Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

COMPARISON OF TWO APPROACHES TO MODELING HEALTH AND THE ENVIRONMENT

Sandra A. Ham[1], Michelle M. Yore[1], Peter Mariolis[1]

ABSTRACT

Public health has a renewed interest in the impact of the environment on health. Ecologic or contextual studies ideally investigate these relationships using public health data augmented with environmental characteristics in multilevel or hierarchical models. In these models, individual health data are the first level and community data are the second. Most public health data use complex sample survey designs that require analyses accounting for clustering, nonresponse, and poststratification to obtain representative estimates of prevalence of health risk behaviors. This study uses the Behavioral Risk Factor Surveillance System (BRFSS), a state-specific U.S. health risk factor surveillance system conducted by the states in collaboration with the Centers for Disease Control and Prevention, which assesses health risk factors in more than 200,000 adults annually. BRFSS data are now available at the metropolitan statistical area (MSA) level and provide quality health information for studies of environmental effects. MSA-level analyses combining health and environmental data are further complicated by joint requirements of the survey sample design and the multilevel analyses. We compare two modeling methods in a study of physical activity and selected environmental factors using BRFSS 2000 data. Each of the methods described here is a valid way to analyze complex sample survey data augmented with environmental information, although each accounts for the survey design and multilevel data structure in a different manner and are thus appropriate for slightly different research questions.

Keywords: surveys, ecologic study, mixed models

1. INTRODUCTION

Public health has a vested interest in the impact of the environment on health (Sallis et al., 1998). Social ecologic theory conceptualizes relationships as persons nested within and affected by sets of contextual or environmental factors leading to hierarchically structured data models (Stokols, 1992; Stokols, 1996; Masse et al., 2002). Ecologic studies ideally investigate these relationships using public health data augmented with contextual effects in multilevel or hierarchical models. In these statistical models, health data from individuals (*i*) living in communities (*j*) are the first level and community data are the second level. Socioeconomic effects on health can be assessed at the individual and community levels as shown by $\text{Health}_{ij} = \text{SES}_{ij} + \text{Environment}_{j}$. Communities may be towns, cities, counties, metropolitan areas, health districts, or any other small geographic unit. Inferences are made by comparing communities across a broad spectrum of combinations of contextual effects. National public health datasets provide the largest sample of communities and are readily available. Hence, national public health datasets are suited to ecologic studies for hypothesis generation and identification of contextual correlates of health behaviors and outcomes.

Few studies (Chou et al., 2001) have combined national health data with contextual factors, such as public safety, climate, and the natural environment. National health surveys have complex sample designs, which require analyses accounting for the clustering, nonresponse, and poststratification to obtain representative estimates of health status and health risk behaviors. Community-level analyses combining health and environmental data are further complicated by joint requirements of the survey sample design and the multilevel analyses. Although statistical software designed for national survey analysis (e.g., SUDAAN and SAS) accounts for complex sample designs, multilevel modeling functions are not available; multilevel analysis software (e.g., MLwiN, HLM, and winBUGS) does not include complex sample design capabilities.

In this study, we compare modeling methods for ecologic public health research that combine the challenges of complex sample designs and data on contextual factors. Methods are presented in the context of an example using physical activity data from the Behavioral Risk Factor Surveillance System (BRFSS) 2000 from the United States.

[1] Centers for Disease Control and Prevention, 4770 Buford Hwy. NE, Atlanta, Georgia, USA 30341

Each of the methods accounts for the complex sample design of the BRFSS and is appropriate for generating hypotheses or evaluating correlates relating environmental contextual factors to health behaviors and outcomes. A discussion of the strengths and limitations includes additional modeling decisions regarding inclusion of correlates and appropriate research questions for each method.

2. METHODS

Our example models Recommended physical activity with three measures of individual socioeconomic status and three contextual measures of community characteristics as shown in Table 1. We chose one measure of contextual socioeconomic status, one measure of climate, and one measure of access to recreation areas. The following two sections describe the example data in more detail followed by presentation of the methods.

Table 1. Data sources for the physical activity example.*

Level	Variable	Source	N
Y	PA (Recommended physical activity)	BRFSS 2000	87,050
1	SEX, AGE (18-24, 25-34, 35-44, 45-54, 55-64, 65+), RACE (White, non-White)	BRFSS 2000	87,050
2	MALES (% Males in MSA)	U.S. Bureau of the Census	100
2	HEATINDEX (Heat index on July 15, 2 pm, average 1961-1990)	The Zunis Foundation	100
2	PARKS (No. of state and federal parks/100,000 population)	Places Rated Almanac	100

*BRFSS = Behavioral Risk Factor Surveillance System.

