BIOMETRICS 55, 652–659
June 1999

Predictive Margins with Survey Data

Barry I. Graubard

Biostatistics Branch, EPS-8024, National Cancer Institute,
Bethesda, Maryland 20892, U.S.A.
email: BG1P@NIH.GOV

and

Edward L. Korn

Biometric Research Branch, National Cancer Institute,
Bethesda, Maryland 20892, U.S.A.

SUMMARY. In the analysis of covariance, the display of adjusted treatment means allows one to compare mean (treatment) group outcomes controlling for different covariate distributions in the groups. Predictive margins are a generalization of adjusted treatment means to nonlinear models. The predictive margin for group r represents the average predicted response if everyone in the sample had been in group r. This paper discusses the use of predictive margins with complex survey data, where an important consideration is the choice of covariate distribution used to standardize the predictive margin. It is suggested that the textbook formula for the standard error of an adjusted treatment mean from the analysis of covariance may be inappropriate for applications involving survey data. Applications are given using data from the 1992 National Health Interview Survey (NHIS) and the Epidemiologic Followup Study to the first National Health and Nutrition Examination Survey (NHANES I).

KEY WORDS: Adjusted mean; Adjusted treatment mean; Analysis of covariance; Logistic regression; Prediction; Sample weights; Survey methods; Survival analysis.

1. Introduction

It is frequently of interest to estimate the average response associated with different risk factors (or treatments) controlling for various covariate imbalances. Consider the analysis of covariance setting where $y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$, $i = 1, \ldots, R$, $j = 1, \ldots, n_i$, and with x_{ij} being the covariate for the jth individual in the ith treatment group. The adjusted treatment mean for group r is defined as $\bar{y}_r - \hat{\beta}(\bar{x}_r - \bar{x}) = \hat{\alpha}_r + \hat{\beta}\bar{x}$, where the hats represent least-squares estimators, $\bar{y}_r$ and $\bar{x}_r$ are means of the y and x observations in the rth group, and $\bar{x}$ is the mean of all the x observations (Snedecor, 1937, Section 12.4). One can interpret the adjusted treatment mean as the expected response for an individual in treatment group r with covariate value $X = \bar{x}$ or as the average predicted response if everyone in the sample had been in treatment group r.

For more complicated models, there are various ways to generalize the notion of a covariate-adjusted outcome. In particular, one can use a conditional or marginal approach, which disagree in nonlinear models. For example, for a simple linear logistic regression model, $\log\{P(y_{ij} = 1)/[1 - P(y_{ij} = 1)]\} = \alpha_i + \beta x_{ij}$, the conditional approach uses $\exp(\hat{\alpha}_r + \hat{\beta}\bar{x})/\{1 + \exp(\hat{\alpha}_r + \hat{\beta}\bar{x})\}$ as an estimator of the expected response for an individual conditional on his belonging to group r and having covariate value $X = \bar{x}$. Alternatively, one could use

$$\frac{1}{n}\sum_{i=1}^{R}\sum_{j=1}^{n_i}\exp(\hat{\alpha}_r + \hat{\beta}x_{ij})/\{1 + \exp(\hat{\alpha}_r + \hat{\beta}x_{ij})\}, \quad n = \sum_{i=1}^{R} n_i,$$

which is an estimator of the predicted response if all the observations had been treated with treatment r (Lee, 1981; Makuch, 1982). Lane and Nelder (1982) refer to such quantities as a predictive margin; Chang, Gelman, and Pagano (1982) describe why this marginal approach may be preferable to the conditional approach.

In this paper, we consider predictive margins estimated from survey data, the use of which adds two complications. The first is that covariate adjustments are possible to different sets of x's, e.g., the values in the population from which the data were sampled versus the values in an external population. Second, since the observed x's are from a sample and are therefore random, their variability needs to be taken into account when estimating standard errors of predictive margins. (This is unlike the experimental situation where the x's are fixed by the investigator.) As an example, consider the predictive margin displayed in column 5 of Table 1, estimated using the 1992 NHIS. This predictive margin, which is based on the logistic regression analysis given in Table 2, estimates

Table 1

Observed sample-weighted proportion and predictive margins for the probability of digital rectal examination as a function of type of health insurance plan; predictive margins control for age, family income, sex, race, education, marital status, and self-reported health status (see Table 2)

