American Journal of Epidemiology
Vol. 164, No. 3
DOI: 10.1093/aje/kwj171
Advance Access publication May 3, 2006

Practice of Epidemiology

Modeling the Relation between Socioeconomic Status and Mortality in a Mixture of Majority and Minority Ethnic Groups

Jim Young[1], Patrick Graham[1], and Tony Blakely[2]

[1] Department of Public Health and General Practice, Christchurch School of Medicine and Health Sciences, University of Otago, Christchurch, New Zealand.
[2] Department of Public Health, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand.

Received for publication June 20, 2005; accepted for publication February 3, 2006.

Ethnic variation in mortality and whether this variation can be explained by socioeconomic status are of substantive interest to social epidemiologists. The authors consider the analysis of mortality data for a mixture of majority and minority ethnic groups. Such data are likely to be coarsely cross-classified by age and socioeconomic status and yet, even then, in some cells of this cross-classification the observed mortality rate will be an imprecise estimate of the underlying rate. The authors illustrate conventional and Bayesian approaches to analysis with data from the 1996 census used by the New Zealand Census-Mortality Study. A conventional approach is exploratory data analysis first followed by Poisson regression. The authors use spline smoothing within a generalized additive model framework as an exploratory data analysis, following a strategy of adding just enough model structure to gain a sensible picture. A Bayesian approach is modeling first and then a description of posterior estimates using exploratory data analysis techniques. The authors use hierarchical Poisson regression and then illustrate their posterior estimates of the mortality rate using the same spline smoothing as before. The advantage of the hierarchical Bayesian approach is that it assesses uncertainty about a Poisson regression model proposed a priori; the conventional approach assumes that the fitted Poisson regression model is correct. All analyses use software that is available at no cost.

ethnic groups; hierarchical model; mortality; nonparametric regression; Poisson regression; smoothing; socioeconomic factors; spline

Abbreviations: CI, credible interval; nMnPI, non-Maori, non-Pacific Island.

Ethnic variation in mortality and whether this variation can be explained by socioeconomic status are of substantive interest to social epidemiologists (1, 2). To develop more realistic models for ethnic and socioeconomic variation in mortality, Kaufman et al. (3) recommended a nonparametric exploratory data analysis. They used kernel smoothing to create a contour plot of the observed mortality rate across dimensions of age and income for each combination of gender and ethnicity. Kaufman et al. imposed as few assumptions as possible so that the data speak for themselves. For this reason, they considered mortality rates for only the main ethnic groups in the United States (Blacks and Whites), even though there were 27,239 Hispanics in the nationwide survey on which their study was based.

It is not obvious how to apply this strategy to mortality data for a mixture of majority and minority ethnic groups. Such data are likely to be coarsely cross-classified, either to ensure confidentiality when releasing official statistics or where ordinal measures of socioeconomic status are used with few categories. Even then, the observed mortality rate

Correspondence to Dr. Jim Young, Vital Statistics Limited, 85B Barrington Street, 8002 Christchurch, New Zealand (e-mail: kreiliger@actrix.co.nz).

may be an imprecise estimate of the underlying rate because of the relatively small number of deaths in some cells of this cross-classification. In addition, conventional statistical inference—the process of generalizing from these data by point or interval estimate—is hard to justify where data are collected without either a randomly assigned intervention or random sampling (4). Without randomization, statistical inference in an observational study has to rely on subjective judgments of exchangeability (5), and then it is logical to take a Bayesian approach to statistical inference.

We consider conventional and Bayesian approaches to modeling mortality data for a mixture of majority and minority ethnic groups. We describe an example where the observed mortality rates for a major ethnic group and two minorities are cross-classified by gender, age, and highest educational qualification. We first illustrate a conventional approach: exploratory data analysis using generalized additive models prior to conventional Poisson regression. We then consider this example from a Bayesian perspective, fitting a hierarchical Poisson regression model and using generalized additive models to illustrate our posterior estimates of the mortality rate. We finish by comparing the two approaches and giving details of the software used in our analyses.

THE NEW ZEALAND CENSUS-MORTALITY STUDY

In this study, New Zealand census data collected every 5 years are anonymously and probabilistically linked to persons who died within the 3 years following each census (6). We use data from the 1996 census for those aged 25–74 years, with 78 percent of subsequent mortality records linked to a census record (7). Mortality rates and person-years at risk are shown for a 240-cell cross-classification of three ethnic groups by gender, age in 5-year categories, and highest educational qualification in four ordered categories (Web appendix A). (This information is described in the first of two supplementary appendices; each is referred to as ''Web appendix'' in the text and is posted on the website of the *Journal* (http://aje.oxfordjournals.org/).) The three ethnic groups are two minorities, Maori (the indigenous people of New Zealand) and Pacific Island (those of Pacific Island descent), and the non-Maori, non-Pacific Island (nMnPI) majority (mostly those of European descent). The ethnic group was categorized as Maori if this was given as one of up to three responses to the census question on ethnicity; otherwise, it was categorized as Pacific Island if this was given as one of the three responses; otherwise, it was categorized as nMnPI (8).

The mortality rate y_i in the ith cell of this cross-classification is estimated from the n_j persons in the cell as:

$$y_i = \frac{\sum_{j=1}^{n_j} w_{ij} z_{ij}}{\sum_{j=1}^{n_j} w_{ij} e_{ij}}, \qquad (1)$$

where z_{ij} is one if the jth person dies in the 3 years after the census and zero otherwise; e_{ij} is the number of years between the census and death for those that die and three otherwise; and w_{ij} is the person's linkage weight (the inverse of the probability of linkage). Not all mortality records can be linked back to a census record, and so mortality and person-years at risk are weighted to account for linkage bias (9). The denominator of equation 1 is a weighted estimate of the person-years at risk e_i.