2.1 Health and Individual-Level Data

The BRFSS is a telephone survey used to track the prevalence of behaviors related to chronic diseases and preventive health practices among the civilian, noninstitutionalized population 18 years of age or older in each state in the United States. It is a joint venture of the Centers for Disease Control and Prevention (Atlanta, GA) and health departments in the 50 states, the District of Columbia, Guam, Puerto Rico, and the U.S. Virgin Islands. In the year 2000, more than 180,000 adults were surveyed with N = 87,050 in 100 census-defined primary metropolitan statistical areas and metropolitan statistical areas (MSAs). BRFSS sampling is state based; therefore, these MSAs included the most populous metropolitan areas in every state and in the District of Columbia; sample sizes for each MSA ranged from 300 to 7100. Survey weights for MSA-level analyses were available for these data. Additional information about the BRFSS may be found at http://www.cdc.gov/brfss.

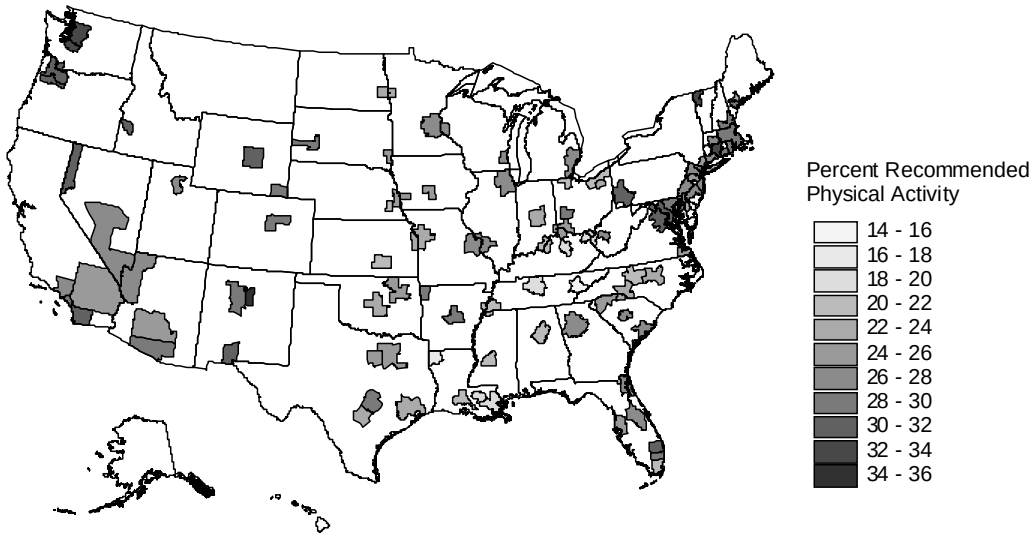
The response variable was a binary measure of meeting recommended leisure-time physical activity level (1 = yes; 0 = no). Respondents were classified as meeting recommendations if they reported participation in five or more days per week of leisure-time physical activity for at least 30 minutes per day, in three or more days per week of vigorous physical activity for at least 20 minutes per day, or in both (CDC, 2001). Prevalence of recommended physical activity for each of the 100 MSAs is shown in Figure 1. Individual-level age, sex, and race/ethnicity were used as level-1 data for this example.

2.2 Contextual Data

Contextual effects at the MSA level used for these examples were the percentage of males in the MSA (MALES), summer heat index (HEATINDEX), and access to parks (PARKS) as shown in Table 1. All data were obtained from relevant sources, input into spreadsheets, and centered around the respective grand means for the 100 MSAs used in the analysis. The percent males living in each MSA was obtained from the 2000 Census of the U.S. Bureau of the Census (http://www.census.gov). The Zunis Foundation (http://www.zunis.org) reported 30-year average heat indices for 2 pm on July 15 for cities using published data for hourly weather observations from 1961 through 1990 for 140 weather stations in the United States, Puerto Rico, and Guam. Weather observation data were compiled by the Solar and Meteorological Surface Observation Network and published by the U.S. National Oceanic and Atmospheric Administration (http://www.noaa.gov). City heat indices were assumed to represent average climate in respective MSAs. The names of state and federal parks in each MSA were obtained from a listing compiled from

federal agencies and published in the Places Rated Almanac (Savageau and D'Agostino, 2000, pp. 509-568). The numbers of all listed state and federal parks were summed and then divided by MSA population totals reported by the U.S. Bureau of the Census in the 2000 Census to calculate the number of state and federal parks per 100,000 population.