Health insurance[a]	Sample size	Observed proportion ± SE	Predictive margin ± SE (Model 1[b])	Predictive margin ± SE (Model 2[b])	Predictive margin ± SE (Model 2[b], pop. = None[c])
None	532	.13 ± .02	.16 ± .02	.14 ± .02	.13 ± .02
FFS (large)	1153	.34 ± .02	.33 ± .02	.33 ± .02	.27 ± .02
FFS (other)	867	.30 ± .02	.29 ± .02	.29 ± .02	.22 ± .02
HMO/PPO	813	.37 ± .02	.37 ± .02	.37 ± .02	.35 ± .03
Public	292	.30 ± .03	.36 ± .04	.45 ± .07	.35 ± .05

[a] Abbreviations indicate the following: None = no private or public health care coverage reported; FFS (large) = one of the 50 largest fee-for-service plans held privately or through employer; FFS (other) = fee-for-service plan held privately or through employer but not one of the 50 largest; HMO/PPO = enrolled in a health maintenance organization or preferred provider organization; Public = Medicaid or other public assistance program but not an HMO/PPO.

[b] Models refer to Table 2.

[c] Standardizing population is subpopulation of individuals who belong to the no health insurance group.

the proportion of individuals in the U.S. who would have an annual digital rectal examination if they had the designated type of health insurance. The standardizing distribution of covariates (ages, income, etc.) is the 1992 noninstitutionalized U.S. population, i.e., the sampled population. Using the (unweighted) covariate distribution of the sampled values would not be the same because individuals were sampled in the survey at differing rates. The standard errors displayed in Table 1 incorporate the fact that the covariate distribution of the population is not known but rather is estimated from the sampled data.

We give more details about this and another example in Section 4. Section 2 defines predictive margins in a general setting, Section 3 describes the estimation of their standard errors, and Section 5 ends with a brief discussion of why it is useful to present predictive margins in addition to estimated regression coefficients.

2. Estimation

We assume that there is a statistical model for the distribution of the response (y) as a function of the risk factor ($r \in \{1, \ldots, R\}$), a vector of covariates (x), and a vector of unknown parameters (θ). (If the risk factor is represented by a continuous variable, then r can refer to a discrete level of that variable.) We denote the quantity for which we wish to predict the margin by $g(r, x, \theta)$. For example, for predicting $\mathrm{E}(y)$ in an analysis of covariance, we have $g(r, x, \theta) = \alpha_r + \beta x$, and for predicting $P(y = 1)$ in a logistic regression, we have $g(r, x, \theta) = \exp(\alpha_r + \beta x)/[1 + \exp(\alpha_r + \beta x)]$; in both cases, $\theta = (\alpha_1, \ldots, \alpha_R, \beta)$ are the regression coefficients. In the nonsurvey setting with grouped data $\{(x_{ij}, y_{ij})\}$, the predictive margin for category r is defined by

$$PM(r) = \frac{1}{n} \sum_{i=1}^{R} \sum_{j=1}^{n_i} g(r, x_{ij}, \hat{\theta}), \tag{1}$$

where n is the total sample size and $\hat{\theta}$ is an estimator of the parameter vector, e.g., least-squares estimators in the analysis of covariance.

There are various generalizations of (1) that are useful in different applications involving survey data. As a general expression for the predictive margin, consider

$$PM(r) = \sum_{i=1}^{k} p_i g(r, z_i, \hat{\theta}), \quad \text{where } \sum_{i=1}^{k} p_i = 1 \text{ and } p_i \geq 0 \tag{2}$$

and $\hat{\theta}$ is an estimator of θ. The formula (1) is a special case of (2) with $k = n$, $p_i \equiv 1/n$, and the covariates $(z_1, \ldots, z_k) = (x_{11}, x_{12}, \ldots, x_{1n_1}, \ldots, x_{R1}, x_{R2}, \ldots, x_{Rn_R})$. With linear models and categorical covariates, the calculation of the predictive margin is a form of direct standardization (Kalton, 1968), with the p_i's determining the standardizing distribution. The population quantity we wish to estimate, which we shall call the population predictive margin, is

$$PPM(r) = \sum_{i=1}^{K} P_i g(r, Z_i, \theta), \quad \text{where } \sum_{i=1}^{K} P_i = 1 \text{ and } P_i \geq 0,$$

where $(Z_1, \ldots, Z_K)$ may or may not be the same as the $(z_1, \ldots, z_k)$, and the P_i are determined by the specific application. We now consider some special cases.

