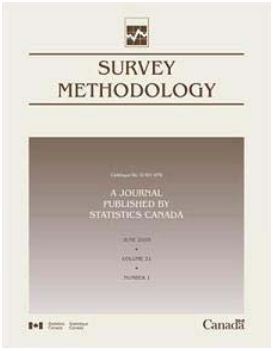
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates

by Alan M. Zaslavsky, Hui Zheng and John Adams



June 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 65-78
Statistics Canada, Catalogue No. 12-001-X

Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates

Alan M. Zaslavsky, Hui Zheng and John Adams [1]

Abstract

We consider optimal sampling rates in element-sampling designs when the anticipated analysis is survey-weighted linear regression and the estimands of interest are linear combinations of regression coefficients from one or more models. Methods are first developed assuming that exact design information is available in the sampling frame and then generalized to situations in which some design variables are available only as aggregates for groups of potential subjects, or from inaccurate or old data. We also consider design for estimation of combinations of coefficients from more than one model. A further generalization allows for flexible combinations of coefficients chosen to improve estimation of one effect while controlling for another. Potential applications include estimation of means for several sets of overlapping domains, or improving estimates for subpopulations such as minority races by disproportionate sampling of geographic areas. In the motivating problem of designing a survey on care received by cancer patients (the CanCORS study), potential design information included block-level census data on race/ethnicity and poverty as well as individual-level data. In one study site, an unequal-probability sampling design using the subjectss residential addresses and census data would have reduced the variance of the estimator of an income effect by 25%, or by 38% if the subjects' races were also known. With flexible weighting of the income contrasts by race, the variance of the estimator would be reduced by 26% using residential addresses alone and by 52% using addresses and races. Our methods would be useful in studies in which geographic oversampling by race-ethnicity or socioeconomic characteristics is considered, or in any study in which characteristics available in sampling frames are measured with error.

Key Words: Descriptive population quantity; Measurement error; Neyman allocation; Regression models; Sample design; Surveys.

1. Introduction

A sample survey is to be designed to obtain data that will be used to estimate coefficients of one or more regression models. Information about the population distribution of the covariates is available, and also some covariate information is available in the sampling frame. How can this information be used to make the survey design more efficient? How much can variance be reduced with such a design, relative to simple random sampling, and how is that answer affected if the frame only provides covariate distributions aggregated over groups, but not for individual subjects?

These questions were motivated by design of a survey of health care processes (such as provision of chemotherapy when appropriate) and outcomes (such as quality of life after treatment) for a large sample of cancer patients at seven sites in the United States, conducted as part of the CanCORS (Cancer Care Outcomes Research and Surveillance) study (Ayanian, Chrischilles, Wallace, Fletcher, Fouad, Kiefe, Harrington, Weeks, Kahn, Malin, Lipscomb, Potosky, Provenzale, Sandler, Vanryn and West 2004). Among the primary objectives of this study was to estimate joint effects of race and income on these measures, using regression models that include both of these patient characteristics. However, only limited data were available when patients were sampled for enrollment in the study. Prior experience suggested that race and residential address might be determined with reasonable accuracy at the time cases were ascertained for possible study recruitment, but income could not be determined until the subject was recruited and interviewed, and could not practically be collected in a screening interview. We undertook the research reported here to determine how the available patient data could be combined with census data on race-income distributions in census blocks to sample patients disproportionately and thereby improve estimates of race and income effects.

Such concerns arise frequently when survey data will be used to estimate coefficients of one or more regression models. For example, the National Health Interview Survey (NHIS) uses geographical oversampling together with a screening interview to oversample Black and Hispanic respondents for improved domain estimation (Botman, Moore, Moriarity and Parsons 2000, page 12); NHIS data have been used extensively in regression analyses, of which domain estimation is a special case. Sastry, Ghosh-Dastidar, Adams and Pebley (2005, pages 1013-1014) oversampled census tracts by minority composition, using simulations to evaluate the power of various designs for regression analyses of interest. The Youth Risk Behavior Surveillance

1. Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.; Hui Zheng, Biostatistics Center, Massachusetts General Hospital, 50 Staniford Street, Boston, MA 02114, U.S.A.; John Adams, RAND Corporation, 1776 Main Street, Santa Monica, CA 90401.

System oversamples *schools* in high-minority PSUs to improve precision of estimates for minority racial/ethnic groups (Eaton, Kann, Kinchen, Ross, Hawkins, Harris, Lowry, McManus, Chyen, Shanklin, Lim, Grunbaum and Wechsler 2006, pages 2-3).

The literature on optimal design of experiments is extensive. Design objectives for surveys, however, differ in important ways from those for experiments, in which the researcher can arbitrarily assign *a priori* identical units to treatments. A strongly model-based approach to estimation of regression coefficients would suggest selection of a suitable set of high-leverage observations, much as in design of an experiment (Royall 1970), but the application of these principles to survey design is controversial (see Sec. 4). The design-based approach requires a sample that is representative, through a known probability mechanism, of a defined population; intermediate positions are also possible (Sec. 2.5, 3.4). From this perspective, the sampler is not free to select, for example, 100 white respondents from a convenient primarily white neighborhood and 100 Hispanic respondents from a convenient primarily Hispanic neighborhood and call the sample "representative" for estimating differences between whites and Hispanics. The objects of design-based inference are quantities that describe the population; in the case of regression we will refer to "descriptive population quantity" or DPQ regressions (Pfeffermann 1993, pages 319-321).