To measure socioeconomic status, we assume the following order among the four categories of highest educational qualification: none, secondary school, trade or vocation, and tertiary. Thus ordered, the highest qualification is then transformed into a ridit score (10): Within each 5-year age category, the educational score associated with a given qualification is the midpoint of the percentages covered by that qualification in the cumulative distribution of qualifications. A ridit transformation is appropriate because more people are gaining higher qualifications over time, so the meaning of a given level of educational achievement in terms of socioeconomic status is different for different age groups.

The resulting data are characteristic of official statistics on mortality where there is a mixture of minority and majority ethnic groups. The total person-years at risk are 5,244,013 for the nMnPI majority and just 385,562 and 182,202 for the Maori and Pacific Island minorities, respectively. The highest qualification, our indicator of socioeconomic status, is coarsely classified into just four categories. The person-years at risk vary in each cell of the cross-classification from over 100,000 person-years to below 10 and, with only a few person-years, the weighted estimate of the mortality rate varies from over 50 percent to zero.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis is recommended as the first step in a conventional analysis (11, 12). Kaufman et al. (3) smooth the mortality rate across the dimensions of age and socioeconomic status for each combination of gender and ethnicity. Their method is equivalent to kernel regression by a gaussian kernel with its bandwidth parameter fixed at $h = \sqrt{1/2}$ (13). They assume both a fixed bandwidth parameter and a relative scale between the dimensions of age and socioeconomic status. The bandwidth parameter controls the amount of smoothing; higher values give greater smoothing (13). Age is divided by 2 years, and income (their measure of socioeconomic status) is divided by its standard deviation where this is calculated separately for each combination of gender and ethnicity.

Kaufman et al. show that their method works well for majority ethnic groups. They work with 27 income categories and with age in 1-year categories. With our data, their method is adequate for the nMnPI majority (figure 1). As an example, the 1 percent mortality rate occurs at a younger age for the Maori relative to the nMnPI majority, with Pacific Islanders intermediate. In each ethnic group, this rate occurs at a younger age in males than in females. A protective effect of education is seen in the nMnPI majority for both genders between the ages of 30 and 40 years. At this point, shifting from no education to a secondary school qualification delays the increase in mortality with age by about 10 years.

However, even in the nMnPI majority, kernel regression leads to contours that change in a stepwise fashion rather

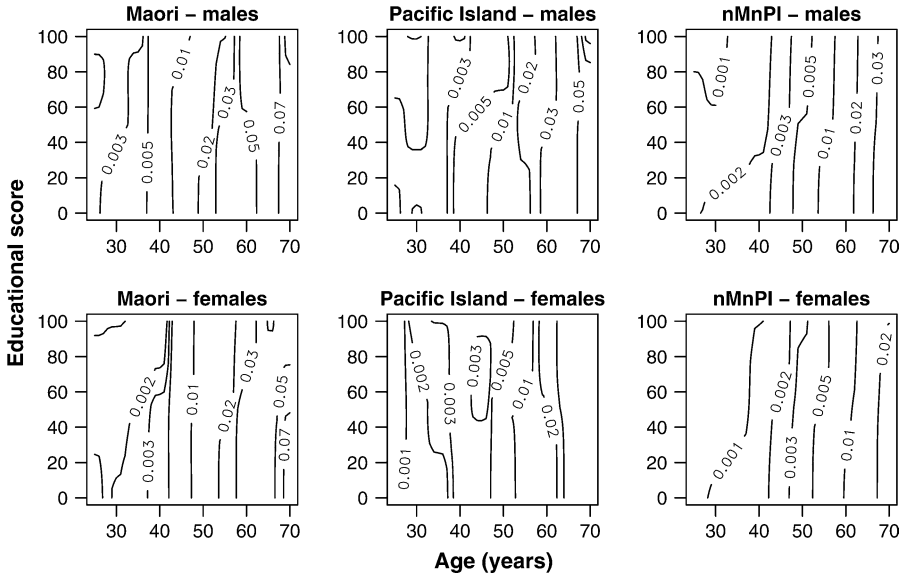

FIGURE 1. Mortality rate contours using kernel regression, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

than smoothly between education categories (figure 1). Kernel regression is essentially a weighted moving average estimating a local constant and, at boundaries in the data, the kernel is asymmetric, and consequently estimates are biased (14). These boundary effects can be mitigated by the use of smoothing methods that estimate a local line or curve, because these provide a more accurate estimate across or into regions where there are no data (14).

One improvement is to smooth using a smoothing spline rather than a kernel. A smoothing spline is a form of nonparametric regression. Observations y_i are modeled as some unspecified (but twice differentiable) function f of a variable x_i with errors ε_i that have zero mean and equal variance (15):

$$y_i = f(x_i) + \varepsilon_i. \quad (2)$$

A spline results from minimizing a modified sum of squares $SS(h)$ (15):

$$SS(h) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + h \int_{x_{\min}}^{x_{\max}} [\partial^2 f(x)]^2 dx, \quad (3)$$

where h is a smoothing parameter, equivalent to the bandwidth parameter in kernel regression. The first term in equation 3 is the error sum of squares, and the second term is a ''roughness penalty'' that is large when $f(x)$ is rough (i.e., when the slope of $f(x)$ changes rapidly over the range of the variable x). Equation 3 represents a compromise between goodness of fit (the first term) and smoothness (the second term). The smoothing parameter h determines the relative importance of these two terms and therefore controls how much the data are smoothed. Parameter h is often chosen by cross-validation (15).