Case 1. Suppose we desire the predictive margin for the population from which the sample was taken or for a specific subpopulation of the sampled population. The population predictive margin is given by

$$PPM(r) = \sum_{i=1}^{N} \delta_i g(r, Z_i, \theta) \Big/ \sum_{i=1}^{N} \delta_i, \tag{3}$$

where N is the population size, $(Z_1, \ldots, Z_N)$ are the population values of the covariate, and δ_i equals one if the ith observation is in the subpopulation and zero otherwise. With

Table 2

Logistic regression of probability of digital rectal exam on age, family income, sex, race, education, marital status, self-reported health status, and type of health insurance using data from individuals between 40 *and* 64 *years of age sampled in the* 1992 *NHIS (sample size* = 3657,[a] *estimated population size* = 57.0 *million)*

	Model 1 (without interaction)			Model 2 (with interaction)		
Variable	Beta	Standard error	*p*-value	Beta	Standard error	*p*-value
Intercept	−3.81	.39	—	−4.18	.42	—
Age (years)	.033	.007	<.001	.033	.007	<.001
Family income (<20 K vs. ≥20 K)	−.24	.12	.048	.35	.30	NA
Sex (men vs. women)	−.53	.08	<.001	−.53	.08	<.001
Married vs. Not Married[b]	.09	.09	.30	.08	.09	.35
Race			.079			.063
White vs. Black	.37	.17		.38	.17	
Hispanic vs. Black	.22	.20		.24	.20	
Education			<.001			<.001
High school vs. <12 years	.26	.15		.28	.15	
>12 years vs. <12 years	.61	.15		.63	.15	
Health status			.085			.078
Fair/poor vs. excellent/very good	.22	.13		.21	.13	
Good vs. excellent/very good	.19	.100		.20	.10	
Health insurance			<.001			NA
FFS (large) vs. None	.98	.18		1.33	.27	
FFS (other) vs. None	.80	.20		1.18	.27	
HMO/PPO vs. None	1.15	.18		1.44	.27	
Public vs. None	1.11	.23		1.95	.47	
Health insurance × income[c]						.016
FFS (large) and <20 K				−.73	.35	
FFS (other) and <20 K				−.99	.37	
HMO/PPO and <20 K				−.23	.41	
Public and <20 K				−1.19	.51	

[a] Sample excludes 120 individuals of other races, 334 individuals with missing information concerning digital rectal exams, 85 individuals with missing covariates other than health insurance, and 69 individuals with military health insurance or missing health insurance information.

[b] Married is married with spouse in the household.

[c] Reference category is no health insurance and income ≥20 K.

sample survey data $\{(z_i, y_i),\ i = 1, \ldots, n\}$, each sampled individual has a sample weight (w_i) that effectively represents the number of people in the population that he represents. The predictive margin is given by

$$PM(r) = \sum_{i=1}^{n} \delta_i w_i g(r, z_i, \hat{\theta}) \Big/ \sum_{i=1}^{n} \delta_i w_i,$$

where $\hat{\theta}$ is a sample-weighted estimator of θ using the full sample. For example, for the analysis of covariance, $PM(r) = \hat{\alpha}_r + \hat{\beta}\bar{z}$, where $(\hat{\alpha}_1, \ldots, \hat{\alpha}_R, \hat{\beta})$ are weighted least-squares estimators using the full sample and

$$\bar{z} = \sum_{i=1}^{n} \delta_i w_i z_i \Big/ \sum_{i=1}^{n} \delta_i w_i.$$

The predictive margins in Table 1 are examples of Case 1 with a logistic regression model.

Case 2. Suppose we want to estimate the predictive margin for an external population for which we know the distribution of the covariates, which takes on S distinct values $Z_1, \ldots, Z_S$. Letting π_i equal the probability that $Z = Z_i$ in the external population, we have

$$PPM(r) = \sum_{i=1}^{S} \pi_i g(r, Z_i, \theta)$$

and

$$PM(r) = \sum_{i=1}^{S} \pi_i g(r, Z_i, \hat{\theta}),$$

where $\hat{\theta}$ is the sample-weighted estimator of θ using the sampled data. For example, for the analysis of covariance, $PM(r) = \hat{\alpha}_r + \hat{\beta}\bar{Z}$, where $(\hat{\alpha}_1, \ldots, \hat{\alpha}_R, \hat{\beta})$ are weighted least-squares estimators using the sample and $\bar{Z} = \Sigma_{i=1}^{S} \pi_i Z_i$. Historically,

demographers and epidemiologists have used this type of standardization in which death rates of one population are standardized to the age distribution of another population using 5-year age categories (Neison, 1844). Another example is given in Section 4 using a proportional hazards regression model and standardizing smoking/sex distribution from an external population.