Since Neyman (1934), the extensive literature on optimal design of surveys (reviewed in standard texts like Cochran 1977 or Särndal, Swensson and Wretman 1992) has primarily focused on estimation of simple quantities such as a mean or ratio, or of several such quantities (Kish 1974; Bellhouse 1984; Chromy 1987). Although variance estimation for design-based estimates of regression coefficients has received considerable attention (Fuller 1975; Fuller 1984; Binder 1981; Binder 1983), relatively little attention has been given to the corresponding optimal sample designs. (Regression-*assisted* estimation of a mean (Cassel, Särndal and Wretman 1976; Särndal, Swensson and Wretman 1992, Sec. 12.2) is a distinct problem.)

Furthermore, characteristics that might be used to define an unequal-probability sampling scheme are likely to be recorded with error in sampling frames, because they are based on aggregated data or because the characteristics associated with a unit (such as an address or a household) change over time. Such errors can greatly affect the efficiency of a putatively optimal sampling scheme; see Morris, Newhouse and Archibald (1979, Sec. III) on stratified sampling for domain estimation and Thomsen, Tesfu and Binder (1986) on probability-proportional-to-size sampling. Waksberg (1973, 1995) considers stratification by census blocks on a single aggregated characteristic for estimation of means for domains such as racial/ethnic groups or the poor, with or without a subsequent screening interview.

Our objective in this article is to describe optimal designs for samples that will be used in DPQ (design-weighted) regression analysis, in the sense of minimizing the weighted sum of variances of some preselected linear combinations of regression coefficients. We also consider some classes of estimands and corresponding estimators that depart from the DPQ approach to improve efficiency. In Section 2, we establish notation and derive optimal sampling rates for DPQ regression under scenarios representative of the individual and area-level information that might be encountered in population surveys with imperfect frames. We first assume that exact design information is available in the sampling frame and then generalize to situations in which some variables are available only as aggregates for subdomains or from inaccurate data. We next consider optimal estimation of combinations of coefficients from more than one model and of flexible combinations of coefficients. In Section 3 we estimate the potential benefits of these methods for a survey in the CanCORS study sites, using block-level census data on race/ethnicity and poverty. Finally, in Section 4 we consider the relevance of the DPQ approach and possible extensions of the methodology.

2. Optimal design calculations

2.1 Notation

Suppose that the target population is divided into cells indexed by $b = 1, 2, ..., B,$ with elements indexed by $k = 1, 2, ..., K_b$ in cell b. With each element is associated a covariate vector $\mathbf{x}_{bk}$ with $\mathbf{x}'_{bk} = (\mathbf{u}'_{bk}, \mathbf{t}'_{bk})$, where $\mathbf{u}_{bk}$ is the component observed for identifiable individuals. The distribution of $\mathbf{t}_{bk}$ in each cell is known but the values for individuals are not observed; thus the cell is the unit of aggregation for some or all of the design variables. Hence we know the finite population values $\mathbf{T}_b = (\mathbf{t}_{b1}, \mathbf{t}_{b2}, ..., \mathbf{t}_{bK_b})'$ but cannot identify the rows with individuals. Define $\bar{\mathbf{t}}_b = 1'\mathbf{T}_b / K_b,$ the mean of $\mathbf{t}$ in cell b.

Associated with sampling each element is a cost c_{bk}. A sampling plan is defined by assigning a probability of selection π_{bk} to each element. Assume a constraint on expected cost,

$$\sum_{b,k} c_{bk}\,\pi_{bk} \le C. \qquad (1)$$

To simplify the presentation, we also assume that the sampling rate is low and potential benefits of stratification are minimal, so the design can be described approximately as unstratified unequal-probability sampling with replacement. We also assume single-stage element sampling. The

population is $U = \{(b, k): b = 1, 2, ..., B; k = 1, 2, ..., K_b\}$ and a sample is $S \subset U$.

The population-descriptive ordinary least squares (OLS) regression coefficient, corresponding to the model $y_{bk} = \boldsymbol{\beta}'\mathbf{x}_{bk} + \varepsilon_{bk}$ with $\varepsilon_{bk} \sim [0, \sigma^2]$, is $\boldsymbol{\beta}_U = (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{X}_U' \mathbf{y}_U$, where subscript U signifies matrices or vectors corresponding to the entire population. (Here $[0, \sigma^2]$ signifies a distribution with mean 0 and variance σ^2, but unspecified form.) Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_S' \mathbf{W}_S \mathbf{X}_S)^{-1} \mathbf{X}_S' \mathbf{W}_S \mathbf{y}_S \qquad (2)$$

is the usual design-based estimator of $\boldsymbol{\beta}$, where S signifies that only the rows corresponding to the sample are included, and $\mathbf{W}$ is the diagonal matrix of weights $1/\pi_{bk}$.

To *design* the survey, we must make some assumptions about the distribution of outcomes y_{bk}, even if we would not rely on the same assumptions in *analysis* of the data. Specifically, we assume that the outcomes are generated by a model $\xi: y_{bk} = \mathbf{x}_{bk}'\boldsymbol{\beta} + \varepsilon_{bk}$, with independent $\varepsilon_{bk} \sim [0, \sigma_{bk}^2]$ and known σ_{bk}^2 (up to a constant factor). Note that the distributions of the design variables $\mathbf{x}_{bk}$ and the residuals are relevant to optimization of the design, but the value of $\boldsymbol{\beta}$ is not since it does not affect the variance of the regression estimators. Furthermore, the assumption of independent residuals from a regression model might be more reasonable then independence of data values. We allow for heteroscedasticity, even when fitting an OLS model. OLS coefficients (including special cases such as the overall mean or domain means) are often useful descriptive statistics even if the OLS model does not actually hold, but if information about heteroscedasticity is available it can be used to make the design more efficient.

