PREDICTIVE MARGINS WITH SURVEY DATA

Edward L. Korn, Barry I. Graubard, National Cancer Institute
Edward L. Korn, Biometric Research Branch, EPN-739, National Cancer Institute, Bethesda, MD 20892

Key Words: Adjusted mean; Adjusted treatment mean; Analysis of covariance; generalized linear model; prediction; sample weights; survey methods; survival analysis

1. Introduction

It is frequently of interest to estimate the average response associated with different treatments (or risk factors) controlling for various covariate imbalances. Consider the simplest 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 j th individual in the i th 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 r th group, and $\bar{x}$ is the mean of all the x observations (Neter, Wasserman and Kutner, 1990, pp 888-890). 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; Lane and Nelder, 1982; Chang, Gelman and Pagano, 1982; Makuch, 1982). Lane and Nelder (1982) refer to such quantities as a "predictive margin"; Chang et al. (1982) describe why this marginal approach may be preferable to the conditional approach.

In this paper, we consider predictive margins estimated from complex survey data. Covariate adjustments to different possible sets of x's are considered in section 2: to the values in the population or a subpopulation from which the data were sampled, to the values of a different population than the one from which the data were sampled, to the sampled values, and to a set of arbitrary values. Standard errors for the predictive margin are discussed in section 3. An important distinction in these derivations is whether the$x$ distribution is considered fixed or random; we consider both cases. To our knowledge, the only standard error formulas for predictive margins that have been available in the literature are in the non-survey setting with the x distribution considered fixed (e.g., Neter et al., 1990, pp 888-890; Gail and Byar, 1986; Flanders and Rhodes, 1987).

We give two applications in section 4: one using data from the first National Health and Nutrition Examination Survey (NHANES I) concerning blood pressure and the place of residence (urban, rural, etc.), and a second using data from the 1992 National Health Interview Survey (NHIS) as to whether the probability of having a digital rectal exam varies according to the type of health insurance a person has.

2. Estimation

We assume that there is a statistical model for the distribution of the response (y) as a function of the risk factor or treatment group ($r \in \{1,\ldots,R\}$), a vector of covariates (x), and a vector of unknown parameters (θ). We denote the quantity for which we wish to predict the margin by $g(r,x,\theta)$. For example, for predicting $E(y)$ 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 non-survey 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

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

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 $(z_1,\dots,z_k)=(x_{11},x_{12},\dots,x_{1n_1},x_{21},\dots,x_{Rn_r})$. The population quantity we wish to estimate, which we shall call the population predictive margin, is

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

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

Case 1: Suppose we desire to estimate the predictive margin for the population from which the sample was taken. The population predictive margin is given by

$$PPM(r)=\frac{1}{N}\sum_{i=1}^{N} g(r,Z_i,\theta) \qquad (3)$$

where N is the population size and $(Z_1,\dots,Z_N)$ are the population values of the covariate. With sample survey data $\{(z_i,y_i), i=1,\dots,n\}$, each sampled individual has a sample weight (w_i) which effectively represents the number of people in the population that he represents. The predictive margin is given by

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

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

Case 2: Suppose we desire the predictive margin for a specific subpopulation of the sampled population. If we let δ_i equal 1 if the ith observation is in the subpopulation, and 0 otherwise, then we have

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

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

where $\hat{\theta}$ is the sample-weighted estimator using the full sample.

Case 3: Suppose we desire to estimate the predictive margin for an external population for which we know the distribution of the covariates, which takes on S distinct values, $Z_1,\dots,Z_S$. Letting π_i equal the probability that $Z=Z_i$ in the external population, we have

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

where $\hat{\theta}$ is the sample-weighted estimator of θ using the sampled data.

Case 4: Suppose a simple random sample of observations is collected and we desire to estimate the predictive margin for the sample distribution of the z's, rather than the sampled population as in Case 1. Then

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

Although the predictive margin is the same as in case 2 ($w_i \equiv 1$), the population predictive margin is different. This has implications for the estimation of the standard errors of the predictive margin as will be explained in section 3.

Case 5: Suppose the data are acquired in an experiment in which the values of the covariates (z's) are fixed by the experimenter. One could adjust to the observed distribution of the z's as in Case 4, or possibly to some other meaningful distribution. For a linear model with only categorical covariates, the "population marginal mean" and "estimated marginal mean" of Searle, Speed and Milliken (1980) are given by

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

where k is the number of combinations of levels of the covariates and $\hat{\theta}$ is the least-squares estimator of θ.