Case 3. Suppose a simple random sample of observations is collected and we estimate the predictive margin for the sample distribution of the z's as

$$PPM(r) = \frac{1}{n}\sum_{i=1}^{n} g(r, z_i, \theta) \text{ and } PM(r) = \frac{1}{n}\sum_{i=1}^{n} g(r, z_i, \hat{\theta}).$$

Note that $PM(r)$ of Case 1 with $\delta_i \equiv 1$ reduces to this $PM(r)$ under simple random sampling (because $w_i \equiv 1$) but that $PPM(r)$ is different for the two cases. We believe that $PPM(r)$ of Case 1 is the more appropriate target parameter. This difference in target parameters has implications for the estimation of standard errors, as will be explained in Section 3.

An implicit assumption in the interpretation of (2) as a predictive margin is that the values of the covariates would be unaffected by assignment of individuals to different risk factor or treatment groups. The importance of not including variables "on the causal pathway" in regression models when estimating causal effects is well known (see Korn and Graubard (1995) for references), and since the predictive margin is predicting a mean if the group were changed, this caution carries over to the present situation. Assuming the covariates z are not causally affected by group, one *can* include interactions between components of z and the group.

We end this section by noting that the "obvious" choice of g for nonlinear models can sometimes be the wrong one. For example, suppose we are interested in the predictive margin for the mean of y, where $\log y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$ and the e_{ij} are independently and identically distributed with normal distributions with mean zero and variance σ^2. One might incorrectly assume that $g(r, x, \theta) = \exp(\alpha_r + \beta x)$ would be the right choice for g. However, note that $\mathrm{E}(y \mid r, x) = \exp(\alpha_r + \beta x + \sigma^2/2)$ since y has a lognormal distribution. Therefore, the correct choice of g to predict the mean response of y is $g(r, x, \theta) = \exp(\alpha_r + \beta x + \sigma^2/2)$, where $\theta = (\alpha_1, \ldots, \alpha_R, \beta, \sigma^2)$.

3. Standard Error Estimation

In the nonsurvey setting with the analysis of covariance $y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$, the variance of the predictive margin, which equals the adjusted treatment mean in this setting, is well known and given in textbooks (Neter, Wasserman, and Kutner, 1990, pp. 888–890) as

$$\mathrm{var}(\hat{\alpha}_r + \hat{\beta}\bar{x} \mid \{x_{ij}\}) = \sigma_e^2 \left(\frac{1}{n_r} + \frac{(\bar{x}_r - \bar{x})^2}{\sum_{i=1}^{R}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2} \right), \quad (4)$$

where σ_e^2 is the variance of the e_{ij}. This variance is actually a conditional variance, conditional on the set of observed $\{x_{ij}\}$, which is appropriate when the target parameter is $PPM(r)$ of Case 3. To compare the conditional variance (4) with the unconditional variance in the analysis of covariance setting, consider a simple random sample of observations in the analysis of covariance setting. The unconditional variance is given by

$$\begin{aligned}&\mathrm{var}(\hat{\alpha} + \hat{\beta}\bar{x})\\ &\quad = \mathrm{E}[\mathrm{var}(\hat{\alpha}_r + \hat{\beta}\bar{x} \mid \{x_{ij}\})] + \mathrm{var}[\mathrm{E}(\hat{\alpha}_r + \hat{\beta}\bar{x} \mid \{x_{ij}\})]\\ &\quad = \mathrm{E}[\mathrm{var}(\hat{\alpha}_r + \hat{\beta}\bar{x} \mid \{x_{ij}\})] + \beta^2\mathrm{var}(\bar{x}),\end{aligned}$$

where the second equality follows since $\mathrm{E}(\hat{\alpha}_r \mid \{x_{ij}\}) = \alpha_r$ and $\mathrm{E}(\hat{\beta} \mid \{x_{ij}\}) = \beta$. The difference in the unconditional and conditional variance estimators is approximately $\beta^2\mathrm{var}(\bar{x})$, which is not of small order compared to $\mathrm{var}(\hat{\alpha}_r + \hat{\beta}\bar{x} \mid \{x\text{'s}\})$. However,

$$\frac{\mathrm{E}[\mathrm{var}(\hat{\alpha}_r + \hat{\beta}\bar{x} \mid \{x_{ij}\})]}{\mathrm{var}(\hat{\alpha} + \hat{\beta}\bar{x})} \geq \left(1 + P_r \frac{R^2}{1 - R^2}\right)^{-1}, \quad (5)$$

where R^2 is the (population) multiple correlation coefficient and P_r is the proportion of the population in group r. The inequality in (5) is an approximate equality if the population mean of the X's in group r is equal to the overall population mean. The ratio of the variances will tend to be close to one unless R^2 is high. For example, with $P_r = 1/3$ and $R^2 = .3$ (.6), the right-hand side of (5) equals .87 (.67).