2.2 Optimal DPQ regression design with individual-level variables only

Consider first the case in which $\mathbf{t}$ is empty, so $\mathbf{x}_{bk} = \mathbf{u}_{bk}$, reflecting a scenario in which all relevant design variables (race and income in our CanCORS design) are available to the researcher before sampling. Since the cells now consist of single cases we drop the subscript b, writing $\hat{\boldsymbol{\beta}} = (\sum_S w_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_S w_k \mathbf{x}_k y_k)$. Then for any fixed linear combination of coefficients with weights $\mathbf{a}$, assuming that the first factor is a design-consistent estimator (after scaling) of $N(\mathbf{X}_U' \mathbf{X}_U)^{-1}$, we have the expectation under sampling of the model-based variance (White 1980) of the estimator,

$$\begin{aligned} V_a &= E_\pi \operatorname{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}} \\ &\approx \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(E_\pi \operatorname{Var}_\xi \sum_{k\in S} \mathbf{x}_k y_k / \pi_k\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a} \\ &= \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(\sum_{k\in U} (\sigma_k^2/\pi_k)\mathbf{x}_k \mathbf{x}_k'\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}. \end{aligned} \qquad (3)$$

For design-based inference, the relevant measure is the average variance under the sampling design over possible populations obtained under the model ξ, $E_\xi \operatorname{Var}_\pi \mathbf{a}'\hat{\boldsymbol{\beta}}$ (the "anticipated variance" of Isaki and Fuller 1982; see also Bellhouse 1984, sec. 1); this quantity is approximately equal to the expected model-based variance (see Appendix for proof and asymptotic conditions).

By the typical Lagrange multiplier argument for optimal allocation problems (*e.g.*, Valliant, Dorfman and Royall 2000, pages 169-170), V_a is minimized subject to the expected cost constraint (1) when $\partial V_a / \partial \pi_k = c_k \lambda$ for some constant λ and all k, so $\pi_k \propto \sigma_k(\mathbf{a}'\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{x}_k)/\sqrt{c_k}$. Thus the optimal sampling rate is higher for cases with greater model variance and lower cost (as in the usual case of estimation of a mean) and also for cases with greater leverage in the regression. This result differs from the standard model-based calculations for optimal experimental design, which would allocate the entire sample to a few high-leverage design points. The design-consistent estimator of the DPQ regression does not assume the correctness of the model and therefore requires that every case have a positive probability of selection. Thus, for estimation of a ratio β under a homoscedastic model $y_k = \beta x_k + \varepsilon_k$, model-based estimation would suggest selection of the units with the largest values of x, but our probabilities of selection are proportional to x.

Typically, more than one estimand will be of interest in a study; CanCORS is intended to estimate both race and income effects. We generalize (3) to simultaneous estimation of several linear combinations of coefficients by optimizing a weighted sum of variances $V = \sum_i d_i V_{\mathbf{a}_i}$, where i indexes the estimands. By the same arguments the optimal sampling probabilities for this objective are

$$\pi_k \propto \sigma_k \left(\sum_i d_i (\mathbf{a}_i'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{x}_k)^2 / c_k\right)^{1/2}. \qquad (4)$$

With some choices of the $\{\mathbf{a}_i\}$, strict adherence to (4) could lead to arbitrarily small π_{bk} (and hence arbitrarily large weights) for cases with leverage approaching zero. To prevent this, we could set a positive floor on the π_k. Alternatively, by making estimation of the population mean one of the objectives (Section 2.4), we guarantee sampling with positive probability over the entire population. Either method makes the design more robust against error in the approximate calculation of leverage and better prepared for possible post hoc decisions to estimate quantities not foreseen in the original design plan (Section 2.6). Furthermore, reasonably good estimation of means is needed to guarantee design-consistency of the first factor of (2).

2.3 Optimal design with individual- and aggregate-level variables

Now suppose that the covariate vector $\mathbf{t}_{bk}$ is nonempty and $\mathbf{u}_{bk}$ is constant in each cell, as when aggregated design information is available for cells corresponding to covariate classes of $\mathbf{u}$ within blocks. In CanCORS, if race ($\mathbf{u}$) but not income ($\mathbf{t}$) is known for individual subjects, and income distributions are available for each race in each census block, we would define cells to consist of people of a single race in a single census block.

Since cases in the same cell cannot be distinguished on covariates we further assume that $\sigma_{bk} = \sigma_b$ and $c_{bk} = c_b$ are *a priori* constant across the cell, so the optimal design also makes $\pi_{bk} = \pi_b$ constant in each cell.

We can now rewrite (3) as

$$V_a \approx \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(E_\pi \mathrm{Var}_\xi \sum_S \mathbf{x}_{bk} y_{bk}/\pi_b\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}$$

$$= \mathbf{a}'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\left(\sum_{b,k}(\sigma_b^2/\pi_b)\mathbf{S}_b\right)(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}, \qquad (5)$$

where

$$\mathbf{S}_b = \begin{pmatrix} \mathbf{u}_b\mathbf{u}_b' & \mathbf{u}_b\bar{\mathbf{t}}_b' \\ \bar{\mathbf{t}}_b\mathbf{u}_b' & \mathbf{S}_{Tb} \end{pmatrix}$$

is the matrix of mean squares and crossproducts in cell b, with $\mathbf{S}_{Tb} = \mathbf{T}_b'\mathbf{T}_b/K_b$. The optimal sampling probabilities corresponding to (4) are then

$$\pi_b \propto \sigma_b\left(\sum_i d_i\,\mathbf{a}_i'(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{S}_b(\mathbf{X}_U'\mathbf{X}_U)^{-1}\mathbf{a}_i/K_b c_b\right)^{1/2}. \qquad (6)$$