Figure 1. Prevalence of recommended physical activity* for 100 metropolitan statistical areas from the Behavioral Risk Factor Surveillance System 2000.



* Recommended physical activity = reported physical activity at least 5 times/week x 30 minutes/time or vigorous physical activity for at least 20 minutes at a time at least 3 times/week.

2.3 Method 1: Aggregated Model

The aggregated model is a two-stage linear regression of level-2, or community level, measures with prevalence of health behavior as the dependent variable and contextual measures as the independent variables. Stage 1 uses survey analysis software to calculate community prevalences using individual weights for community-level analyses. Stage 2 is the regression of the prevalences on the contextual measures using standard statistical software after merging all data into one dataset.

Let Y_j be the prevalence of the health behavior for community j with normal distribution $N \sim (\mu,\sigma^2)$. The equation for this model is $Y_j = \beta_0 + X\beta + e_j$, where X is a vector $X_1, \ldots X_k$ of level-2 covariates for community j, β is a vector $\beta_1, \ldots \beta_k$ of level 2 coefficients, and e_j is residual error assumed to be normally distributed $N \sim (0, \sigma^2)$. This model does not use any individual-level covariates; all variables are at the community level.

For our example, we first estimated the prevalence of individual respondents i who met the recommended physical activity levels in each MSA j using standard methods of accounting for the complex sample design using SUDAAN 8.0. The data were weighted and poststratified for each MSA. Next, the prevalence estimates were merged with a database of MSA-level contextual variables: percent males, heat index, and number of state and federal parks per 100,000 population. The level-2 model was

$$Y_j = \beta_0 + \beta_1(\text{MALES}) + \beta_2(\text{HEATINDEX}) + \beta_3(\text{PARKS}) + e_j. \quad (1)$$

SAS 8.2 was used for the linear regression analysis of the level-2 model.

2.4 Method 2: Multilevel Random-intercepts model

Next, we present a multilevel model in which the complex sample design was modeled (Korn and Graubard, 1999, pp. 177-181) using multilevel modeling software and each community had a unique random intercept defined by the contextual covariates. Survey weights for community-level analyses and sample design variables are incorporated, except when they are confounders as discussed below. We present a logistic version of this method because most health measures from surveys are categorical. Here we refer to the individual-level covariates as level 1 and the contextual covariates as level 2.

Let y_{ij} be the health behavior measure from a binomial distribution $y_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij})$ and π_{ij} be the probability that the *i*th survey respondent in the *j*th community reports the health behavior. Let X be a vector of level-1 covariates $X_{1ij},\ldots, X_{pij}$ for respondent *ij* and β be a vector of level-1 coefficients $\beta_{1j,\ldots,}\beta_{pj}$. Let Z be a vector $Z_{1j},\ldots, Z_{pj}$ of level-2 contextual covariates for community *j* and γ be a vector of level-2 coefficients $\gamma_1, \ldots, \gamma_p$. A general model for the probability of reporting the health behavior - given covariates is $P(X) = \Pr(Y = 1|X, Z)$.

The level-1 model for the probability of reporting a health behavior is $y_{ij} = \pi_{ij} + e_{0ij,}$ where the variance of level-1 random terms e_{0ij} is $\pi_{ij}(1 - \pi_{ij})$. The probability of the health behavior for respondent *ij* is a logit function :
$\text{logit}(\pi_{ij}) = \beta_{0j} + X\beta.$

We then add a second level to the model in which the intercept β_{0j} for an individual is a function of contextual or level-2 covariates measured at the community level in $\beta_{0j} = \gamma_0 + Z\gamma + :_{0j}$. Level-2 random effects $:_{0j}$ are normally distributed with mean 0 and variance equal to the between-community variance in health behavior as in $[:_{0j}] \sim N(0, \Omega_:)$, $\Omega_: = [\sigma^2_{:0}]$. Residual individual-level error is normally distributed with mean 0 and variance Ω_e as in $[e_{0ij}] \sim N(0, \Omega_e)$, $\Omega_e = [\sigma^2_{e0}]$.

All covariates may be centered around the grand means for simpler model interpretation (Raudenbush and Byrk, 2002, pp. 31-35). Cross-level effects or interactions between level-1 and level-2 covariates may be added to the multilevel model to test hypotheses about contextual effects for population subgroups. The percent of variance explained by the multilevel random-intercepts model compares the level-2 variance $\sigma^2_{:0}$ of an unconditional model of the survey design without covariates to the level-2 variance of models that include covariates.