As a consequence of equations 2 and 3, the spline is a series of cubic polynomial curves; these curves join at knots, and the knots are constructed so that the ''join'' is smooth. Fitting the spline requires estimates of the four coefficients that describe each cubic polynomial (15). A full thin-plate spline is a multivariate generalization of this smoothing spline (15), and a thin-plate regression spline is an approximation of the full thin-plate spline; the approximation is quicker to fit and more stable (16, 17). With a bivariate thin-plate regression spline for age and educational score (figure 2), the boundary effects disappear, although the contours for ''Pacific Island–males'' are clearly unrealistic.

Further extensions lead to the generalized additive model (18). First the observations y_i may be an additive function of several variables, where the functional form of each remains unspecified:

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) \cdots + f_J(x_{iJ}) + \varepsilon_i. \quad (4)$$

Second observations may come from the exponential family of distributions, so that the error sum of squares is replaced by a different function of errors, and a link function g is chosen to restrict the range of the expected curve:

$$g(E[y]) = \alpha + \sum_{j=1}^{J} f_j(x_j), \quad (5)$$

where each $f_j(x_j)$ has zero mean, a constraint ensuring that the model is identifiable (19). For simplicity, equations 4 and 5 are shown as the sum of univariate splines, but some or all of these splines could be multivariate.

The generalized additive model is a useful framework for adding and subtracting model structure, following a strategy of adding just enough structure to gain a sensible picture. We could, for example, construct three generalized additive models, one for each ethnic group, with each generalized additive model having an additive difference between male

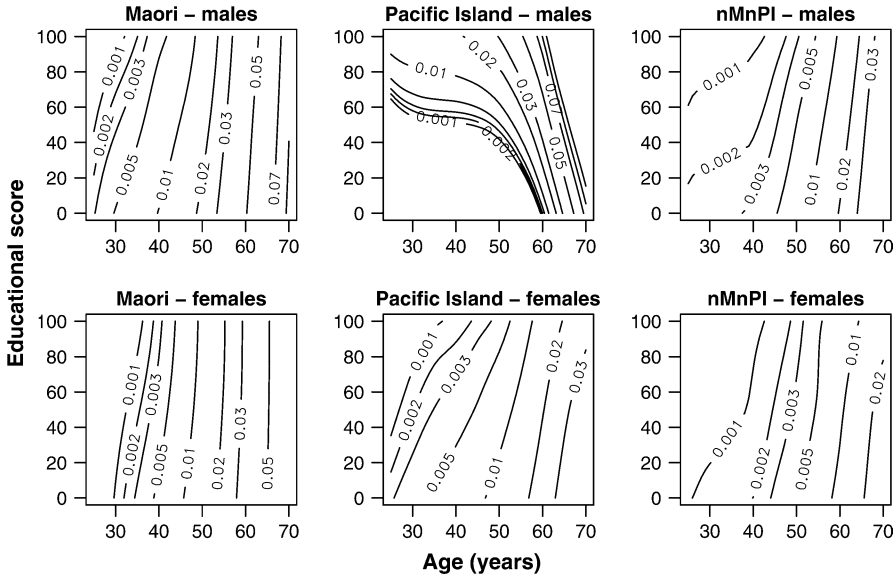

FIGURE 2. Mortality rate contours using a bivariate thin-plate regression spline, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

and female in the form of the bivariate spline for age and educational score. However, with our data, we can still produce sensible contour plots even if we smooth each combination of gender and ethnicity separately.

It is reasonable to view death as random and therefore a Poisson process (20). We expect variation in mortality rates between cells of the cross-classification because of both observed and unobserved covariates (20). This suggests that, for each combination of gender and ethnicity, the number of deaths will follow an ''overdispersed'' Poisson distribution, where the variance in the number of deaths is approximately some multiple of the Poisson mean (21, p. 199) and where the Poisson mean is some unspecified function of age and education. We also expect that the number of deaths will be directly proportional to the person-years at risk. We choose a log-link function, so that the expected number of deaths must be greater than zero. The Poisson generalized additive model that meets these specifications is equivalent to smoothing the mortality rate on a log scale. For each combination of gender and ethnicity, we smooth the mortality rate on a log scale across the dimensions of age and education using a bivariate thin-plate regression spline (figure 3).

Up to this point, we apply the same relative scale between age and education as used by Kaufman et al. However, we do not need to assume a relative scale if we use a bivariate tensor-product spline, a bivariate spline formed from the tensor product (a type of vector multiplication (22)) of univariate spline smoothing in each dimension (23–25). The default univariate spline has five knots, but we set the number of knots for the education dimension to three to ensure at least a degree of smoothing (figure 4).

In summary, we suggest three improvements to the smoothing proposed by Kaufman et al. These improvements should give sensible contour plots even when the data are coarsely cross-classified and highly variable. We replace kernel smoothing (figure 1) with spline smoothing (figure 2), smooth the mortality rate on a log scale rather than on a linear scale (figure 3), and use a bivariate spline that is appropriate when variables are measured in different units (figure 4).

For data of this sort, the second step in a conventional analysis might be model building and statistical inference using Poisson regression (9). As an example, at the end of the next section, we consider the hypothesis that the protective effect of education differs between ethnic groups.