If $\mathbf{t}$ is measured through a census of each cell, then $\bar{\mathbf{t}}_b$ and S_{Tb} are known exactly. The same principles apply, however, if S_{Tb} is not directly observed but instead is estimated under a model ζ. We then replace $\bar{\mathbf{t}}_b$ and $\mathbf{S}_b$ in (5) with predictive expectations $\tilde{\bar{\mathbf{t}}}_b = E_\zeta\,\bar{\mathbf{t}}_b$ and $\tilde{\mathbf{S}}_b = E_\zeta\,\mathbf{S}_b$. Examples might include the following situations: (1) data for each cell are only available for a sample, (2) design data are old and the distribution of design variables in the cell may have changed over time, or (3) data on individual elements are measured with error. Similarly, the distribution of $\mathbf{t}$ might be available only for a supercell that contains multiple values of $\mathbf{u}$ (for example, race and census block of residence are known for each individual, but the income distribution is known for the block as a whole but not for each race within the block), so $\bar{\mathbf{t}}_b$ and $\mathbf{S}_{Tb}$ must be estimated under a model.

2.4 More than one model

The preceding development assumes that all estimands of interest are combinations of parameters of a single model. More generally, the contemplated analyses might involve fitting several models, and V might sum the variances of combinations of parameters from these models. An obvious special case is estimation of a population mean (as suggested in Section 2.2), the coefficient of the model $y_k = \beta_0 \cdot 1 + \varepsilon_{bk}$, together with some regression coefficients. Another simple example is estimation of the means of variously defined domains, that is the coefficients of models of the form $y_k = \boldsymbol{\beta}'_{(m)}\mathbf{x}_{(m)bk} + \varepsilon_{(m)bk}$ where $\mathbf{x}_{(m)bk}$ is a vector of domain membership indicators with alternative domain definitions indexed by $m = 1, \ldots, M$, or contrasts of these means. For example, we might be interested in estimating mean outcomes both by race and by age.

If each of the combinations of interest only includes parameters of a single model, then each combination has its own design matrix, so the model index m can be identified with the estimand index i. Thus in (5) and (6) we replace $\mathbf{X}_U$ with $\mathbf{X}_{U(i)}$ and replace $\mathbf{S}_b$ with $\mathbf{S}_{(i)b}$.

If some estimands combine parameters from different models, we stack the estimators $\hat{\boldsymbol{\beta}}_{(m)}$ for the different models. Then in (5) and (6) we replace $\mathbf{X}_U'\mathbf{X}_U$ with $\mathrm{diag}(\mathbf{X}_{U(m)}'\mathbf{X}_{U(m)}, m = 1, \ldots, M)$ and redefine $\mathbf{S}_b$ as the combined sums of squares and crossproducts matrix for all of the models, with blocks

$$\mathbf{S}_{b(m,m')} = \begin{pmatrix} \mathbf{u}_{b(m)}\mathbf{u}_{b(m')}' & \mathbf{u}_{b(m)}\bar{\mathbf{t}}_{b(m')}' \\ \bar{\mathbf{t}}_{b(m')}\mathbf{u}_{b(m)}' & \mathbf{S}_{Tb(m,m')} \end{pmatrix}.$$

The remainder of the optimization is unchanged from Section 2.3.

2.5 Flexible contrast weights

In CanCORS, we are interested in the income effect controlling for race and averaged across races. It is less important to us how the races are weighted in that average, since the study areas are not representative of national proportions by race. Then we might estimate the poverty/nonpoverty income effect for each race and combine them with weights chosen to minimize the variance of the estimator of the weighted average of within-race income effects.

In general, we consider situations in which scientific interest is directed at estimating or testing *any* combination $\mathbf{a}_i = \mathbf{A}_i\mathbf{f}_i$ where $\mathbf{A}_i$ is fixed and each $\mathbf{f}_i$ is arbitrary (and not necessarily all of the same dimension) subject to the constraints $\mathbf{1}'\mathbf{f}_i = 1$, $f_{ij} \geq 0$. In our motivating example, the underlying model includes eight indicator variables for each of the groups defined by four race groups crossed with dichotomous poverty level, and $\mathbf{A}_1$ is an 8×4 matrix in which each column contains a 1 and -1 for the contrast between poor and nonpoor within one race. Then $\mathbf{f}_1$ contains the weights given to the contrast in each race, and $\mathbf{a}_1' = (f_{11}, -f_{11}, f_{12}, -f_{12}, f_{13}, -f_{13}, f_{14}, -f_{14})$ is the weighted contrast of the eight indicator coefficients.

Substituting into (5)-(6), we optimize over both sampling probabilities $\boldsymbol{\pi} = \{\pi_k\}$ and combining weights $\mathbf{f} = \{\mathbf{f}_i\}$. With multiple models, we use either of the formulations of Section 2.4, depending on whether the combinations of

interest include coefficients of one or several models. The definition of $\mathbf{a}_i$ is thus determined in part by scientific considerations and in part by the information available from the population at hand.

