Published in final edited form as:

Biometrics. 2008 December ; 64(4): 1054–1061. doi:10.1111/j.1541-0420.2008.00997.x.

Incorporating Marginal Covariate Information in a Nonparametric Regression Model for a Sample of $R \times C$ Tables

Joan G. Staniswalis

University of Texas at El Paso, Department of Mathematical Sciences, 500 West University Avenue, Bell Hall 124, El Paso, Texas 79902, U.S.A. email: joan@math.utep.edu

Summary

Nonparametric regression models are proposed in the framework of ecological inference for exploratory modeling of disease prevalence rates adjusted for variables, such as age, ethnicity/race, and socio-economic status. Ecological inference is needed when a response variable and covariate are not available at the subject level because only summary statistics are available for the reporting unit, for example, in the form of $R \times C$ tables. In this article, only the marginal counts are assumed available in the sample of $R \times C$ contingency tables for modeling the joint distribution of counts. A general form for the ecological regression model is proposed, whereby certain covariates are included as a varying coefficient regression model, whereas others are included as a functional linear model. The nonparametric regression curves are modeled as splines fit by penalized weighted least squares. A data-driven selection of the smoothing parameter is proposed using the pointwise maximum squared bias computed from averaging kernels (explained by O'Sullivan, 1986, *Statistical Science* 1, 502–517). Analytic expressions for bias and variance are provided that could be used to study the rates of convergence of the estimators. Instead, this article focuses on demonstrating the utility of the estimators in a study of disparity in health outcomes by ethnicity/race.

Keywords

Ecological inference; Incomplete $R \times C$ tables; P-splines; Randomized response

1. Introduction

Records on health outcomes, demographic variables, and environmental exposures are becoming readily available as a result of a concerted effort among government agencies to make the best use of all available data (Mather et al., 2004). Many of these databases are not a random sample of the population and information on confounding variables is needed to reduce bias when making inferences for the population. Yet, socio-economic information is not available at the subject level due to Title 13 preventing the disclosure of U.S. Census data; Title 13 guarantees confidentiality of census information, and establishes penalties for disclosing this information. However, U.S. Census data can be obtained in summary form for state-county units. Thus, disparate databases with health outcomes and demographic information can be merged by linking records at the state-county level, and health outcomes studied using statistical analysis methods from ecological inference.

Ecological inference refers to methods of statistical inference when subject-level data are not available, because only summary statistics are reported for groups such as state-county units. A distinguishing feature of the ecological data sets obtained in this way is that the joint information is not available for those variables imported from different databases. Methodologic problems encountered by performing an analysis with ecological data, such as specification and aggregation bias, are reviewed in Web Appendix A. Specification and

aggregation bias may be reduced by including additional information, from both the subject level and reporting unit, in the regression model for ecological inference.

This article proposes methods for nonparametric adjustment of disease rates by covariates, given marginal information on the joint distribution of the response and covariates. The joint distribution of subject-level covariates is assumed known for each reporting unit. The literature on ecological inference, given a sample of $R \times C$ tables, has focused mostly on evaluating environmental exposures or on understanding election voting patterns (entire issue of *Journal of Royal Society, Series A*, 2001, 164(1); King, Rosen, and Tanner, 2004). Rosen et al. (2001) consider parametric covariate adjustments using both Bayesian and frequentist approaches. King (2004) offers a free software package for predicting the cell probabilities without covariate adjustment, given marginal data for a sample of 2×2 contingency tables. Natário and Knorr-Held (2003) studied nonparametric modeling for ecological data, given complete counts in $R \times C$ contingency tables on each reporting unit.

The remaining sections of the article are organized as follows. The data that motivated the study in this article are introduced in Section 2. Section 3 has the proposed model. Statistical methods for estimation are explained in Sections 4 and 5. The data analysis is described in Section 6, where the utility of the nonparametric regression tool is demonstrated for the ecological data.

2. The Data Motivating This Investigation

Pazmiño and Pazmiño (2003) reported a diabetic and renal failure belt along the U.S.–Mexican border. Current thinking is that the extra risk of end-stage renal disease (ESRD) in Hispanics is explained by complications of diabetes, and in African Americans by complications of hypertension, but poverty and lack of access to medical care may be contributing to the poor health outcomes. Thus, it is of interest to estimate the ESRD prevalence rate to explore the interaction between ethnicity and county median household income in Texas.

For the $m = 80$ Texas counties with Medicare-approved dialysis centers in 1998, define the variables as follows: n county population; ω county median household income; Z number of ESRD patients treated, excluding kidney transplants. These data were published in Pazmiño and Pazmiño (2003; table 1); the original sources were the U.S. Census 2000, 1997 Texas Almanac, and the offices of the End-Stage Renal Disease Network of Texas, respectively. The marginal counts for renal failure as a complication of diabetes (ESRD-DM) were also reported. As an example, Table 1 provides the marginal counts available for El Paso County.