The textbook formula for the variance of an adjusted treatment mean is (4) with the least-squares estimate of σ_e^2 substituted for σ_e^2. As just described, even with a simple random sample, this formula is only valid when the inference is conditional on the sampled x_{ij}'s (Case 3) and not when the inference is for the population from which the sample was taken (Case 1). In survey applications, when an inference for a particular population is desired, the unconditional variance is appropriate, implying that the textbook variance formula should not be used even in the case of simple random sampling.

The difference between variance estimators for the adjusted treatment mean that are, and are not, conditional on the sampled x_{ij}'s is surprising because it is not seen for some other common regression parameters. For example, in a simple linear regression, $y_{ij} = \alpha + \beta x_{ij} + e_{ij}$, the unconditional variance of the slope and intercept can be expressed as

$$\mathrm{var}(\hat{\beta}) = \mathrm{E}[\mathrm{var}(\hat{\beta} \mid \{x\text{'s}\})] \quad \text{and} \quad \mathrm{var}(\hat{\alpha}) = \mathrm{E}[\mathrm{var}(\hat{\alpha} \mid \{x\text{'s}\})],$$

showing that the unconditional variance estimators [e.g., an estimator of $\mathrm{var}(\hat{\beta})$] would be expected to be close to the conditional variance estimators [e.g., an estimator of $\mathrm{var}(\hat{\beta} \mid \{x\text{'s}\})$].

To estimate the variance of $PM(r)$ for general $g(\cdot)$ and complex sampling designs, one can use Taylor series linearization. We sketch the approach for Case 1; details and a computer program for linear and logistic regression are available from the authors. Using

$$PM(r) \cong \left[\frac{1}{n}\sum_{i=1}^{n} \delta_i w_i\right]^{-1}$$

$$\times \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_i w_i g(r, z_i, \theta_0) + \frac{1}{n} \sum_{i=1}^{n} \delta_i w_i \left(\frac{\partial g(r, z_i, \theta)}{\partial \theta} \bigg|_{\theta_0} \right)' (\hat{\theta} - \theta_0) \right\},$$

a linearized variance estimator is given by $\widehat{\text{var}}[PM(r)] = \nu' S \nu$, where $\nu' = (1/\omega, -\hat{\gamma}/\omega^2, \hat{\Delta}'/\omega)$ and

$$\omega = \frac{1}{n} \sum_{i=1}^{n} \delta_i w_i,$$

$$\gamma = \frac{1}{n} \sum_{i=1}^{n} \delta_i w_i g(r, z_i, \theta),$$

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \delta_i w_i g(r, z_i, \hat{\theta}),$$

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \delta_i w_i \left(\frac{\partial g(r, z_i, \theta)}{\partial \theta} \bigg|_{\hat{\theta}} \right),$$

and S is the covariance matrix of $(\gamma, \omega, \hat{\theta})$, estimated using the survey design, evaluated at $\theta_0 = \hat{\theta}$.

A jackknife procedure is an alternative to linearization. Let the first stage of sampling involve the stratified selection of m_h primary sampling units (PSUs) from stratum h, $h = 1, \ldots, L$. Consider a new dataset that excludes the observations in the ith sampled PSU of the hth stratum and increases the sample weights of the retained observations in the hth stratum by a factor of $m_h/(m_h - 1)$. Let $PM(r; -hi)$ be the rth category of the predictive margin based on this new dataset. The jackknife variance estimator is given by

$$\widehat{\text{var}}_{JK}[PM(r)] = \sum_{h=1}^{L} \frac{m_h - 1}{m_h} \sum_{i=1}^{m_h} [PM(r; -hi) - PM(r)]^2.$$

In our experience, the linearization and jackknife variance estimators both behave reasonably well when $L - \Sigma m_h \geq 20$.