A natural approach to jointly optimizing $\boldsymbol{\pi}$ and $\mathbf{f}$ is alternately to minimize V with respect to $\boldsymbol{\pi}$ using the modified (6) and with respect to $\mathbf{f}$, observing the constraints on $\mathbf{f}$. In the optimization, $\mathbf{f}_i$ appears in an expression of the form $\mathbf{f}_i'\mathbf{D}_i(\pi)\mathbf{f}_i$. Minimizing subject to the constraint $\mathbf{f}_i' \cdot \mathbf{1} = 1$ using Lagrange multipliers, we obtain $\hat{\mathbf{f}}_i = \mathbf{D}_i^{-1}(\pi)\mathbf{1}/(\mathbf{1}'\mathbf{D}_i^{-1}(\pi)\mathbf{1})$ as long as $\pi_{bk} > 0$ and the nonnegativity constraints are not binding. If the nonnegativity constraints are binding, quadratic programming methods can be used.

2.6 Precision of unanticipated analyses

A design that is intended to be optimal for one regression coefficient might be very inefficient for other regression coefficients in the same or different models. Making the population mean one of the estimands helps to control this risk. We illustrate this by an example with design variables x_k, z_k with joint distribution

$$\zeta\colon (X, Z) \sim N\left(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right)$$

fully observed for individuals (indexed by k as in Section 2.2) and the following constant and univariate regression models:

Model 0: $y_k = \alpha_0 + \varepsilon_k^0$, $\varepsilon_k^0 \sim [0, \sigma_0^2]$

Model 1: $y_k = \alpha_X + \beta_X x_k + \varepsilon_k^X$, $\varepsilon_k^X \sim [0, \sigma_X^2]$

Model 2: $y_k = \alpha_Z + \beta_Z z_k + \varepsilon_k^Z$, $\varepsilon_k^Z \sim [0, \sigma_Z^2]$. To simplify notation we assume $\sigma_0^2 \approx \sigma_X^2 \approx \sigma_Z^2 \approx 1$, $\bar{x}_U = \bar{z}_U = 0$ and costs c_k are constant.

Consider the sample design optimized for $V = dV(\hat{\alpha}_0) + V(\hat{\beta}_X), d \geq 0$. By (4), the optimal design has $\pi_k \propto \sqrt{d + x_k^2}$. Under this design, the anticipated variance is approximated by $V(\hat{\beta}_Z) \approx n^{-1}\sigma_Z^2(Z_U'Z_U)^{-1}(Z_U'W_UZ_U)(Z_U'Z_U)^{-1}$ where $Z_U = (z_1, \ldots, z_N)'$ and $W_U = \text{diag}(\pi_1^{-1}, \ldots, \pi_N^{-1})$. Then $E_\zeta nV(\hat{\beta}_Z) \approx c_0 E_\zeta(Z^2/\sqrt{d + X^2})$ where c_0 depends only on d so $E_\zeta nV(\hat{\beta}_Z)$ depends on ρ and d. If $d = 0$ (no weight is attached to the estimation of mean), $E_\zeta nV(\hat{\beta}_Z) = \infty$ unless $\rho = \pm 1$. Thus unless the objective gives some weight to the variance of the mean estimator, the design is potentially very poor for the coefficients attached to covariates that are not in the span of variables of the optimized models. But if $d > 0$ we can decompose Z into components parallel and orthogonal to X, $Z = Z_1 + Z_2$ where $Z_1 = \rho X$ and $Z_2 = Z - \rho X$, so $Z_1 \perp Z_2$, $Z_2 \perp X$ and $E_\zeta Z_2 = 0$. Then $E_\zeta nV(\hat{\beta}_Z) = \rho^2 c_0 E_\zeta(X^2/\sqrt{d + X^2}) + (1 - \rho^2) c_0 E_\zeta(1/\sqrt{d + X^2}) = \rho^2 E_\zeta V_{\text{opt}}(\hat{\beta}_X) + (1 - \rho^2) E_\zeta V_{\text{opt}}(\hat{\alpha}_0)$.

In words, the variance of the coefficient of the new model is a combination of the two variances that were controlled in the optimization. This suggests that a design that includes estimation of the overall population mean in the optimization gives some protection against extreme inefficiency for other models with variables that were not considered in the original design, although the simple results given here do not necessarily generalize to cover every case.

3. Application: Regressions on race and poverty status

3.1 Description of sites and data

The CanCORS project (Ayanian *et al.* 2004) consists of five geographically-defined sites (northern California, Los Angeles, Alabama/Georgia, North Carolina, and Iowa) and two organizationally-based sites. The northern California site consists of 9 counties extending from the San Francisco Bay area to semirural Placer County on the Nevada border. This site is ethnically diverse and geographically varied and therefore best illustrates the methods. We describe results for this site in detail and then summarize results for other sites.

Our data were based on the 2000 U. S. Decennial Census "long form" sample and were extracted for the 9 counties of our target area (Alameda, Contra Costa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara, and Solano) from SF-3, Tables 159a-159i, "Poverty Status in 1999 by Age." We cross-tabulated the sampled residents at least 65 years old of each census block group (a small contiguous area roughly equivalent to several city blocks, henceforth referred to as a block) by race/ethnicity and income, using census sampling weights. The age restriction roughly corresponds to the ages of most incident cancer cases eligible for the study. Household income was dichotomized as exceeding or falling below the standard poverty line. The census included separate items on Hispanic ethnicity and race; we classified the population as Hispanic or as non-Hispanic white, Black, or Asian-American. A heterogeneous "Rest" category constitutes the remaining 3% of the elderly population. (For conciseness we henceforth refer to these as "race" categories.) The study site contained 844,560 over-65 individuals in 5,098 block groups, or an average of 166 per block group.