3. Standard error estimation

In the non-survey 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, et al., 1990, pp 888-890):

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

where σ_e^2 is the variance of the e_{ij}. This variance is actually a conditional variance, conditional on the set of observed $\{x_{ij}\}$, a fact that is usually not explicitly stated. In this section, we compare the conditional variance (4) with the unconditional variance in this simple analysis of covariance setting.

Consider a simple random sample of observations in the analysis of covariance setting. The conditional variance is given by (4), the unconditional variance by

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

$$= E[Var(\hat{\alpha}_r + \hat{\beta}\bar{x} | \{x_{ij}\})] + \beta^2 Var(\bar{x})$$

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

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

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.

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 4 of section 2), and not when the inference is for the population from which the sample was taken (case 1). In some applications, adjusted treatment means are used as a convenient display of group differences, with the choice of standardizing population not being important. In this situation, one could argue that the variability of the x_{ij}'s need not be considered when estimating standard errors of adjusted treatment means (or a predictive margin). However, 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. In other applications, when an inference for adjusted treatment mean for a particular population is desired, the unconditional variance will be appropriate, implying that the textbook variance formula should not be used.

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 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

$Var(\hat{\beta}) = E[Var(\hat{\beta} | \{x's\})]$ and

$Var(\hat{\alpha}) = E[Var(\hat{\alpha} | \{x's\})]$

showing that the unconditional variance estimators (e.g., an estimator of $Var(\hat{\beta})$) would be expected to be close to the conditional variance estimators (e.g., an estimator of $Var(\hat{\beta} | \{x's\})$).

To estimate the variance of $PM(r)$ for general $g(\cdot)$ and complex sampling designs, one can use a Taylor series approximation around $\theta_0 = \lim \hat{\theta}_n$ or a jackknife procedure; details available from the authors.

4. Examples

In this section we present examples of using predictive margins involving linear and logistic regressions using data from NHANES I and the 1992 NHIS.

Example 4.1: Systolic blood pressure and size of place of residence

Table 1 presents a linear regression analysis of systolic blood pressure on size of place of residence (urban area with population one million or more, urban area with population under one million, rural area), age, body-mass-index, and sex. The data used for this analysis are from individuals 25 years or older sampled in NHANES I, which was a survey of the civilian noninstitutionalized population of the United States conducted in 1971-75 (Miller, 1973; Engel et al., 1978). The regression coefficients and their standard errors presented in Table 1 were computed taking into account the sample weights and clustering of NHANES I; the design can be approximated by the sampling of three PSU's from each of 35 strata (Ingram and Makuc, 1994). Table 2 displays the observed means and predictive margin for place of residence. Notice that because this is a linear regression, the differences between the predictive marginal values in Table 2 are equal to the corresponding regression coefficients in Table 1, e.g., 132.10-130.72 = 1.38. The predictive margin for the three place-of-residence groups is less spread out than the observed means, but the differences between the observed means and predictive margin are small.

The standard errors in Table 2 were computed as described in section 3 (case 1), and account for the sampling variability of the independent variables. (The relatively low standard error for the category "urban with $\geq$ million" is at first glance surprising given the relatively low sample size for this group, but is because the sampling design sampled many large urban areas with certainty.) One could argue for this example that the population values of the independent variables are not of great importance, implying that their sampling variability should not be taken into account. To see the effect of this sampling variability on the variability of the predictive margin, Table 3 computes the predictive margin pretending the sample was a simple random sample. The standard errors in this table are calculated three ways: allowing for the randomness of the x's (case 1 previously discussed), assuming the x's are fixed and allowing for heteroscedastic error variances in the linear regression (case 4), and assuming the x's are fixed and assuming homoscedastic error variances in the linear regression. The decreases in the standard errors when assuming the x's are fixed are small, but consistent with (5) since R^2 is .32 for this linear regression. If one included diastolic blood pressure as an additional independent variable, then $R^2 = .60$ and the difference in standard errors would be larger, e.g., .29 and .24 for "urban $\geq$ million" for the random and fixed x's standard errors, respectively.

Example 4.2: Digital rectal exams and type of health insurance coverage

The American Cancer Society recommends annual digital rectal exams for individuals aged 40 or over for cancer screening (American Cancer Society, 1993). Of interest is the association of the probability an individual has had an annual digital rectal exam with his type of health insurance; a full analysis including other types of cancer screening is given elsewhere (Potosky et al., in press). Table 4 presents two logistic regression analyses of this probability on the type of health insurance, 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 United States (Benson and Marano, 1994).