HIERARCHICAL BAYESIAN POISSON REGRESSION

Christiansen and Morris (26) describe an appropriate framework for Bayesian inference, where the analyst views death as random and therefore a Poisson process with a different rate in each cell of the cross-classification. By use of the notation $x \sim D[a, b]$ to represent a random variable x distributed D with mean a and variance b, their hierarchical Poisson regression model for the full cross-classification has three levels:

$$d_i|e_i,\lambda_i \sim \text{Poisson}\,[e_i\lambda_i, e_i\lambda_i], \tag{6}$$

$$\lambda_i|X_i,\beta,\zeta \sim \text{gamma}\,[\mu_i, \mu_i^2/\zeta], \quad \log(\mu_i) = X_i\beta, \tag{7}$$

$$(\beta,\zeta) \sim \pi. \tag{8}$$

At the first level, the number of deaths d_i is distributed Poisson with a mean and a variance $e_i\lambda_i$, where e_i and λ_i are the person-years at risk and mortality rate, respectively, in the ith cell. At the second level, the mortality rate λ_i is

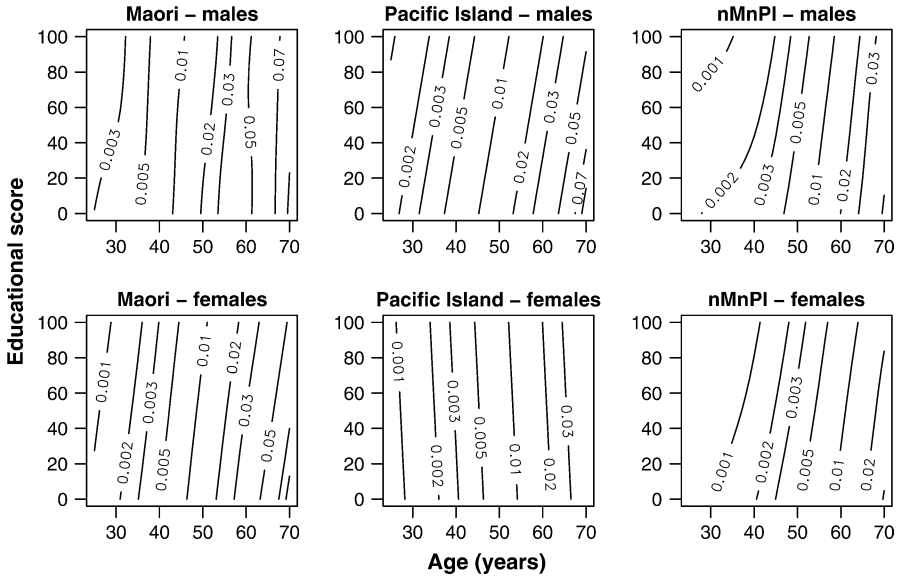

FIGURE 3. Mortality rate contours using Poisson generalized additive models with smoothing by a bivariate thin-plate regression spline, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

distributed gamma, with a mean μ_i that depends, through a log-link function, on a prior structure given by covariates X_i with parameters β estimated from the data. The variance of the mortality rate (μ_i^2/ζ) depends on ζ (the shape parameter of the gamma distribution) and, at the third level, a prior distribution π is required for parameters β and ζ.

In the Bayesian model, the prior covariate structure influences the mean of the posterior rate, but the degree of influence depends on the overall support for this prior structure and on how much local information is available. How this works can been seen from the conditional posterior distribution for the Poisson rate parameters, although the process is more complicated in the marginal posterior

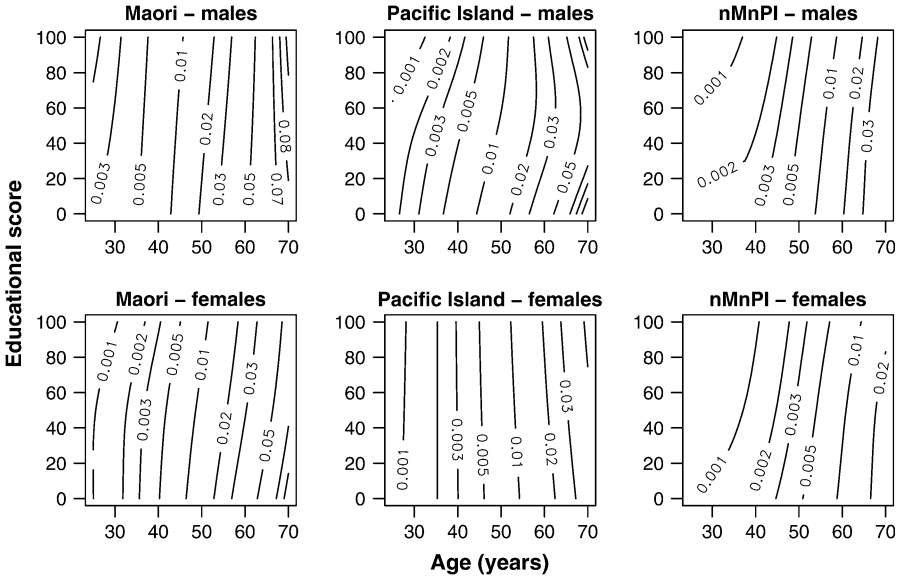

FIGURE 4. Mortality rate contours using Poisson generalized additive models with smoothing by a bivariate tensor-product spline with knot constraints, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