Table 1 summarizes the distribution of race and income in the northern California site. Blacks have the highest overall poverty rate and are also the most segregated (largest coefficient of variation of percent Black by block), consistent with national patterns of residential segregation (Denton and Massey 1993). Hispanics have the most relative geographical variation in poverty rates (largest coefficient of variation of poverty rates by block).

Table 1
Distributions of race and poverty for those with age ≥ 65 years, by census block group in the northern California site. (CV = coefficient of variation)

	White	Black	Asian	Hispanic	Rest	Total
Percent of population	65.70	6.40	16.80	8.20	3.00	100.00
Percent poor	5.20	14.20	10.10	10.60	11.30	7.20
CV block percent in race	0.46	2.94	1.21	1.73	2.28	-
CV block percent poor	1.53	1.37	1.58	1.89	2.30	1.16

3.2 Design conditions: Available information and design objectives

We calculated the efficiency relative to simple random sampling (SRS) of the optimal design for scenarios defined by two conditions: (1) the choice of objective function, and (2) the assumptions about the information available for determining sampling probabilities.

We considered six possible assumptions about available information for race (unavailable, or available at the individual level) and income (unavailable, only available by block, or available at the individual level). Because race is more often recorded in hospital records than income, we excluded the case where individual income group is known but race is only known by block group. Each assumption corresponds to a definition of the cell for the development of Sections 2.3 and a corresponding definition of variables $\mathbf{t}$ and $\mathbf{u}$:

1. No design information available: the cell is the entire population and $\mathbf{u}$ includes race and income. (Columns headed "SRS" in Table 2.)
2. Race alone: the cell is a race category, $\mathbf{u}$ contains race variables, and $\mathbf{t}$ is income. (Columns headed "Race.")
3. Block-aggregated data alone: the cell is a census block group, $\mathbf{u}$ is empty and $\mathbf{t}$ includes race and income. (Columns headed "Block.")
4. Individual race, block-aggregated income data by race: the cell is the population of one race in a block group, $\mathbf{u}$ is race, and $\mathbf{t}$ is income. (Columns headed "Race+Block.")
5. Individual income, no race data: the cell is an income group, $\mathbf{u}$ is income and $\mathbf{t}$ is empty. (Columns headed "Income.")
6. Race and income both available for each individual: the cell is a race by income category, $\mathbf{u}$ includes race and income, and $\mathbf{t}$ is empty. (Columns headed "Race+Income.")

We calculated optimal sampling rates under each assumption about available information, with a variety of objective functions. Each of the objective functions we considered weights together variances of coefficient estimates in some or all of four regression models: (1) the "intercept only" model whose single parameter is the population mean, (2) a race model parametrized as a white mean and contrasts for differences between whites and each of the other major race groups (Blacks, Hispanics, and Asians), (3) an income model parametrized as a nonpoor mean and a contrast between poor and nonpoor, and (4) an additive joint model including race and income effects. Every objective includes weight $d_{\text{mean}} > 0$, which guarantees that all $\pi_{bk} > 0$, avoiding numerical problems in the optimization. Thus, at least two models are represented in each objective (Section 2.4). When the objective weights both income and race effects, the single income effect is given weight $d_{\text{income}} = 3$ to match the three race effects with weights of 1.

We explored a selection of objective weights that emphasized estimation of race effects, income effects, or both. Each panel of Table 2 represents a single choice of objective weights d_i (third column) for the contrast coefficients $\mathbf{a}_i$ (second column) of a series of models (first columns). The fourth column shows the variance (normalized to unit sample size) $nV_{\mathbf{a}_i}$ for estimation of that coefficient under SRS assuming residual variance $\sigma^2 = 1$. The remaining columns present design effects, the ratios of the normalized variance $nV_{\mathbf{a}_i}$ for the optimized design with various assumptions about available design information to the variance under SRS. Rows with objective weight $d_i = 0$ do not affect the optimization but are included to illustrate the effect of each design on efficiency for estimating a coefficient that is not included in the objective function. The final row summarizes the weighted design effect correspondding to the loss function, that is, the weighted combination of variances.

3.3 Efficiency with fixed models

The first two objective functions optimize for estimation of race contrasts and the overall mean. Using individual race greatly improves efficiency for estimating Black and Hispanic effects. The greatest gains are for the Black effect (the smallest of the three major racial minorities), whose variance is reduced to 43% of its value under SRS. Conversely there is no gain for Asian-Americans, whose population representation is close to the optimal sampling rate. With this objective, once race is available, additional design information (block or individual income) is irrelevant to the optimization. If individual race is unknown, using block of residence can help with oversampling of Blacks (the most segregated group residentially), reducing the variance of the estimated Black effect to about 65% of that under SRS, but oversampling by block only slightly reduces the variance of the estimated Hispanic effect. Knowing income by itself is of little use to improve sampling for estimation of race effects.