It should be noted that standard errors for adjusted treatment means are not needed for, nor should they be used for, testing group differences. These differences should be tested using the model parameters. For example, in the analysis of covariance, one would test $\alpha_1 = \alpha_2 = \cdots = \alpha_R$ using the appropriate covariance matrix of the α_i that takes into account the survey design. This is not to say that the presentation of standard errors for the predictive margin is useless; the situation is analogous to the analysis of variance where one would present standard errors for group means even though they would not be used for testing group differences.

4. Examples

Example 4.1. *Digital Rectal Exams and Type of Health Insurance Coverage*

This example was briefly discussed in Section 1. The probability an individual has had an annual digital rectal exam is important, as the American Cancer Society recommends annual digital rectal exams for individuals aged 40 or over for cancer screening (American Cancer Society, 1993). For public policy reasons, there is interest in the association of this probability with the type of health insurance the individual has (Potosky et al., 1998). Table 2 presents two logistic regression analyses of the probability on the type of health insurance and age, family income, sex, race, education, and self-reported health status. Model 1 contains only the main effects, while model 2 additionally contains the health insurance-by-income interaction. The data used for the analyses are from the Cancer Control Supplement to the 1992 National Health Interview Survey, a survey of the civilian noninstitutionalized population of the U.S. (Benson and Marano, 1994).

Based on model 1 in Table 2, one can see that the probability of having a digital rectal exam is lowest for those with no health insurance (since the base group is no health insurance) and highest for the health maintenance organization (HMO)/preferred provider organization (PPO) insurance group. We find that these differences are much easier to interpret by displaying the predictive margin in column 4 of Table 1. With the interaction (model 2), we find the improvement in interpretability offered by the predictive margin even larger (column 5 of Table 1). Additionally, as a statistical model builder, one might be interested in the effect of the inclusion of the interaction on the primary question. This is difficult to see from Table 2, but comparison of the predictive margins in Table 1 for the models shows that the major effect was to increase the predicted probability of the exams if everyone was using public insurance. The standard errors of the predictive margins in Table 1 were calculated using linearization.

The third predictive margin displayed in Table 1 addresses the question of predicted probability of digital rectal exams if the individuals with no insurance had instead one of the other types of insurance. This predictive margin was calculated by using as the population for the adjustment only those individuals with no insurance (Case 1). The interesting relative differences in the predictive margins for groups (e.g., the HMO/PPO and Public groups) are due to the fact that individuals with no insurance tend to have lower income than the population as a whole and that there is an income-by-group interaction included in model 2.

Example 4.2. *Lung Cancer Incidence and Size of Place of Residence*

There is interest in urban/rural differences in lung cancer rates (see Kafadar and Tukey [1993] for references). Because of changes in urban/rural smoking patterns over time (S. Devesa, personal communication, 1998), estimating urban/rural lung cancer rates adjusted for smoking is of scientific importance. Table 3 presents a proportional hazards regression analysis of the incidence of lung cancer on smoking status (current smoker, former smoker, never smoked), sex, and size of place of residence. For this semiparametric modeling, age rather than time from the baseline survey is used as the time scale (Korn, Graubard, and Midthune, 1997). The model is

$$\lambda(a \mid I_{s1}, I_{s2}, I_{\text{sex}}, I_{r1}, I_{r2}) = \lambda_0(a) \exp\{\beta_{s1} I_{s1} + \beta_{s2} I_{s2} + \beta_{\text{sex}} I_{\text{sex}} + \beta_{r1} I_{r1} + \beta_{r2} I_{r2}\},$$

where $\lambda(a \mid I_{s1}, \ldots, I_{r2})$ is the hazard of an individual at age a and the I's are dummy indicator variables corresponding to the categorical variables in Table 3. The analysis is cause

Table 3

Cause-specific proportional hazards regression of lung cancer incidence on smoking status at baseline, sex, and place of residence using data from the epidemiologic follow-up of NHANES I (sample size = 12,939,[a] estimated population size = 96.8 million)

Variable	Beta	Standard error	p-value
Smoking at baseline survey			<.001
Current vs. Never	1.87	.20	
Former vs. Never	.76	.21	
Sex (men vs. women)	.57	.16	<.001
Place of residence			.051
Urban ($<10^6$) vs. Urban ($\geq 10^6$)	.05	.16	
Rural vs. Urban ($\geq 10^6$)	.38	.17	

[a] Sample excludes 1355 individuals with missing smoking information and 13 individuals whose data of lung cancer diagnosis was undetermined or before the date of the baseline survey.

specific, so deaths are treated as censored observations. The data used for the analysis are from the 1992 follow-up of the NHANES I Epidemiologic Followup Study (Ingram and Makuc, 1994), which is a continuing followup of the NHANES I sample who were aged 25–74 years at the baseline survey. Therefore, the analyses presented in Table 3 and to be discussed below are conditional on the individuals not having lung cancer at age 25.