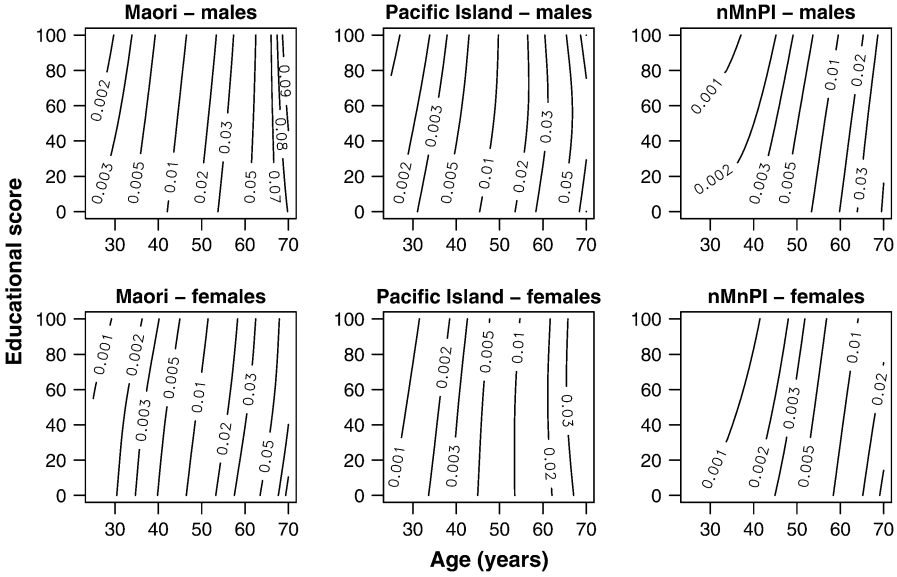

FIGURE 5. Posterior point estimate contours using gamma generalized additive models with smoothing by a bivariate tensor-product spline with knot constraints, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

distribution. The conditional posterior distribution is gamma with mean:

$$\mathrm{E}[\lambda_i|\mathrm{data},\beta,\zeta] = B_i\mu_i + (1-B_i)y_i, \tag{9}$$

where $y_i = d_i/e_i$ is the observed mortality rate in the ith cell and where

$$B_i = \zeta/(\zeta + e_i\mu_i). \tag{10}$$

The B_i lie between zero and one and are known as "shrinkages" because values near one shrink the posterior mean rate away from the observed rate toward the prior structure. The gamma shape parameter acts as a measure of confidence in the prior structure. Large values of ζ lead to shrinkages close to one, and more weight is attached to the prior structure. The shrinkages also depend on the amount of information in the cell through the expected number of deaths, $e_i\mu_i$; cells with more information lead to shrinkages close to zero, and more weight is attached to the observed rate in the cell.

In a hierarchical Bayesian analysis, the second-stage parameters β and ζ are given a prior distribution. Christiansen and Morris assume a priori that β and ζ are independent and use a flat uniform prior for the β parameters, so that $p(\beta,\zeta) = p(\xi|\beta)p(\beta) = p(\xi)$. They then use a "uniform shrinkage prior" for ζ where

$$B_0 = \zeta/(\zeta + d_0) \sim \mathrm{uniform}\,(0,1), \tag{11}$$

and where d_0 is chosen to represent one's confidence in the prior structure. This uniform distribution transforms to a prior distribution for ζ with d_0 as its median (26). This suggests a strategy for choosing d_0: Set it equal to an expected number of deaths at which one is ambivalent about the weight attached to the prior structure and to the observed rate in a cell. This prior is relatively noninformative (27), and posterior inference seems to be relatively robust to the choice of d_0 (refer also to Albert's chapter in the book edited by DeGroot et al. (28)).

If the posterior estimate of ζ is large, this implies strong support for the prior covariate structure. There is then little variance (μ_i^2/ζ) in the mortality rate λ_i around its expected value μ_i (equation 7). As ζ tends to infinity, the hierarchical Poisson model reduces to a conventional Poisson regression model with $\log(\mu_i) = X_i\beta$. In this way, ζ is a measure of uncertainty about the prior covariate structure, and this structure represents the usual Poisson regression model (29).

Having read the analysis by Kaufman et al. (3), we consider the following prior structure for our data: mortality depends on gender, on age but with a different association for different ethnic groups, and on education but with an association that varies with ethnic group and with age. This structure implies a log-linear model for the expected mortality rate (equation 7), with terms for age, sex, Maori ethnicity, Pacific Island ethnicity, educational score, and interaction terms for age and ethnicity, education and ethnicity, and age and education. With a cell count of 10 deaths, we might be ambivalent about the weight attached to this structure and to the observed rate in a cell. This implies that, in cells with higher expected counts, we would want the observed mortality rate to be given more weight than the prior model, and we would want the reverse in cells with lower expected counts.

These prior considerations lead to posterior estimates of the mortality rate (Web appendix A), which we then smooth using a generalized additive model for each combination of gender and ethnicity (figure 5). Our use of generalized additive models in this context is to interpolate continuous age by education surfaces at points where no observation was made, because these surfaces are easier to interpret than a table of 240 cells. Gelman (30) suggests applying the ideas

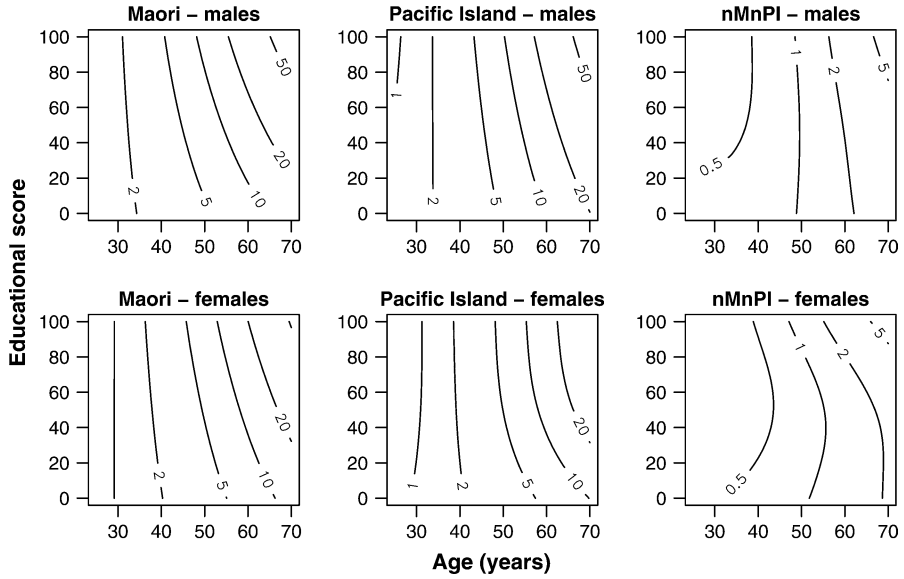

FIGURE 6. Ninety-five percent credible interval width contours ($\times 10^3$) using gamma generalized additive models with smoothing by a bivariate tensor-product spline with knot constraints, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

and methods of exploratory data analysis to structures other than raw data, such as plots of parameter inferences; comparing observed (figure 4) and predicted (figure 5) mortality rates may suggest ways in which the fitted model departs from the data. Our approach to the analysis of social variation in health has much in common with the analysis of spatial variation in health (31, 32).