See Table 2 for the five number summaries based on the data for the $m = 80$ Texas counties. The first row of Figure 1 displays exploratory plots for the responses, ESRD and ESRD-DM prevalence per million versus median household income. The prevalence of renal failure appears to be decreasing with median household income. It is of interest to determine whether this pattern is evident across ethnic/racial groups. The second row in Figure 1 summarizes the distribution of ethnicity/race and age obtained from the 2000 U.S. Census website (www.census.gov). Hispanic (Other) representation declines as median household income of the county increases (decreases), with a high representation in counties with low (high) median household income. The highest representation of African Americans is in the middle-class communities where there is roughly a comparable presence of Hispanics. Hispanics and African Americans have similar age distributions with more youth and fewer elderly people as compared with Other.

3. The Proposed Model

Covariates enter the proposed nonparametric model in one of two ways: through a varying coefficient model for reducing aggregation bias, or a functional linear model to reduce specification bias. A varying coefficient model (Hastie and Tibshirani, 1993) for a scalar response Y in terms of a covariate X is given by $Y = \beta(\omega)X + \varepsilon$, where the covariate ω is allowed to have a modulating effect on the coefficient β for X . A form of the functional linear model (Ramsay and Silverman, 1997) for a scalar response Y in terms of a covariate $X(t)$, $t \geq 0$, is given by $Y = \int_0^\infty \beta(t)X(t)dt + \varepsilon$. The functional linear model replaces the usual inner product $X\beta$ on a finite-dimensional vector space, with an inner product $\langle X, \beta \rangle$ on a Hilbert space, say on L_2 .

Notation is now introduced to specify the model for the ESRD counts. For $j = 1, \dots, n_i$ and $i = 1, \dots, m$, consider the triple (u_{ij}, v_{ij}, z_{ij}) : z_{ij} is the indicator of ESRD status; u_{ij} is the categorical variable for ethnicity/race (1: Hispanic, 2: African American, 3: Other); and v_{ij} is the age, for the j th person in county i . Suppose that ω_i is the county-level covariate median household income reflecting access to medical care, and the $z_{ij}|\omega_i$, u_{ij} , v_{ij} are independent realizations of Bernoulli random variables with the parameter given as an additive model (Hastie and Tibshirani, 1990) in θ and ψ :

$$g[\theta(u, \omega) + \psi(v)], \quad (1)$$

where only the function g is known and $E_V[\psi(V)] = 0$. If the subject-level data (u_{ij}, v_{ij}, z_{ij}) were available, then logistic regression could be used for nonparametric estimation of $\theta(\cdot)$ and $\psi(\cdot)$.

However, the ESRD status of subjects ($Z_{ij} = z_{ij}$) is not available, only the total number of cases $z_i = \sum_{j=1}^{n_i} z_{ij}$ for each county $i = 1, \dots, m$. Further, the ethnicities/races and ages (u_{ij}, v_{ij}) are part of a U.S. Census database that cannot be disclosed at the subject level (U.S. Title 13), and anyway is not linked with the ESRD subject status z_{ij} . Let $G(v | u, \omega_i)$ denote the conditional distribution of age, given ethnicity/race, and $F(u | \omega_i)$ the distribution of ethnicity/race, for the i th county with county median household income ω_i . Let $I(\cdot)$ denote the indicator function that takes on the value 1 whenever the argument is true, and 0 otherwise. Then

$n_{ik} = \sum_{j=1}^{n_i} I(u_{ij}=k)$ denotes the number of people in each ethnic/racial group in county i , and the respective proportions are $p_{ik} = n_{ik}/n_i$, $k \in \{1, 2, 3\}$. Apply the model for the Bernoulli mean given by equation (1):

$$\begin{aligned} & E[Z_i/n_i | \omega_i, n_i, \{(u_{ij}, v_{ij})\}_{j=1, \dots, n_i}] \\ &= \sum_{k=1}^3 \frac{n_{ik}}{n_i} \sum_{j=1}^{n_i} \frac{1}{n_{ik}} I(u_{ij}=k) E[Z_{ij} | \omega_i, u_{ij}, v_{ij}] \\ &= \sum_{k=1}^3 p_{ik} \int_0^\infty g[\theta(k, \omega_i) + \psi(v)] dG(v | u=k, \omega_i) \end{aligned} \quad (2)$$

$$\begin{aligned} &= \int_0^\infty \int_0^\infty g[\theta(u, \omega_i) + \psi(v)] dG(v | u, \omega_i) dF(u | \omega_i) \\ &= \eta(\omega_i, F; \theta + \psi), \end{aligned} \quad (3)$$

where $F = (F, G)$ and the sums are now written in terms of Riemann–Stieltjes integrals. The expected number of cases for ethnic/racial group $u = k$ and county i would be given by

$$(n_i p_{ik}) \int_0^\infty g[\theta(k, \omega_i) + \psi(v)] dG(v|u=k, \omega_i) \quad (4)$$

for Hispanics ($k = 1$), African Americans ($k = 2$), and Others ($k = 3$).

Although not written explicitly as (3), this regression model was first considered by Prentice and Sheppard (1995) in the context of ecological inference with parametric models for $\theta(u, \omega)$ and $\psi(v)$. This is reviewed in Wakefield and Salway (2001) along with some other related approaches that used parametric models for $G(v|u, \omega)$ (Richardson, Stucker, and Hémon, 1987).