Table 2
Normalized variances and objective functions for optimal designs for various objective weights and design information assumptions

Objective 1: Optimized for race effects

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	0.1	1.0	181	119	181	100	181
Race	Black	1	17.2	43	65	43	99	43
	Asian	1	7.5	100	106	100	100	100
	Hispanic	1	13.7	55	90	55	100	55
Income	Poor	0	15.0	182	104	182	81	182
Race+Income	Black	0	17.4	44	65	44	99	44
	Asian	0	7.5	100	106	100	100	100
	Hispanic	0	13.8	55	90	55	100	55
	Poor	0	15.2	182	104	182	81	182
Total = $nV = n\sum d_i V_{a_i}$			38.6	59	82	59	99	59

Objective 2: Optimized for race effects and overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	136	115	136	100	136
Race	Black	1	17.2	44	66	44	99	44
	Asian	1	7.5	100	104	100	100	100
	Hispanic	1	13.7	56	90	56	100	56
Income	Poor	0	15.0	121	101	121	82	121
Race+Income	Black	0	17.4	45	66	45	99	45
	Asian	0	7.5	100	104	100	100	100
	Hispanic	0	13.8	56	90	56	100	56
	Poor	0	15.2	122	102	122	82	122
Total = $nV = n\sum d_i V_{a_i}$			41.5	65	84	65	100	65

Objective 3: Optimized for income effect

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	0.001	1.0	103	154	173	173	173
Race	Black	0	17.2	75	119	152	163	163
	Asian	0	7.5	90	144	173	170	170
	Hispanic	0	13.7	86	142	196	168	168
Income	Poor	3	15.0	97	74	60	27	27
Race+Income	Black	0	17.4	75	119	153	164	164
	Asian	0	7.5	90	144	174	171	171
	Hispanic	0	13.8	86	143	197	169	169
	Poor	0	15.2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			45.0	97	74	60	27	27

Objective 4: Optimized for income effect and overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	103	134	147	151	151
Race	Black	0	17.2	76	107	128	142	142
	Asian	0	7.5	91	127	145	148	148
	Hispanic	0	13.7	86	125	161	147	147
Income	Poor	3	15.0	97	75	61	27	27
Race+Income	Black	0	17.4	76	107	129	143	143
	Asian	0	7.5	91	127	146	149	149
	Hispanic	0	13.8	86	125	162	147	147
	Poor	0	15.2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			48.0	97	79	66	35	35

Table 2 (continued)
Normalized variances and objective functions for optimal designs for various objective weights and design information assumptions

Objective 5: Optimized for separate race effects, income effect and overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	111	117	135	114	150
Race	Black	1	17.2	54	74	55	109	53
	Asian	1	7.5	95	106	109	112	116
	Hispanic	1	13.7	67	96	69	112	67
Income	Poor	3	15.0	101	82	72	38	37
Race+Income	Black	0	17.4	55	74	55	109	52
	Asian	0	7.5	95	106	109	113	115
	Hispanic	0	13.8	67	96	69	112	66
	Poor	0	15.2	101	82	72	39	35
Total = $nV = n\sum d_i V_{a_i}$		0	86.4	86	86	74	73	56

Objective 6: Optimized for race effects and income effect in two-factor model and for overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	111	119	138	114	156
Race	Black	0	17.2	55	74	55	108	53
	Asian	0	7.5	95	107	109	112	114
	Hispanic	0	13.7	67	96	69	111	67
Income	Poor	0	15.0	101	82	72	38	37
Race+Income	Black	1	17.4	55	74	55	109	52
	Asian	1	7.5	95	107	109	113	113
	Hispanic	1	13.8	67	96	69	112	66
	Poor	3	15.2	101	81	71	39	35
Total = $nV = n\sum d_i V_{a_i}$			87.2	86	86	73	73	54

Objective 7: Optimized for income effect in two-factor model and for overall mean

Model	Effect	Weight (d_i)	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	103	135	149	148	156
Race	Black	0	17.2	77	100	98	139	97
	Asian	0	7.5	91	124	132	145	132
	Hispanic	0	13.7	86	122	135	144	122
Income	Poor	0	15.0	97	75	62	28	28
Race+Income	Black	0	17.4	77	100	98	140	96
	Asian	0	7.5	91	124	132	146	132
	Hispanic	0	13.8	86	122	135	144	121
	Poor	3	15.2	97	75	62	29	27
Total = $nV = n\sum d_i V_{a_i}$			48.6	97	79	67	36	35

Disproportionate sampling, tuned to optimize for estimation of race effects, inflates the variances of the other parameter estimators. When minimal weight is given to the mean in the optimization objective (Objective 1), this inflation can be quite large: a factor of 181% for the mean and income effects. Giving more weight to the mean (Objective 2) moderates this effect, reducing the variance inflation to 136% for the mean and 121% for the income effect, while only slightly increasing variances for the race effects.

The minimum possible normalized variance for estimation of the income effect (Objective 3) is 4 (27% of the variance under SRS), attained when income is known for individuals under a design that divides the sample equally between poor and nonpoor. With block-level information, variance can be reduced to 74% of that under SRS. Although knowing race alone has little benefit for this objective, adding individual race to block-level information further reduces the variance of the estimated income effect to 60% of that under SRS. Variances of estimates of the mean and of race effects are substantially increased under

these designs, but increasing the weight of the mean (Objective 4) substantially ameliorates the variance inflation for the mean and race effects, only slightly increasing the variance of the estimated income effect.

Including both race and income effects in Objective 5 yields designs that are not quite as good as the optimal designs for either alone, but still much better than SRS. For example, variances of the race effects with race and block of residence known are 10% to 24% higher than with the designs using the same design information but separately optimized for race or income. When only individual race or only individual income is known, the design essentially optimizes for the effects corresponding to the available variable, inflating the variance of estimated effects of the other variable.

The design optimized for joint race and income effects in the two-factor additive model (Objective 6) is quite close to that optimizing for race and income effects in separate marginal models (Objective 5). When optimizing for separate effects, variances of these effects are slightly smaller than those of the corresponding effects in the two-factor model. When optimizing for effects in the joint model, their variances are reduced although in most cases still slightly larger than those of the corresponding effects in marginal race and income models, due to the partial confounding of race and income effects.