Because the marginal posterior distribution for the mortality rate λ_i is approximately gamma (26), we smooth these using a gamma generalized additive model with a log link (figure 5). In the same way, we also smooth the widths of the 95 percent credible intervals (Web appendix A; figure 6). The shrinkages B_i are distributed approximately beta and estimated as $a_i/(a_i + b_i)$, where a_i and b_i are estimates in each cell of the beta distribution parameters (26). The beta distribution does not belong to the exponential family of distributions and so cannot be fit as a generalized additive model. Instead, we smooth a_i ''successes'' in $(a_i + b_i)$ ''trials'' as an ''overdispersed'' binomial generalized additive model with a logit link (figure 7), so that the first and

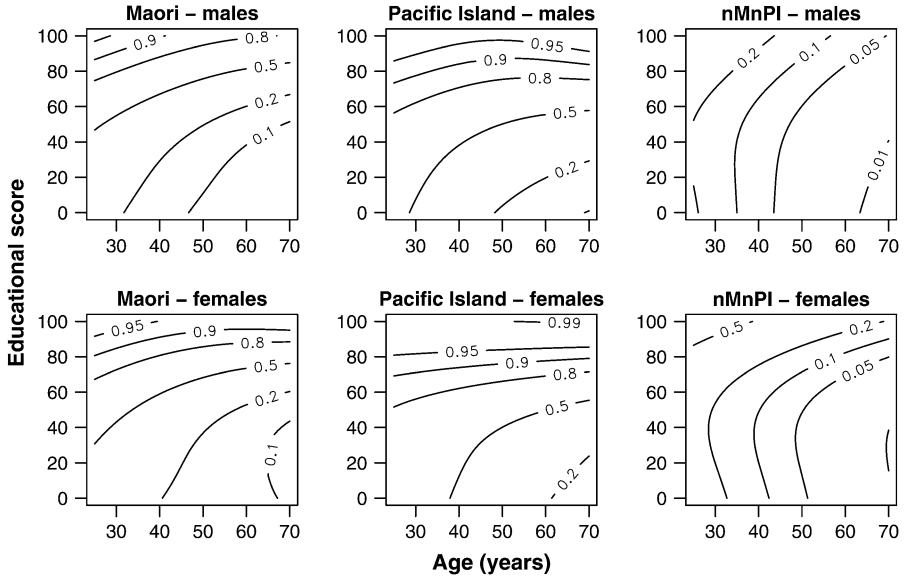

FIGURE 7. Shrinkage contours using binomial generalized additive models with smoothing by a bivariate tensor-product spline with knot constraints, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

second moments of our generalized additive model are the same as those of the beta distribution (33, 34).

To summarize our Bayesian approach, our posterior point estimates of the mortality rate λ_i (figure 5) are similar to those suggested by exploratory data analysis (figure 4). Of course, exploratory data analysis by itself does not allow a formal comparison of the differences (e.g., between ethnic groups); for this, we need to consider the variance in estimates. The marginal posterior distribution for the λ_i is approximately gamma with a variance that depends on the posterior mean rate and on the person-years at risk (26). Therefore, credible intervals become wider with age (figure 6), because the mortality rate increases with age. Credible intervals are also wider at higher educational scores, where there are fewer person-years at risk, and for this reason, they are much wider for the two minorities than for the nMnPI majority. Shrinkages show that our prior structure has a strong influence on posterior estimates for both minorities in the region of higher educational scores and younger ages (figure 7), with values in this region close to the maximum value of one.

Posterior estimates of β parameters are often of interest. We consider the hypothesis that the protective effect of education differs between ethnic groups. Using the prior structure previously described and with age centered at 50 years, we find that education appears to have a protective effect such that, in the nMnPI majority, the expected mortality rate at an educational score of zero is 2.35 (95 percent credible interval (CI): 1.95, 2.83) times the expected mortality rate at a score of 100. However, for Maori and Pacific Islanders, the expected mortality rates at a score of zero are only 1.54 (95 percent CI: 1.19, 2.00) and 1.37 (95 percent CI: 0.96, 1.95) times the respective rates at a score of 100.

Credible intervals for the hierarchical Poisson model are wider than confidence intervals for the equivalent conventional Poisson model (table 1). The hierarchical model assesses support for a prior Poisson regression model, so its credible intervals reflect both parameter uncertainty and uncertainty about this prior model. The conventional confidence interval is a conditional inference: It assumes that the fitted Poisson model is correct. This is unrealistic, so estimates from a hierarchical model are typically more accurate—with a lower mean squared error (35, 36)—than those from a conventional model. Here, the conventional model leads to contours without the well-defined curvature in the nMnPI majority that suggests that a secondary school qualification has a strong protective effect (figure 8). This curvature remains in the hierarchical Poisson model (figure 5), even when this conventional model is used as its prior structure, because there is strong support from the data for this curvature and therefore little shrinkage towards the prior structure in this region of the data (figure 7). This curvature suggests that the association among mortality, age, and education is more complex than we anticipated.