That $\eta(\omega_i, F; \theta + \psi)$ in (3) has features of a varying coefficient model is seen in equation (2), because the regression parameter θ , which goes with the county covariate p_{ik} , is modulated by ω_i . That $\eta(\omega_i, F; \theta + \psi)$ has features of a functional linear model is attributed to the integral over $v | u, \omega_i$; see equation (3). The functional linear model adjusts for continuous variables (age), whereas the varying coefficient model allows for interactions between categorical and continuous variables (ethnicity/race and income).

The usual optimal rates of L_2 convergence for nonparametric estimators have been established for the estimation of $\beta(\cdot)$ in varying coefficient models estimated with regression B-splines (Lu and Mao, 2004) and by local linear modeling (Fan and Zhang, 1999). However, functional linear models fit with penalized B-splines had a slower rate of convergence for the estimation of $\beta(\cdot)$ (Cardot, Ferraty, and Sarda, 2003), comparable to rates obtained in deconvolution problems (Rice and Rosenblatt, 1983). Rates of convergence have not been established when features of both varying coefficient models and functional linear models are included. This article leaves this as an open problem, while demonstrating the utility of the models in ecological inference.

3.1 Model Fitting Criteria

The full model (3) is fit by weighted least squares. This is justified by considering that the contribution of a single observation z_i to a Poisson likelihood: $z_i \log \mu_i - \mu_i$, is approximated by $(z_i - \mu_i)^2 / \mu_i$ for large μ_i , where $\mu_i = n_i \eta_i$ (McCullagh and Nelder, 1989). The weighted least squares approach is justified as a moment-matching estimation procedure. Rosen et al. (2001) also implemented parametric model fitting, given a sample of $R \times C$ contingency tables from this frequentist approach based on first moments.

3.2 Linearization Step

Let $v(u, \omega, v) = \theta(u, \omega) + \psi(v)$ in equation (3) for η . This model for η is not in general linear in v . O'Sullivan (1986) suggests sidestepping this difficulty by using a linear approximation η^* that is fit to adjusted observations z^* .

To this end, the Gâteaux derivative η^* about the curve v_o and the adjusted observations are computed for this example to be

$$\eta^*(\omega, F; v) = \int_0^\infty \int_0^\infty g'[v_o(u, \omega, v)] v(u, \omega, v) dG(v|u, \omega) dF(u|\omega), \quad (5)$$

$$z_i^* = z_i/n_i - \eta(\omega_i, F_i; v_o) + \eta^*(\omega_i, F_i; v_o). \quad (6)$$

The integrals in equation (5) are expected values over the categorical variable u (ethnicity/race) that may be computed by summing, and expected values over the continuous variable v (age) that are approximated by summing over the binned data, for example, using the class interval frequencies provided by the U.S. Census.

The adjusted observations (6) are interpreted as prevalence rates of the disease, rather than the count of disease cases. Because $\eta(\omega, F; v) \approx \eta(\omega, F; v_o) + \eta^*(\omega, F; v - v_o)$ and under $E[z/n | \omega, F] = \eta(\omega, F; v)$, the mean of the pseudo-observations is approximately $\eta^*(\omega_i, F_i; v)$. The variance of the pseudo-observations is obtained by modeling the conditional distribution of the disease counts for each county as Poisson distributed. Conditionally on the covariates, $z_1, \dots, z_m$ are independent and the variance of the adjusted observations given by equation (6) is $\sigma^2 = \eta/n$.

3.3 Background for Nonparametric Curve Fitting

The nonparametric curve fitting uses penalized splines (Eilers and Marx, 1996) based on the cubic B-spline functions with n knots. Cubic B-splines (order $r = 4$) provide a convenient basis for the representation of smooth functions (Eubank, 1988). The B-splines have compact local support, provide a flexible basis for curve fitting, and have overlapping support that is exploited in the P-spline curve fitting technology; this is described next.

An objective function to be minimized $\mathcal{S}(\beta; \text{data})$ for curve fitting is penalized by finite differences of the cubic B-spline basis coefficients β . That is, to estimate β from the data, minimize $\mathcal{S}(\beta; \text{data}) + \beta' P_{n+r}(h) \beta$ with respect to β , where h is a scalar smoothing parameter that controls the roughness of the fit, $P_{n+r}(h) = h(D^2)^t D^2$, and D is the matrix representation of the first-order differencing operator acting on IR^{n+r} : $D\beta = (\beta_2 - \beta_1, \dots, \beta_{n+r} - \beta_{n+r-1})^t$. Eilers and Marx (1996) argue that this penalty is roughly equivalent to the smoothing spline penalty that uses an L_2 norm on the second derivative of the regression curve represented in a B-spline basis. By roughly equivalent it is meant that the difference penalty has the same flavor as the usual penalty based on derivatives. The difference penalty has the effect of imposing smoothness by requiring that the heights of neighboring B-splines “hold hands” to withstand erratic fluctuations in the data.