Complex sample designs may be modeled as an alternative to the usual survey analysis methods. When modeling the design, the survey weights and all of the design variables that do not confound the analysis should be included in the model. Research questions requiring geographic units of analysis may be confounded with the geographic stratification in the complex sample design when the geographic unit of interest is similar to but different from the geographic strata. The problem is evident when the design variables explain much of the between-community variance before the contextual measures of interest are included in the model. When this confounding occurs, the design variables for geographic stratification are not used in the multilevel random-intercepts model.

For our physical activity example, we used a weighted model with standardized weights with mean of 1.0 for each MSA. The BRFSS complex sample design includes stratified sampling, poststratification, and weighting. Geographic strata variables were excluded because these variables confounded the interpretation of the results when using MSA as our level-2 unit of analysis. Our model for physical activity is

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_1(\text{SEX})_{ij} + \beta_2(\text{AGE})_{ij} + \beta_3(\text{RACE})_{ij} \quad (2)$$

$$\beta_{0j} = \gamma_0 + \gamma_1(\text{MALES})_j + \gamma_2(\text{HEATINDEX})_j + \gamma_3(\text{PARKS})_j + :_{0j}. \quad (3)$$

Dummy variables were created for categories of SEX, AGE, and RACE. Analyses were performed with the use of MLwiN 1.10 using first-order marginal quasi-likelihood estimation to ensure that the model would converge followed by second-order penalized quasi-likelihood estimation to obtain the least biased estimators. An extra-binomial distribution option was available but not used after testing the unconditional model and finding no evidence of over-dispersion or under-dispersion in this BRFSS dataset. If we were to use raw weights, the weighted

MSA population size would be a factor in the model and the data would have an extra-binomial distribution. Three models were run: an unconditional model with the design but no covariates, a model with individual effects, and a full model with individual and contextual effects.

Parameter estimates, standard errors, *P* values, and percent of between-MSA variance explained by the aggregated and multilevel random-intercepts models are shown in Table 2. Although parameter estimates cannot be directly compared, we can, however, illustrate several points using this table. Parameter estimates for individual-level covariates were not available for the aggregated model. The direction of contextual-effect estimates was consistent for both models, positive for MALES and PARKS, and negative for HEATINDEX. *P* values for the contextual effects differ between the models, leading to the conclusion that PARKS is a significant correlate from the aggregated model but that PARKS is not significant using the multilevel model. The magnitude of the individual effects is greater than the contextual effects in the multilevel model showing relative importance of individual versus contextual effects for recommended physical activity. The level-2 variance explained by the aggregated model was 31%, was 22% for the multilevel model compared with the unconditional model, and was 29% when compared with the model with individual effects.

Table 2. Parameter estimates for aggregated and multilevel random-intercepts models.[†]

	Aggregated Model (linear)		Multilevel Model (logistic)	
	β (SE)	*P* value	β (SE)	*P* value
Fixed effects				
SEX	NA	NA	-0.10 (0.02)	<0.0001
AGE (35-44y[*])	NA	NA	-0.23 (0.03)	<0.0001
RACE	NA	NA	-0.32 (0.02)	<0.0001
MALES	1.54 (0.47)	0.0014	0.09 (0.02)	0.0007
HEAT INDEX	-0.19 (0.06)	0.0028	-0.01 (0.00)	0.012
PARKS	1.03 (0.37)	0.0057	0.04 (0.02)	0.066
Variance components				
Between-MSA variance explained	31%		22% (full model vs unconditional model) 29% (full model vs individual effects model)	

[†]SE = standard error; NA = not applicable; MSA = metropolitan statistical area.
[*]Estimate for age group 35-44y compared with age group 18-24y is shown. Other age group estimates were less significant.

4. DISCUSSION

Two methods for using complex sample survey data in ecologic public health research were presented. The first method was an aggregated model for relating prevalence to contextual measures using all components of complex sample design. Then we presented a multilevel random-intercepts model using most of the components of the complex sample design and all of the available individual and contextual-level data. Each of the methods described here is a valid way to analyze complex sample survey data augmented with environmental information, although each accounts for the survey design and multilevel data structure in a different manner and is thus appropriate for slightly different research questions. There is no single, easy, ideal method or software package for all types of hypothesis and research questions. We used three software packages, SAS, SUDAAN, and MLwiN, for the analyses.