Based on model 1 in Table 4, 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 HMO/PPO insurance group. We find that these differences are much easier to interpret by displaying the predictive margin (Table 5). With the interaction (model 2), we find the improvement in interpretability offered by the predictive margin even larger. 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 4, but comparison of the predictive margins in Table 5 for the models shows that the major effect was to increase the predicted probability of the exams if everyone was using public insurance.

The third predictive margin displayed in Table 5 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 2 previously discussed). 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 the model 2.

Since in this example we are interested in inferences for the population, the standard errors for the predictive margins in Table 5 account appropriately for the variability of the independent variables (cases 1 and 2 of section 2).

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

Engel, A., Murphy, R. S., Maurer, K. and Collins, E. (1978). Plan and operation of the HANES I Augmentation Survey of adults 25-74 years, United States, 1974-1975. National Center for Health Statistics. Vital Health Stat 1(14).

Flanders, W. D. and Rhodes, P. H. (1987). Large sample confidence intervals for regression standardized risk, risk ratios, and risk differences. Journal of Chronic Diseases, 40, 697-704.

Gail, M. H. and Byar, D. P. (1986). Variance calculations for direct adjusted survival curves, with applications to testing for no treatment. Biometrical Journal, 28, 587-599.

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 Stat 2(121).

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.

Miller, H. W. (1973). Plan and operation of the Health and Nutrition Examination Survey, United States, 1971-1973. National Center for Health Statistics. Vital Health Stat 1(10a).

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

Potosky, A. L., Breen, N., Graubard, B. I. and Parsons, P.E. (in press). Health care insurance coverage and the use of cancer screening examinations in the U.S., Medical Care (in press).

Searle, S. R., Speed, F. M. and Milliken, G. A. (1980). Population marginal means in the linear model: an alternative to least squares means. American Statistician, 34, 216-221.

Table 1: Linear regression of systolic blood pressure on age, body-mass-index (kg/m^2), sex, and place of residence using data from NHANES I (sample size=14333, estimated population size=104.7 million)

Variable	Beta	Standard Error	P-value
Intercept	68.50	1.32	---
Age	0.68	0.02	<.001
Body-mass-index	1.24	0.05	<.001
Sex (Men vs Women)	1.93	0.46	<.001
Place of residence			.064
Urban($<10^6$) vs Urban($\geq 10^6$)	1.38	0.69	
Rural vs Urban ($\geq 10^6$)	1.40	0.67	

Table 2: Observed sample-weighted mean and predictive margin for systolic blood pressure as a function of place of residence; predictive margin controls for age, body-mass-index and sex (see Table 1)

Place of residence	Sample size	Observed mean ± SE	Predictive margin ± SE
Urban ($\geq 10^6$)	3907	130.17 ± 0.44	130.72 ± 0.38
Urban ($<10^6$)	5291	132.11 ± 0.72	132.10 ± 0.63
Rural	5135	132.61 ± 0.70	132.11 ± 0.64

Table 3: Predictive margin for systolic blood pressure as a function of place of residence, controlling for age, body-mass-index and sex, treating the sample as a simple random sample

		Standard errors calculated with:		
Place of residence	Predictive margin	x's random	x's fixed	x's fixed, σ_ϵ^2 constant
Urban($\geq 10^6$)	133.22	0.35	0.32	0.32
Urban($<10^6$)	134.63	0.30	0.28	0.28
Rural	134.39	0.30	0.28	0.28

Table 4: 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, estimated population size=57.0 million); only coefficients for Family income, Health Insurance, and Health Insurance X Income are displayed in this table.

	Model 1 (without interaction)			Model 2 (with interaction)		
Variable	Beta	Standard Error	P-value	Beta	Standard Error	P-value
...	...	...	...	...	...	...
Family income (<20K vs ≥20K)	-.24	.12	.048	.35	.30	N.A.
Health Insurance[a]			<.001			N.A.
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 X Income[b]						.016
FFS(large) & <20K				-.73	.35	
FFS(other) & <20K				-.99	.37	
HMO/PPO & <20K				-.23	.41	
Public & <20K				-1.19	.51	

(a) Abbreviations for types of health insurance are: None = no private or public health care coverage reported; FFS (large) = one of the 50 largest fee-for-service plan 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; and Public = Medicaid or other public assistance program, but not a HMO/PPO

(b) Reference category is Health Insurance = "None" and Income = "≥20K"

Table 5: 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

Health Insurance	Sample size	Observed prop. ± SE	Predict. Margin ± SE (Model 1)	Predict. Margin ± SE (Model 2)	Predict. Margin ± SE (Model 2, pop.=None[b])
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) Standardizing population is subpopulation of individuals who belong to the health insurance="None" group.