4. Reduced Model Fitting

Consider first estimation in the reduced model (3) that does not involve ψ , and which only has one unknown function $\theta(u)$ not dependent upon ω , assumed to satisfy $\theta \in \Theta = \text{span } \mathcal{B}$, where $\mathcal{B} = \{B_k(\cdot)\}_{k=1}^K$. The variable u could be continuous or categorical. When u is a continuous variable, the basis $\mathcal{B}$ is taken to be a cubic B-spline basis (order $r = 4$) with n pre-specified knots on the support of the variable u , so that $K = n + r$. On the other hand, when u is a categorical variable with K levels, then the basis is $B_k(u) = 1$ for $u = k$, and 0 otherwise. Working first with the reduced model serves to simplify the presentation for model fitting. The simultaneous estimation of $\theta(u, \omega)$ and $\psi(v)$ is a straightforward generalization, developed in Section 5.

4.1 Method of Penalized Weighted Least Squares

Suppose independent observations $\{(\omega_i, F_i, z_i^*)\}_{i=1}^m$ are available satisfying: (i) $z_i^* = \eta^*(\omega_i, F_i; \theta) + \varepsilon_i$; (ii) ε_i is independent and identically distributed (i.i.d.) with mean 0 and variance σ_i^2 ; and (iii) $\eta^*(\omega_i, F_i; \cdot): \Theta \rightarrow IR$, a continuous linear functional. Define $X_{m \times K} =$

(x_{ik}) to be the matrix with rows $x_1^t, \dots, x_m^t$, and $x_{ik} = \eta^*(\omega_i, F_i; B_k)$. So that X is the image of the basis $\mathcal{B}$ under the linear mappings $\eta^*(\omega_i, F_i; \cdot)$; $i = 1, \dots, m$.

Set $w_i = 1/\sigma_i^2$ and $\mathbf{W}_{m \times m} = \text{diagonal}(w_1, \dots, w_m)$. Let the working weight matrix used at the estimation step be denoted by $\mathbf{W}_o = \text{diagonal}(w_{1o}, \dots, w_{mo})$, where $w_{io} = 1/\sigma_{io}^2$; $i = 1, \dots, m$. Estimation of θ would seek to find an estimator by minimizing over $\beta \in \mathbb{R}^K$,

$$\frac{1}{m}(\mathbf{z}^* - \mathbf{X}\beta)^t \mathbf{W}_o (\mathbf{z}^* - \mathbf{X}\beta) + \beta^t \mathbf{P}_K(h)\beta, \quad (7)$$

where $\mathbf{z}^* = (z_1^*, \dots, z_m^*)^t$. It is straightforward to verify that the minimizer of (7) is given by $\hat{\beta} = \{\mathbf{X}^t \mathbf{W}_o \mathbf{X} + m \mathbf{P}_K(h)\}^{-1} \mathbf{X}^t \mathbf{W}_o \mathbf{z}^*$ assuming the required inverse exists.

Set $\mathbf{B}_u^t = (B_1(u), \dots, B_K(u))$. The estimator of θ is thus given by $\hat{\theta}(u) = \mathbf{B}_u^t \hat{\beta} = \mathbf{S}_u^t \mathbf{z}^*$, where

$$\mathbf{S}_u = \mathbf{W}_o \mathbf{X} \{\mathbf{X}^t \mathbf{W}_o \mathbf{X} + m \mathbf{P}_K(h)\}^{-1} \mathbf{B}_u. \quad (8)$$

Alternately, $\hat{\theta}(u)$ may be expressed as

$$\hat{\theta}(u) = \sum_{i=1}^m s_i(u) z_i^*, \quad (9)$$

where the *impulse response functions* $s_i(u) = w_{io} x_i^t \{\mathbf{X}^t \mathbf{W}_o \mathbf{X} + m \mathbf{P}_K(h)\}^{-1} \mathbf{B}_u$ are given by the entries of the $m \times 1$ column vector $\mathbf{S}_u$.

4.2 Iteration

The final parameter estimates are obtained by updating and iterating to convergence. The iteration updates the functional linearization step (5) and construction of the adjusted observations (6). However, the value of the weight matrix $\mathbf{W}_o$ is not updated in the iteration. The working weight matrix $\mathbf{W}_o$ was not updated so that when η is a linear functional, the fitting criteria only depends on the choice of the smoothing parameter through the penalty term. The sandwich estimator of variance is used to account for misspecification of the weight matrix $\mathbf{W}_o$.

5. Full Model Fitting

Now we address the generalization for fitting equation (3) to data using penalized weighted least squares as in Section 4.1. Suppose J, L are very large, so that for practical purposes of numerical computation of the estimators, it can be assumed that $\theta(u, \omega)$ and ψ are in the span of $\{N_j(\omega) B_k(u)\}_{k=1, \dots, K}^{j=1, \dots, J}$ and in the span of $\{Q_l(v)\}_{l=1}^L$, respectively. Set the notation $\mathbf{B}_u^t = (B_1(u), \dots, B_K(u))$, $\mathbf{N}_\omega^t = (N_1(\omega), \dots, N_J(\omega))$, and $\mathbf{Q}_v^t = (Q_1(v), \dots, Q_L(v))$ for the bases evaluated at a point in the respective domains. Because U is a categorical variable that takes on the coded values $1:K$, $\{B_k(u)\}_{k=1}^K$ is taken as $B_k(u) = 1$ for $u = k$, and 0 otherwise. The bases

$\{N_j(\omega)\}_{j=1}^J$ and $\{Q_l(v)\}_{l=1}^L$ are taken to be cubic B-splines for a given sequence of knots on the support of the respective variables.

