F	1976	1166	738	501	365	286	240	216	208	216	240

From Table 6.1 the independent table can be generated (cf. Table 5.2) to use as data for Step D.

We can compare these predictions with the predictions for the saturated model either from SPREE using iterative proportional fitting, or via the generalized linear model algorithm outlined in Section 5.

7. Discussion

We have shown that all of the traditional methods of small area estimation can be simply expressed as members of the class of generalized linear models, and are thus closely related. This modeling approach provides explicit parameter estimates, and explicit model specification which ensures that the model assumptions are more transparent and amenable to checking.

We have introduced an alternative algorithmic approach to the traditional iterative proportional fit for SPREE models to reestimate cell values in a contingency table when new margins from sample survey data are available. This new approach has been to model the contingency table explicitly as a log-linear model, estimate all of the parameters for census data, and then adjust the parameters which can be accurately estimated from the new survey data via revised parameter estimates from fitting the new survey margins.

There are two distinct advantages to this approach. The first is that the log-linear model used to model the contingency table and the assumptions of this model are explicitly stated. The approach also allows extensions into problems which the iterative proportional fit cannot solve, since the new algorithm as outlined can be applied to any generalized linear model, not only log-linear models. An iterative proportional fit solution is only possible for categorical variables which can be described by a contingency table, while the generalized linear model can be used with continuous variables. With this new approach the general concept underlying SPREE can be extended. Any census data which can be modeled by a generalized linear model, and for which there exists survey data to adjust some of the parameter estimates initially determined from the census data, is amenable to this method.

Although variance estimation has not been discussed in this article in detail, by using replication methods such as balanced repeated replicates or the jacknife to provide a set of alternative survey margins, and a range of census values to allow for census variation over time, it is also possible to estimate parameter variances in the generalized linear models detailed in the earlier part of the article. (See Green et al. 1998 for further detail.) By extension variance estimates for estimated means are also possible. Such variance estimates by allowing for complexities in the survey design and for uncertainty in the census values are much better than variances and standard errors available in standard statistical packages from fitting the generalized linear model to a single set of census data and overall survey estimates as new margins.

Appendix

Computing in M1Win, Splus and SAS

M1Win

Set-up data:
- Census data
- Design matrix
- Dummy variables for constants etc. These are required to set up a hierarchical model with no variation at the higher levels. A constant is used so that all of the data is nested in the same higher level, since M1Win generally fits multilevel models.
- New data from the survey repeated the appropriate number of times.

Model data – Run IGLS estimation using the census data.

Enter constraints – In the constraints window. Enter 1 by the variable and the value of the coefficient for that variable at the bottom of the column.

Re-estimate new model

– Using the new data and the constraints.

Predict the new cells – Using the prediction window.

Splus

The data can be modeled in Splus using the glm function and a poisson distribution. Splus includes the constant term by default so the design matrix does not need the constant term. The census data is easily modeled:

$$\text{census} \leftarrow \text{glm}\ (y \sim X,\ \text{family} = \text{poisson})$$

This gives the coefficients for those terms which will remain the same. They are then used to produce an ''offset'' for each y value by multiplying those coefficients by the columns of the design matrix for those effects and interactions. Using the sample data, a design matrix for the effects which will change and the offset, the new coefficients are estimated.

$$\text{sample} \leftarrow \text{glm}\ (\text{sampledata} \sim Xa + \text{offset(offsets)},\ \text{family} = \text{poisson})$$

The new coefficients are estimated and predicted cell counts can be calculated from the full model.

SAS

SAS can be used in a similar way to Splus except it will not allow data to be negative or non-integer. This causes a small rounding problem with the sample data as the predicted cell counts are not whole numbers and the census data may also require some interpolation if regional boundaries for the census and sample data do not match exactly.

8. References

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis. MIT Press.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment to a Sampled Frequency Table when the Expected Marginal Totals Are Known. Annals of Mathematical Statistics, 11, 427–444.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Sciences, 9, 55–93.

Goldstein, H. (1995). Multilevel Statistical Models. Kendall's Library of Statistics 3, Edward Arnold (www.arnoldpublishers.com/support/goldstein.htm).

Green, A., Haslett, S.J., and Zingel, C. (1998). Small Area Estimation Given Regular Updates of Census Auxiliary Variables. Proceedings of the New Techniques and Technologies for Statistics Conference. Eurostat. November, Sorrento, Italy, 206–211.

Marker, D.A. (1999). Organization of Small Area Estimators Using a Generalized Linear Regression Framework. Journal of Official Statistics, 15, 1–24.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society, Series A, 135, 370–384.

Purcell, N.J. and Kish, L. (1980). Postcensal Estimates for Local Areas (or Domains). International Statistical Review, 48, 3–18.

del Pino, G. (1989). The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms. Statistical Science, 4, 394–408.

Rao, J.N.K. (1999). Some Recent Advances in Model-Based Small Area Estimation. Survey Methodology, 25, 175–186.

Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. Biometrika, 78, 719–727.

Thompson, R. and Baker, R.J. (1981). Composite Link Functions in Generalized Linear Models. Applied Statistics, 30, 125–131.

Received December 2000
Revised January 2002