We consider the predictive margin for the cumulative incidence of lung cancer as a function of place of residence, controlling for sex and smoking, and consider standardization to two different populations (Table 4). One population is the 1971–1975 U.S. population aged 25–74 years as sampled by NHANES I. The other is 1992–1993 U.S. population aged 25–74 years as sampled by the Current Population Survey (CPS). The CPS is a continuing monthly survey of the U.S. For 1 month in 1992 and 2 months in 1993, the Tobacco Use Supplement acquired information on smoking on a total of 222,442 individuals, which is the basis of the estimated proportions in Table 4 (Shopland et al., 1996).

Table 5 displays the observed incidence and predictive margins for cumulative incidence for lung cancer at age 70. Although the predictive margin yields a cumulative incidence for each age ($\geq$ 25), we present only the age 70 incidence. The standard errors of the predictive margin were computed using a jackknife because no linearization variance formulas have been derived for this type of analysis. For calculating the standard errors when the standardizing distribution was from the 1992–1993 CPS, the sampling variability of the smoking-by-sex proportions in Table 4 was ignored since it is of small order because of the large sample size of the CPS compared to NHANES I. In Table 5, we note that the incidences in the predictive margin (NHANES I) are closer to each other than are the observed incidences. This is because fewer individuals smoked in rural areas than in the urban areas and fewer men smoked in the Urban ($\geq 10^6$) areas than in the Urban ($< 10^6$) areas (data not shown). The incidences in the predictive margin (1992–1993 CPS) are lower than in the predictive margin (NHANES I) since there was less smoking in 1992–1993 than in 1971–1975 (Table 4).

5. Discussion

The display of predictive margins can benefit a presentation of estimated regression coefficients associated with treatment or risk factor groups in several ways.

(1) The predictive margin may convey the scale of group differences better than regression coefficients. For example, the presentation of the probabilities of rectal exams in Table 1 may be easier to interpret than the log odds ratios in Table 2.

(2) With more than two groups, the display of group differences via regression coefficients of 0–1 dummy variables requires designating one of the groups as the baseline group, even though there may be no natural baseline group. Comparison of nonbaseline groups is then slightly inconvenient, requiring subtraction of regression coefficients. A predictive margin treats all the groups symmetrically, avoiding this problem. Predictive margins can also be used with a continuous treatment or risk variable by fixing that variable to be specific values.

(3) In some applications there is interest in the magnitude of the effects of inclusion of certain covariates in the model on group differences. By performing the analysis with and without the covariate, one can determine the changes in the predictive margin for each of the groups. For example, one may find that the inclusion of a covariate has a large effect on the predictive margin for group 1 but not for groups 2 and 3. One cannot easily see this by examining the regression coefficients.

Table 4

Estimated population proportions of smokers by sex for the U.S. population ages 25–74 years sampled at two different times with NHANES I and the Current Population Survey (CPS)

		Smoking status (%)		
		Never smoked	Former smoker	Current smoker
NHANES I (1971–1975)	Men	12.1	13.8	21.6
	Women	27.9	7.1	17.5
1992–1993 CPS	Men	20.5	14.1	13.4
	Women	29.3	10.6	12.1

Table 5

Observed cumulative cause-specific incidence and predictive margin for cumulative cause-specific incidence of lung cancer at age 70 as a function of place of residence; predictive margins control for smoking status at baseline and sex (see Table 3), standardizing to populations sampled by NHANES I (1971–1975) and the 1992–1993 Current Population Survey (CPS)

Place of residence	Sample size	Observed incidence[a] ± SE (%)	Predictive margin ± SE (NHANES I) (%)	Predictive margin ± SE (1992–1993 CPS) (%)
Urban ($\geq 10^6$)	3402	3.6 ± .6	4.0 ± .5	3.1 ± .4
Urban ($<10^6$)	4771	4.3 ± .7	4.2 ± .5	3.3 ± .4
Rural	4766	5.4 ± .8	5.8 ± .8	4.5 ± .6

[a] Observed incidence is based on a sample-weighted Fleming–Harrington estimator of the cause-specific survival distribution.