Assume the following unique real-valued vectors exist:

- i. $\beta^{(1)} \in IR^{JK}$ depending upon θ , such that

$$\theta(u, \omega) = \sum_j \sum_k \beta_{jk} N_j(\omega) B_k(u) = (\mathbf{B}_u \otimes \mathbf{N}_\omega)^t \beta^{(1)},$$

where $\beta^{(1)}$ is partitioned into K vectors $\beta^{(1)} = (\beta_1^t, \dots, \beta_K^t)^t$ with each vector in the partition identified as $\beta_k = (\beta_{1k}, \dots, \beta_{Jk})^t \in IR^J$, for $k = 1, \dots, K$.

- ii. $\beta^{(2)} = (\beta_{K+1}, \dots, \beta_{K+L})^t \in IR^L$ depending on ψ , such that

$$\psi(v) = \sum_l \beta_{K+l} Q_l(v) = \mathbf{Q}_v^t \beta^{(2)}.$$

Then

$$\begin{aligned} E[z^*] &= \mathbf{X}\beta, \text{ where } \mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}] \text{ and } \beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix}, \\ \mathbf{X}_{m \times JK}^{(1)} &= \begin{bmatrix} (\eta^*(\omega_1, \mathbf{F}_1; B_1), \dots, \eta^*(\omega_1, \mathbf{F}_1; B_K)) \otimes \mathbf{N}_{\omega_1}^t \\ \vdots \\ (\eta^*(\omega_m, \mathbf{F}_m; B_1), \dots, \eta^*(\omega_m, \mathbf{F}_m; B_K)) \otimes \mathbf{N}_{\omega_m}^t \end{bmatrix} \end{aligned}$$

and

$$\mathbf{X}_{m \times L}^{(2)} = \begin{bmatrix} \eta^*(\omega_1, \mathbf{F}_1; Q_1), \dots, \eta^*(\omega_1, \mathbf{F}_1; Q_L) \\ \vdots \\ \eta^*(\omega_m, \mathbf{F}_m; Q_1), \dots, \eta^*(\omega_m, \mathbf{F}_m; Q_L) \end{bmatrix}.$$

As described earlier, β is found to minimize the penalized weighted least squares given by equation (7) over $\beta \in IR^{JK+L}$, except now the penalty is $\beta^t \mathbf{P}(h_1, \dots, h_{K+1}) \beta$, where $\mathbf{P}(h_1, \dots, h_{K+1})$ is a block diagonal matrix given by

$$\mathbf{P}(h_1, \dots, h_{K+1}) = \text{diagonal}(\mathbf{P}_j(h_1), \dots, \mathbf{P}_j(h_K), \mathbf{P}_L(h_{K+1})).$$

The B-spline bases elements satisfy $\sum_j N_j(\omega) = 1$ and $\sum_l Q_l(v) = 1$, so the additive model $\theta(u, \omega) + \psi(v)$ is over-parameterized. This really is not a problem, because the Penrose–Moore inverse can be used in place of the matrix inverse in the following expression for $\hat{\beta}$:

$$\hat{\beta} = \{\mathbf{X}^t \mathbf{W}_o \mathbf{X} + m \mathbf{P}(h_1, \dots, h_{K+1})\}^{-1} \mathbf{X}^t \mathbf{W}_o \mathbf{z}^*.$$

The estimates of θ and ψ may not be unique, but the estimate of $\theta(u, \omega) + \psi(v)$ is unique, so that relative prevalences could still be reported. Alternately, a constraint such as $E_V[\psi(V)] =$

0 could be imposed so that the estimators of θ and ψ are uniquely determined. Finally, we can compute estimates of the components in the additive model:

$$\begin{aligned}\widehat{\theta}(u, \omega) &= (B_u \otimes N_\omega)' \widehat{\beta}^{(1)} \\ \widehat{\psi}(v) &= Q_v' \widehat{\beta}^{(2)},\end{aligned}\tag{10}$$

where $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ are the elements in the partition of $\widehat{\beta}$ into vectors of length JK and L , respectively.

Implementation of the model fitting proceeds by updating the linearization (5) and adjusted observations (6) as follows:

Step 1

Set $v_o(u, \omega, v) = \theta_o(u)$ obtained from a parametric fit in the reduced model without dependence on v and ω . A nonparametric update $\widehat{\theta}(u, \omega)$ to $\theta_o(u)$ is obtained under $\psi_o(v) = 0$.

Step 2

Set $\theta_o(u, \omega) = \widehat{\theta}(u, \omega)$ and $\psi_o(v) = 0$ so that $v_o = \widehat{\theta}(u, \omega)$ from Step 1. Obtain an update $\widehat{\theta}(u, \omega)$ to θ_o , and an update $\widehat{\psi}(v)$ to $\psi_o(v)$ under the full model.