TABLE 1. Mortality rate for males and females aged 50 years who had an educational score of zero as a multiple of their mortality rate with a score of 100, New Zealand, 1996–1999

Ethnic group	Hierarchical Poisson model		Conventional Poisson model	
	Mortality rate ratio	95% credible interval	Mortality rate ratio	95% confidence interval
Non-Maori, non-Pacific Island majority	2.35	1.95, 2.83	2.16	2.04, 2.30
Maori	1.54	1.19, 2.00	1.53	1.34, 1.75
Pacific Island	1.37	0.96, 1.95	1.47	1.14, 1.91

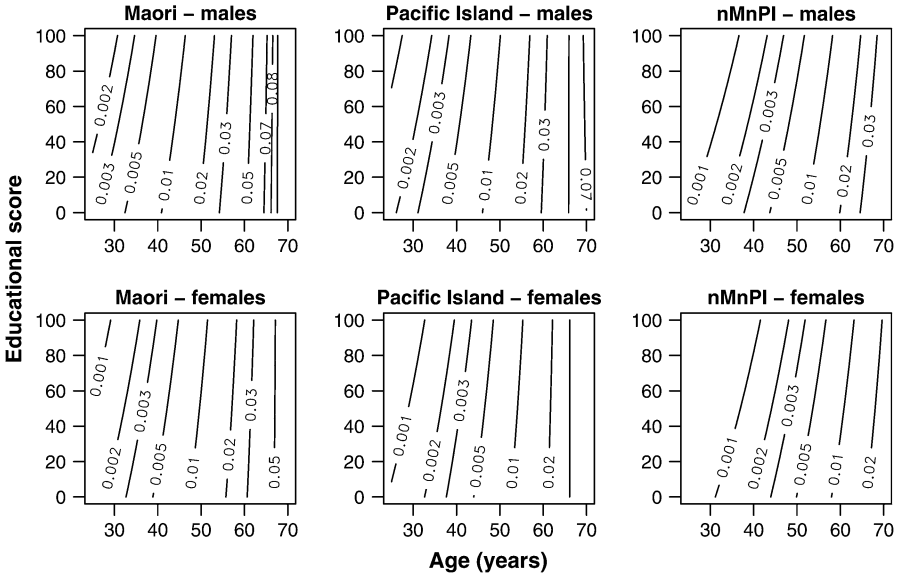

FIGURE 8. Point estimate contours for conventional Poisson regression using Poisson generalized additive models with smoothing by a bivariate tensor-product spline with knot constraints, New Zealand, 1996–1999. nMnPI, non-Maori, non-Pacific Island majority.

DISCUSSION

Spline smoothing is likely to give a clearer exploratory data analysis than is kernel smoothing if data are coarsely cross-classified and highly variable within some cells of that cross-classification. The generalized additive model is a useful framework for adding and subtracting model structure following a strategy of adding just enough structure to gain a clear picture. With these tools, the conventional approach—exploratory data analysis followed by modeling and statistical inference—is possible with mortality data from a mixture of majority and minority ethnic groups.

In hierarchical Bayesian Poisson regression, we add model structure by specifying a prior covariate structure. However, both the amount of local information and the overall fit of the prior structure determine the degree to which this prior structure influences posterior estimates of the mortality rate. Markov chain Monte Carlo methods can be used to fit a hierarchical Poisson regression model. However, the method described by Christiansen and Morris is much quicker, so that it is easy to carry out sensitivity analyses using other prior covariate structures or with a different level of confidence (d_0) in a given prior structure.

Conventional statistical inference, at least in theory, considers support for hypotheses proposed a priori, rather than for those suggested by exploratory data analysis. In practice, ''the best analyses are those that combine both, flagrantly moving easily from ideas the investigator initially proposed to ideas suggested by the data'' (37, p. 780). By comparing observed and predicted patterns of mortality, the investigator can identify a variety of models that appear to be consistent with the data (3). However, the investigator may be mislead into reporting false positive results by chance variation in the data (38). The advantage of the hierarchical Bayesian analysis is that its statistical inference is not conditional on specifying the correct Poisson regression model; rather, its intervals reflect both parameter uncertainty and uncertainty about a Poisson regression model proposed a priori. In addition, prior information about the likely direction and magnitude of covariate effects can be incorporated into a hierarchical model by using an informative prior at the highest level of the model (Web appendix B). When the prior evidence for a hypothesis is strong, a positive study is more likely to be a true positive. ''The mistake is to confuse an increment in support from a positive study with cumulatively strong support for the hypothesis'' (39, p. 958). Focusing on cumulative support for a hypothesis is the key to avoiding spurious findings in epidemiology.

SOFTWARE

All analyses and plots use the R system for statistical computation and graphics version 1.9.1 (40). Generalized additive models were fit with an add-on package, mgcv version 1.1–5. Both R and mgcv are available from the Comprehensive R Archive Network website (http://cran.R-project.org/); further information on mgcv is available from its author, Simon Wood (http://www.maths.bath.ac.uk/~sw283/). The hierarchical Poisson regression model of Christiansen and Morris was fit by use of their Splus code (PRIMM), available from the ''Statlib'' website (http://lib.stat.cmu.edu/S/). Minor changes are needed to make this code run within the R system.

SUMMARY OF STATISTICS NEW ZEALAND SECURITY STATEMENT

The full security statement is published at http://www.wnmeds.ac.nz/nzcms-info.html.