(4) In some applications, there is interest in the magnitude of the effects of different possible transformations of the dependent variable on group differences. These effects are difficult to interpret by examining regression coefficients (because they are on different scales with different transformations) but are easy to interpret with predicted quantities like predictive margins (Carroll and Ruppert, 1981).

(5) With group-by-covariate interactions in the model, the predictive margin allows one to display the overall group effect on the outcome. This effect is very difficult to see from the regression coefficients. Because of this difficulty, we suspect that many analysts inappropriately avoid including group-by-covariate interactions in their models.

(6) Applications involving the effects of group changes on outcome for specific populations are easily handled.

With appropriate consideration of the standardizing population of the covariate distribution, the presentation of predictive margins with their standard errors can be a useful addition to most analyses of group effects on outcome.

Acknowledgements

The authors thank C. E. McCulloch, E. Russek-Cohen, R. Williams, and a referee for their helpful comments and D. Midthune for his help with the computer programming.

Résumé

Dans l'analyse de la covariance, les moyennes des traitements ajustées permettent de comparer les moyennes par groupe (traitement) après contrôle des distributions des diifférentes covariables dans les groupes. Les marges "prédites" sont une généralisation des moyennes des traitements ajustées aux modèles non linéaires. La marge "prédite" pour le groupe r représente la réponse prédite moyenne si tout l'échantillon était dans le groupe r. Ce papier discute l'utilisation des marges prédites dans le cadre de données de survie complexes, où l'on s'intéresse principalement au choix de la distribution des covariables utilisée pour standardiser la marge "prédite." Il est suggéré que la formule classique pour l'écart-type de la moyenne des traitements ajustées à partir d'une analyse de covariance peut être inadaptée dans le cadre de données d'enquêtes. Des applications ont été réalisé à partir de données du National Health Interview Survey (NHIS) de 1992 et du Epidemiologic Followup Study to the first National Health and Nutrition Examination Survey (NHANES I).

References

American Cancer Society. (1993). *Guidelines for the cancer-related checkup.* Report 80-1MM-Rev.2/93-No.2070-LE. American Cancer Society, Atlanta.

Benson, V. and Marano, M. A. (1994). Current estimates from the National Health Interview Survey, National Center for Health Statistics. *Vital Health Statistics* **10,** 189.

Carroll, R. J. and Ruppert, D. (1981). On prediction and the power transformation family. *Biometrika* **68,** 609–615.

Chang, I.-M., Gelman, R., and Pagano, M. (1982). Corrected group prognostic curves and summary statistics. *Journal of Chronic Diseases* **35,** 669–674.

Ingram, D. D. and Makuc, D. M. (1994). Statistical issues in analyzing the NHANES I Epidemiologic Followup Study, National Center for Health Statistics. *Vital Health Statistics* **2,** 121.

Kafadar, K. and Tukey, J. W. (1993). U.S. Cancer death rates: A simple adjustment for urbanization. *International Statistical Review* **61,** 257–281.

Kalton, G. (1968). Standardization: A technique to control for extraneous variables. *Applied Statistics* **17,** 118–136.

Korn, E. L. and Graubard, B. I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A* **158,** 263–295.

Korn, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal followup of a survey: Choice of the time-scale. *American Journal of Epidemiology* **145,** 72–80.

Lane, P. W. and Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics* **38,** 613–621.

Lee, J. (1981). Covariance adjustment of rates based on the multiple logistic regression model. *Journal of Chronic Diseases* **34,** 415–426.

Makuch, R. W. (1982). Adjusted survival curve estimation using covariates. *Journal of Chronic Diseases* **35,** 437–443.

Neison, F. G. P. (1844). On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census. *Journal of the Royal Statistical Society of London* **7,** 40–68.

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Models,* 3rd edition. Homewood, IL: Irwin.

Potosky, A. L., Breen, N., Graubard, B. I., and Parsons, P. E. (1998). The association between health care coverage and the use of cancer screening tests. *Medical Care* **36,** 257–270.

Shopland, D. R., Hartman, A. M., Gibson, J. T., Mueller, M. D., Kessler, L. G., and Lynn, W. R. (1996). Cigarette smoking among U.S. adults by state and region: Estimates from the Current Population Survey. *Journal of the National Cancer Institute* **88,** 1748–1758.

Snedecor, G. W. (1937). *Statistical Methods, Applied to Experiments in Agriculture and Biology.* Ames, IA: Collegiate Press.

Received August 1997. *Revised May* 1998.
Accepted May 1998.