The accuracy of the linearization step in Section 3.2 depends on the accuracy of the linear approximation for $g[v]$, which in turn depends on the magnitude of

$$|v - v_o| \leq \begin{cases} |\theta(u, \omega) - \theta_o(u)| + |\psi(v) - 0| & \text{in Step 1} \\ |\theta(u, \omega) - \widehat{\theta}(u, \omega)| + |\psi(v) - 0| & \text{in Step 2,} \end{cases}$$

under the constraint $E_V[\psi(V)] = 0$. For the ESRD data analysis described in the next section, $g(v) = \exp(v)$ is approximated by $\exp(v_o) [1 + (v - v_o)]$, which should be satisfactory because for the most part $|v - v_o| \leq 1$ and v had a range in $[4, 9]$.

6. ESRD Prevalence Data Analysis

The knots for the cubic B-spline basis on the support of the variable v (age) were set at the center of the class intervals used by the U.S. Census frequency tables for age. The knots for ω (median household income) were set regularly spaced on the interval $[16, 55]$ that contained the bulk of the observations, and padded with knots at the endpoints 14.17 and 65.91 of the support of ω . Knots were not placed between median household incomes of 55 and 65.91 (units of thousands of dollars) because of unavailability of data.

Equation (3) was first fit to the data using the η^* and z^* defined by equations (5) and (6) for the full model without an adjustment for age, that is, under the assumption $\psi = 0$. The starting value $\theta_o(u)$ equal to 7.3, 7.9, and 6.05 for $u = 1, 2, 3$, respectively, was taken from a parametric fit of constant prevalence in the reduced model. The initial estimate and the updated estimate of $\theta(u, \omega)$ without the age adjustment are shown in Figure 2 using dashed lines. Then, equation (3) was fit for the full model to the data using the starting values $\psi_o(v) = 0$ and $\theta_o(u, \omega) = \widehat{\theta}(u, \omega)$ obtained in the previous step. The final estimate of $\theta(u, \omega)$ from fitting the full model with an age adjustment is shown in Figure 2 using thick solid lines. Dimmit County was a highly influential point that was omitted from this analysis. Its inclusion affected the fit of the ln prevalence component for the ethnic/racial group "Others" by making it even steeper at lower

values of median household incomes. However, conclusions and general findings agree with and without this influential point.

The final nonparametric estimate of $\ln$ prevalence by ethnicity/race for ESRD are displayed again in Figure 3 (first three frames) along with plus/minus two pointwise standard errors (solid lines). Median household income of the county (ω) modulates the effect of ethnicity/race, and a linear model $\hat{\theta}$ across ω is not supported for Hispanics. The fourth frame in Figure 3 is the nonparametric adjustment $\hat{\psi}(v)$ to $\ln$ prevalence for age (v) (solid lines). This age adjustment requires integrating over the county's age distribution by ethnicity/race using the functional linear model features of the ecological regression model to reduce pure specification bias. A linear model for the age adjustment would appear to be questionable for ESRD. The fits in Figure 3 may now be used to predict the number of ESRD cases by ethnicity/race in El Paso County using equation (4). The expected counts are 717, 49, and 161 for Hispanic, African American, and Other, respectively.

The nonparametric estimates of $\ln$ prevalence for ESRD-DM along with plus/minus two pointwise standard errors are also shown in Figure 3 (dashed line). The difference between the ESRD and ESRD-DM prevalence rates in Figure 3 is smallest for Hispanics, consistent with the current belief that diabetes is more of a contributing factor to ESRD among Hispanics. The relative risk of old to young is highest for ESRD-DM, because the age adjustment has a steeper slope as compared with ESRD. While holding race fixed, the prevalence ratio of ESRD for an age of 68 years as compared with 20 years is about 2.6, but for ESRD-DM it is a 4.9 prevalence ratio of old to young.

African Americans have the highest age-adjusted prevalence of ESRD across all values of the county median household income. Age-adjusted ESRD (ESRD-DM) prevalence ratios of "Hispanic to Other" and "African American to Other" have maximum values of 2 and 8 (6 and 21), respectively, with the maximum occurring at the highest county median income. The reasons for this cannot be discerned from these data; nevertheless, the results are interesting in light of the findings of a national survey of households that described their ethnicity/race as black, Latino, or white (Haas et al., 2004). The survey results suggest that blacks and Latinos perceive fewer barriers to care when they live in a county with a high prevalence of people of similar ethnicity/race, whereas whites perceive more difficulty receiving care when they live in an area with a high prevalence of Latinos. Figure 3 is consistent with those survey findings in that ESRD prevalence rates are high for Hispanics and African Americans in counties with high median household income, whereas ESRD prevalence rates are high for Others in counties with low median household income. Median household income in Texas was shown to be highly associated with the distribution of ethnicity/race, that is, counties with high median household incomes have low representation of Hispanics and African Americans, whereas counties with low median household incomes have high representation of Hispanics.