The aggregated model has strengths of being a simple technique that models prevalence of health effects and uses all survey sample design information. The effect estimates compare favorably with the multilevel model for hypothesis generation but should be used with caution. Limitations of the model include the potential for biased estimates, ignoring spatial autocorrelation, and ignoring the variability in prevalence estimates. The aggregated model limits the research questions that can be answered to those related to prevalence because individual-level variables cannot be included in the models.

The multilevel random-effects model is a conservative method that may be used for hypothesis generation and to identify correlates of individual health. Cross-level interactions allow for analyses of contextual effects for population subgroups. The data were treated hierarchically in a model of the probability of individual health effects in our logistic regression example. Limitations of this method include ignoring the geographic stratification in the complex sample design and ignoring spatial autocorrelation.

After deciding on a modeling method, additional consideration should be given to the potentially confounding effects of spatial autocorrelation inherent in the data. Spatial autocorrelation, or correlation between geographic areas due to close proximity to each other, may confound estimates generated from the models relating health to the environment. Survey data for communities is considered by spatial statisticians to have a lattice structure due to the units of analysis being areas that can be drawn on a map using polygons. Health survey data may include contiguous or noncontiguous geographic units. Whenever spatial analysis is required, the methods would be incorporated into the main model or as an added stage to the modeling process. SAS, MLwiN, and winBUGS have spatial analysis capabilities that are described in the software documentation and will not be discussed here but are left for future work.

Spatial analysis controls for correlation among geographic units due to close proximity and is used in combination with the aggregated or multilevel modeling techniques. Spatial patterns may have several interpretations and different analytic solutions. We might interpret the correlation with BRFSS data as simply that some MSAs near each other are more similar due to close spatial proximity than are MSAs farther apart, or we may interpret the clustering as the effects of regions of similar regional topography, climate, and culture. The former interpretation may lead one to control for spatial correlation, the latter interpretation may be a valuable inference from the analysis, and control for this correlation may be undesirable. Community-level health survey data may exhibit spatial correlation, and researchers should consider whether to perform spatial analysis based on the research question and interpretation of contextual effects with and without controlling for spatial correlation.

These methods are useful for public health research ecologic studies using complex survey data with contextual measures. The methods presented here have applications in public health research using both survey data and contextual measures of environmental factors. An aggregated model can relate contextual effects to prevalence across communities using traditional survey analysis software and methods. Multilevel analysis models the complex sample design in national health surveys to examine how environment affects individual health behaviors or outcomes. In addition, spatial analysis controls for the spatial correlation inherent in geographic data.

ACKNOWLEDGEMENTS

We thank the following persons for their input: Tom Schmid, William Kalsbeek, Fuzhong Li, Alastair Leyland, Michael Daniels, Constantine Gatsonis, and Danny Pfefferman.

REFERENCES

Centers for Disease Control and Prevention (2001), "Physical Activity Trends--United States, 1990-1998", *Morbidity and Mortality Weekly Report*, 50, pp. 166-169.

Chou, S.Y., M. Grossman, and H. Saffer (2001), "An Economic Analysis of Adult Obesity: Results from the Behavioral Risk Factor Surveillance System." Paper presented at the Third International Health Economics Association Conference, York, U.K.

Korn, E.L. and B.I. Graubard (1999), *Analysis of Health Ssurveys*. New York, NY: John Wiley & Sons.

Masse L.C., C. Dassa, L. Gauvin, B. Giles-Corti, and R. Motl (2002), "Emerging Measurement and Statistical Methods in Physical Activity Research", *American Journal of Preventive Medicine*, 23(2 Suppl), pp. 44-55.

Raudenbush S.W. and A.S. Bryk (2002), *The logic of hierarchical linear models. Hierarchical linear models: applications and data analysis methods*, Thousand Oaks, CA: Sage Publications.

Sallis, J.F., A. Bauman, and M. Pratt (1998), "Environmental and Policy Interventions to Promote Physical Activity", *American Journal of Preventive Medicine*, 15, pp. 379-397.

Savageau, D. and R. D'Agostino (2000), *Places Rated Almanac: Millenium Edition*, Foster City, CA: Macmillan General Reference USA.

Stokols D. (1992), "Establishing and Maintaining Healthy Environments: Toward a Social Ecology of Health Promotion", *American Psychology*, 47, pp. 6-22.

Stokols D. (1996), "Translating Social Ecologic Theory into Guidelines for Community Health Promotion", *American Journal of Health Promotion*, 10, pp. 282-298.