The New Zealand Census-Mortality Study is a study of the relation between socioeconomic factors and mortality in New Zealand, based on the integration of anonymous population census data from Statistics New Zealand and mortality data from the New Zealand Health Information Service. The project was approved by Statistics New Zealand as a Data Laboratory project under the Microdata Access Protocols in 1997. The data sets created by the integration process are covered by the Statistics Act and can be used for statistical purposes only. Only approved researchers who have signed Statistics New Zealand's declaration of secrecy can access the integrated data in the Data Laboratory. For further information about confidentiality matters in regard to this study, please contact Statistics New Zealand.

ACKNOWLEDGMENTS

This project was supported by a University of Otago research grant.

The authors thank June Atkinson and Richard Penny for sharing their knowledge of the data used in this project and Simon Wood for his help with mgcv.

Conflict of interest: none declared.

REFERENCES

1. Smith GD. Learning to live with complexity: ethnicity, socioeconomic position, and health in Britain and the United States. Am J Public Health 2000;90:1694–8.
2. Nazroo JY. The structuring of ethnic inequalities in health: economic position, racial discrimination, and racism. Am J Public Health 2003;93:277–84.
3. Kaufman JS, Long AE, Liao Y, et al. The relation between income and mortality in U.S. blacks and whites. Epidemiology 1998;9:147–55.
4. Greenland S. Randomization, statistics, and causal inference. Epidemiology 1990;1:421–9.
5. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986;15:413–19.
6. Blakely T, Woodward A, Salmond C. Anonymous linkage of New Zealand mortality and census data. Aust N Z J Public Health 2000;24:92–5.
7. Hill S, Atkinson J, Blakely T. Anonymous record linkage of census and mortality records: 1981, 1986, 1991, 1996 census cohorts. Wellington, New Zealand: Department of Public

Health, Wellington School of Medicine and Health Sciences, University of Otago, 2002.
8. Blakely T, Robson B, Atkinson J, et al. Unlocking the numerator-denominator bias. I. Adjustments ratios by ethnicity for 1991–94 mortality data. The New Zealand Census-Mortality Study. N Z Med J 2002;115:39–43.
9. Blakely T, Kawachi I, Atkinson J, et al. Income and mortality: the shape of the association and confounding New Zealand Census-Mortality Study, 1981–1999. Int J Epidemiol 2004; 33:874–83.
10. Bross IDJ. How to use ridit analysis. Biometrics 1958;14:18–38.
11. Tukey JW. We need both exploratory and confirmatory. Am Stat 1980;34:23–5.
12. Chatfield C. The initial examination of data. J R Stat Soc (A) 1985;148:214–53.
13. Michels P. Asymmetric kernel functions in nonparametric regression: analysis and prediction. Statistician 1992;41:439–54.
14. Hastie T, Loader C. Local regression: automatic kernel carpentry. Stat Sci 1993;8:120–43.
15. Silverman BW. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. J R Stat Soc (B) 1985;47:1–52.
16. Wood SN. mgcv: GAMs and generalized ridge regression in R. R News 2001;1:20–5.
17. Wood SN. Thin plate regression splines. J R Stat Soc (B) 2003;65:95–114.
18. Hastie T, Tibshirani R. Generalized additive models; some applications. J Am Stat Assoc 1987;82:371–86.
19. Wood SN, Augustin NH. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol Modell 2002;157:157–77.
20. Brillinger DR. The natural variability of vital rates and associated statistics. Biometrics 1986;42:693–712.
21. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London, United Kingdom: Chapman and Hall, 1989.
22. Rougier J. What's the point of 'tensor'? R News 2001;1:26–7.
23. Barry D. Nonparametric Bayesian regression. Ann Stat 1986; 14:934–53.
24. Gu C, Wahba G. Discussion: multivariate adaptive regression splines. Ann Stat 1991;19:115–23.
25. Wood SN. Low rank scale invariant tensor product smooths for generalized additive mixed models. Glasgow, Scotland: Department of Statistics, University of Glasgow, 2004. (Technical report 04-13).
26. Christiansen CL, Morris CN. Hierarchical Poisson regression modeling. J Am Stat Assoc 1997;92:618–32.
27. Daniels MJ. A prior for the variance in hierarchical models. Can J Stat 1999;27:567–78.
28. Albert JH. Bayesian estimation of Poisson means using a hierarchical log-linear model. In: DeGroot MH, Lindley DV, Smith AFM, et al, eds. Bayesian statistics 3: proceedings of the Third Valencia International Meeting, June 1–5, 1987. Oxford, United Kingdom: Oxford University Press, 1989: 519–31.
29. Albert JH. Computational methods using a Bayesian hierarchical generalized linear model. J Am Stat Assoc 1988;83: 1037–44.
30. Gelman A. Exploratory data analysis for complex models. J Comput Graph Stat 2004;13:755–79.
31. Pascutto C, Wakefield JC, Best NG, et al. Statistical issues in the analysis of disease mapping data. Stat Med 2000;19: 2493–519.
32. Lawson AB. Disease map reconstruction. Stat Med 2001;20: 2183–204.
33. Kieschnick R, McCullough BD. Regression analysis of variates observed on (0,1): percentages, proportions, and fractions. Stat Model 2003;3:193–213.
34. Papke LE, Wooldridge JM. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. J Appl Econometrics 1996;11: 619–32.
35. Witte JS, Greenland S. Simulation study of hierarchical regression. Stat Med 1996;15:1161–70.
36. Greenland S. Principles of multilevel modelling. Int J Epidemiol 2000;29:158–67.
37. Hertz-Picciotto I. What you should have learned about epidemiologic data analysis. Epidemiology 1999;10:778–83.
38. Mills JL. Data torturing. N Engl J Med 1993;329:1196–9.
39. Savitz DA. Commentary: prior specification of hypotheses: cause or just a correlate of informative studies? Int J Epidemiol 2001;30:957–8.
40. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2004.