Figure 4 is a plot of the component bias due to nonparametric smoothing. The rug plot of county median household income in Figure 4 displays only those counties with the highest (top 50%) ethnic/racial representation of Hispanic, African American, and Other, respectively, for each frame. Only the bias from nonparametric smoothing is quantified in Figure 4 because it was computed from equation (10.6) in Web Appendix B with a pilot estimate of the regression curves. Any other sources of bias, such as specification or aggregation bias from working with ecological data, are present in the pilot estimator used in place of the true underlying regression curve. Referring back to Figure 4 and paying attention to the rug plot, the component bias from smoothing is largest at those levels of median household income that have the lowest representation of the respective ethnic/racial group. Likewise the magnitude of the bias for the age adjustment is the smallest for ages with the highest representation, between 20 and 40 years of age; the bias is the largest near the lower and upper boundaries of age.

The advantages of the proposed method are demonstrated by fitting a parametric log-linear model in SAS `PROC GENMOD` (SAS Institute, Inc., 2003) to z with mean $n\eta$, $\eta = \exp[\beta_1 + \beta_2 p_2 + \beta_3 p_3 + \gamma_1 \omega + \gamma_2 p_2 \omega + \gamma_3 p_3 \omega]$, without an age adjustment by minimizing the Poisson deviance. The fitted parametric log-linear prevalence displayed in Figure 2 (straight solid line) is to be compared with the nonparametric fit without an age adjustment (dashed curve). The only terms that are not significantly different than 0 at the 0.05 level of significance are β_2 and γ_2 , suggesting that the prevalence rates for Hispanics and African Americans are similar, but different for Others. The prevalence for Others decreases with increasing median household income, but this is not so for Hispanics and African Americans. The fitted slope of median household income for Hispanics is positive, contradicting the locally decreasing relationship between $\ln \eta$ and median household incomes above \$30,000 (first frame of Figure 2, dashed curve, and solid straight line). There are two reasons for this contradiction: (1) The dependence of $\ln \eta$ on median household income changes direction from a positive to a negative locally linear relationship between \$25,000 and \$35,000; and (2) The counties with the highest representation of Hispanics have median household incomes near or below \$30,000, so the positive locally linear dependence of $\ln \eta$ on ω is driving the global fit. This illustrates that when the true underlying regression curve is nonlinear, both the magnitude and the sign of a globally fitted linear term can depend on the distribution of the covariate. Only the fitted slope of median household income for Others agrees with the nonparametric fit (third frame of Figure 2).

In the time that has passed since this investigation was begun, the ESRD counts by ethnicity/race have become publicly available through the U.S. Renal Data Extraction and Referencing System `RenDER`, a new online data querying application available through the USRDS website (usersds.org). Data obtained using `RenDER` from the United States Renal Data System website were used to compute the distribution of Hispanics, African Americans, and Others among ESRD cases. For the most part, the predicted proportions are underestimating the ones observed for Hispanics and African Americans; this is most likely from the negative bias in $\hat{\eta}$ due to smoothing. For example, in Figure 4 for Hispanics the bias of the age-adjusted prevalence rate of ESRD is negative for counties with median household incomes below roughly \$35,000, which are known to have a higher representation of Hispanics. A similar exploratory plot was not drawn for ESRD-DM because most of the `RenDER` counts by ethnicity/race were below the threshold level of confidentiality.

A limitation of this analysis is that the entire Texas population does not reside in the 80 reported counties equipped with Medicare-approved dialysis centers, so the estimated disease prevalence rates for 1998 may be somewhat inflated. This may have been inadvertently compensated for by using the 2000 population sizes instead of the smaller 1998 population sizes. The year 2000 population for these 80 Texas counties is roughly 18 million, whereas the 2000 population for all 254 Texas counties is closer to 21 million.

7. Discussion

This article proposes and demonstrates the use of an exploratory nonparametric regression model to estimate disease prevalence, given a sample of $R \times C$ tables with only marginal information. A unified approach is given for including marginal information on covariates available at both the level of the reporting unit and subject. The penalized weighted least squares approach undertaken for model fitting was advantageous because it allowed for the immediate application of O'Sullivan (1986). Explicit analytic expressions for the estimators were available. The averaging kernel development (O'Sullivan, 1986) allowed for the computation of the pointwise maximum squared bias of the nonparametric estimators of the components of the model, a term used as part of a data-driven smoothing parameter selection procedure; see Web Appendices B and C. Also, the plug-in estimate of bias from smoothing

taken together with the standard error of the estimators avoids cumbersome bootstrapping as in Staniswalis (2006).

The nonparametric methods in this article could be extended to apply locally over geographic location by incorporating an additional kernel smoothing step as in Nakaya et al. (2005). They fit parametric models for covariate adjustment to disease rates, but the coefficients in the log-linear models were estimated locally over geographic location. The validity of the additive regression model (equation 3) can only be checked with subject-level data. The general model was written for two variables available at the subject level, namely, age and ethnicity/race, and a third variable available at the county level, median household income. This is easily extended to either handle more variables in the additive model for the marginal covariates by conditioning, or to include more variables in the county-level covariates. Further, equation (3) has application to nonparametric modeling of covariates in a randomized response model (Scheers and Dayton, 1988), using the logistic transformation g , although such a demonstration is not carried out here.

8. Supplementary Materials

Background on specification and aggregation bias, as well as technical matters, such as quantifying the bias from nonparametric smoothing, standard error, and a data-based smoothing parameter selection procedure, are in Web Appendices A–C that are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author's effort on this project was partially supported by NIH-SCORE 2S06 GM08012.

References

- Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model. *Statistica Sinica* 2003;13:571–591.
- Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996;11:89–121.
- Eubank, RL. *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker, Inc; 1988.
- Fan J, Zhang W. Statistical estimation in varying coefficient models. *Annals of Statistics* 1999;27:1491–1518.
- Haas JS, Phillips KA, Sonneborn D, McCulloch CE, Baker LC, Kaplan CP, Perez-Stable EJ, Liang SY. Variation in access to health care for different racial/ethnic groups by the racial ethnic composition of an individual's county of residence. *Medical Care* 2004;42:707–714. [PubMed: 15213496]
- Hastie, TJ.; Tibshirani, RJ. *Generalized Additive Models*. London: Chapman & Hall; 1990.
- Hastie TJ, Tibshirani RJ. Varying-coefficient models. *Journal of Royal Statistical Society, Series B* 1993;55:757–796.
- King G. EI: A program for ecological inference. *Journal of Statistical Software* 2004;11:1–41.
- King, G.; Rosen, O.; Tanner, MA. *Ecological Inference: New Methodological Strategies*. Cambridge, U.K: Cambridge University Press; 2004.
- Lu Y, Mao S. Local asymptotics for B-spline estimators of the varying coefficient model. *Communications in Statistics: Theory and Methods* 2004;33:1119–1138.

- Mather FJ, White LE, Langlois EC, Shorter CF, Swalm CM, Shaffer JG, Hartley WR. Statistical methods for linking health, exposure, and hazards. *Environmental Health Perspectives* 2004;112:1440–1445. [PubMed: 15471740]
- McCullagh, P.; Nelder, JA. *Generalized Linear Models*. Vol. 2. London: Chapman & Hall; 1989.
- Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease associated mapping. *Statistics in Medicine* 2005;24:2695–2717. [PubMed: 16118814]
- Natário I, Knorr-Held L. Non-parametric ecological regression and spatial variation. *Biometrical Journal* 2003;6:670–688.
- O’Sullivan F. A statistical perspective on ill-posed inverse problems. *Statistical Science* 1986;1:502–517.
- Pazmiño PA, Pazmiño AK. Diabetic renal failure in Texas: Influence of ethnicity and household income. *Texas Medicine* 2003;99:57–65.
- Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. *Biometrika* 1995;82:113–125.
- Ramsay, JO.; Silverman, BW. *Functional Data Analysis*. New York: Springer; 1997.
- Rice J, Rosenblatt M. Smoothing splines: Regression, derivatives and deconvolution. *The Annals of Statistics* 1983;11:141–156.
- Richardson S, Stucker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: Some methodologic considerations. *International Journal of Epidemiology* 1987;16:111–120. [PubMed: 3570609]
- Rosen O, Jiang W, King G, Tanner MA. Bayesian and frequentist inference for ecological inference: The $R \times C$ case. *Statistica Neerlandica* 2001;55:134–156.
- SAS Institute, Inc. *PROC GENMOD*. Cary, North Carolina: SAS Institute, Inc; 2003. Version 9.1
- Scheers NJ, Dayton CM. Covariate randomized response models. *Journal of the American Statistical Association* 1988;83:969–974.
- Staniswalis J. On fitting generalized non-linear models with varying coefficients. *Computational Statistics & Data Analysis* 2006;50:1818–1839.
- Wakefield J, Salway R. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A* 2001;164:119–137.

ESRD Prevalence

ESRD-DM Prevalence

Ethnic Proportions

Age Density Histogram

Figure 1.

Exploratory plots of data for estimation of prevalence of end-stage renal disease in Texas.

Figure 2.

Components of ln-Prevalence (dashed lines, reduced model initial parametric estimate and update to include ω ; thick solid curve, full model final estimate including both ω and age-adjustment; solid straight line, parametric log-linear model). The spline knots are displayed in the rug plot.

Figure 3.

Full model components of ln-Prevalence + two pointwise standard errors (solid line, ESRD; dashed line, ESRD-DM). The spline knots are displayed in the rug plot.

Figure 4.

Plug-in estimates of bias for full model components of ln-Prevalence (solid line, ESRD; dashed line, ESRD-DM). The rug plot of county median household income displays the counties with the highest (top 50%) ethnic representation of Hispanic, African American, and Other, respectively, for each frame.

Table 1
ESRD data by ethnicity available for El Paso County

El Paso County Ethnicity	ESRD status		Row totals
	Cases	Controls	
Hispanic	—	—	531,464 (78.2%)
African American	—	—	18,350 (2.7%)
Other	—	—	129,808 (19.1%)
Column totals	839 (0.12%)	678,783 (99.88%)	679,622

Summary statistics for ESRD, median household income, and population for the $m = 80$ counties in Texas with Medicare-approved dialysis centers in 1998

Table 2

Variable	Five number summary for $m = 80$ counties				
	Min	Q_{25}	Median	Q_{75}	Max
ESRD counts Z	12	46	75	160	3461
ESRD-DM counts Z	4	16	31	58	1232
Income ω (thousands \$)	14	28	31	34	65
Population n (thousands)	8	38	77	189	3401