Likewise, optimization for the income effect in the two-factor model (Objective 7) is fairly similar to optimization for the univariate income effect (Objective 4) when no race data are available. Making race data available together with either block or individual income, however, considerably reduces variances for race effects under the design for the two-factor model. Because of the partial confounding of race and income effects under this model, this design adapts to estimate the former more efficiently, accumulating more data at the design points that are critical to unconfounding these effects.

3.4 Efficiency with flexible contrast weights

We next consider the potential benefits of estimating income effects under a flexible weighting scheme (Table 3). The objective function considers coefficients of two models, the constant model whose parameter is the population mean, and a model with indicator variables for each race-by-income cell. The income effect within each race is estimated as the difference of the coefficients for poor and nonpoor within that race, and these estimates are combined with flexible weights to estimate an overall income effect. This strategy is most nearly parallel to Objective 7, which also estimates income effects controlling for race. The flexible-contrast analysis is less model-dependent than the two-factor model in that it does not rely on that model's additivity assumption. On the other hand, the way the races are combined does not necessarily reflect population proportions. The weights given to the income contrast in each race, estimated as described in Section 2.5, are presented in the lower panel of Table 3 to demonstrate how this approach allows us to modify the estimand to exploit available design information. (The alternating-optimization algorithm converged to adequate accuracy within 7 iterations.)

Under SRS the variance of the income effect under the flexible-weights model is slightly larger than in the two-factor model (15.91 versus 14.99). The weight given to the white contrast under this design (51%) is less than the white proportion of the population (66%) because relatively few whites are poor and therefore the income contrast among whites is relatively imprecise. Conversely, the weight for the Black income contrast (12%) is almost twice that group's share of the population, because of the disproportionately high poverty rates in that group.

Using individual race in the design accentuates this disproportion: more sample, and much more weight (75%), is given to the Black group, with the highest percentage in poverty. Thus flexible weighting makes possible a large reduction in the variance of the estimated income effect (to 63% of that under SRS) using only race, which was not possible under the more restrictive two-factor DPQ model.

Block-level information is slightly less useful for this design than race information. The combination of block and race information, however, is very powerful, reducing the variance of the income effect to 48% of that under SRS. Under this design, much more weight (46%) is given to the Hispanic income contrast, which can be estimated efficiently because of the greater income segregation among Hispanics (Table 1). When individual income is available (with or without race), the contrasts weights approximate the proportions by race, since efficient income contrasts can be obtained within any race and the inclusion of the overall mean in the objective pulls the design toward proportionate sampling. Thus, the design is dramatically different under alternative assumptions about availability of design information.

3.5 Comparisons across sites

Table 4 compares the gains for disproportionate sampling at four CanCORS sites, excluding the nongeographical sites and one site (Iowa) that was almost all white. At each site we optimized for unit ($d_i = 1$) weighting of variances of overall mean and the income effect in the two-way model (proportional to Objective 7), under alternative assumptions about available design information. The theoretical minimum for this objective with a balanced population is 5 ($V_{\text{mean}} = 1$, $V_{\text{income}} = 4$). SRS is inefficient at every site,

especially in Alabama and northern California, and race information alone would be of little help. Conversely, the best variance attainable using full race and income information on individuals is between 5.60 and 5.72 at each site. Oversampling based on block-level income information would substantially reduce variances, with substantially greater gains in Alabama and northern California than in the other sites.

4. Discussion

To develop design alternatives for a health services study, we extended previous methods for optimal design in domain estimation to show how an optimal unequal-probability sampling scheme can be designed for estimation of regression coefficients in one or more models. In our application, substantial reductions in variance were possible even if some variables were only available for geographical aggregates. Particularly large gains were possible for categorical regressors (poverty status, race) with very imbalanced distributions.

In essence, our approach to survey design with imprecisely measured design variables uses the predictive distribution of the design variables for each sampled unit, specifically the expectations of the variables and of their squares and cross-products. This concept unites design using cell aggregates (estimated from census or sample data), using variables measured with error, or using a sampling frame whose units might have changed their characteristics over time.

Table 3
Normalized variances and contrast weights for optimal DPQ designs with flexible weighting of income contrasts by race. Lines for fixed contrasts are included to demonstrate the effect of various choices of flexible weights, for comparison to fixed-weight objective scenarios

Variances for "flexible-weight" estimate and for contrasts represented in Table 2								
				Variance as percent of variance under SRS (by available design information)				
Model	Effect	Weight (d_i)	Variance Under SRS	Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	233	139	206	152	152
Flexible contrast	Income	3	15.9	63	74	48	28	26
Race	Black	0	17.2	34	90	61	143	139
	Asian	0	7.5	197	124	172	149	146
	Hispanic	0	13.7	172	124	61	148	144
Income	Poor	0	15.0	209	77	124	27	39
Race+Income	Black	0	17.4	35	90	61	144	139
	Asian	0	7.5	197	124	173	150	147
	Hispanic	0	13.8	172	124	61	148	144
	Poor	0	15.2	210	77	125	29	39
Total = $\sum d_i V_{a_i}$			50.7	73	78	57	35	33

Optimum weights (as percent, 100% × f_i) of each within-race income contrast in calculation of the combined estimate of the income effect, under each design information assumption. (Columns may not sum to 100% due to roundoff error.)

	Design information assumptions					
Contrast	SRS	Race	Block	Race+Block	Income	Race+Income
Black	12	75	17	25	9	6
Asian	24	9	26	16	21	17
Hispanic	12	5	13	46	11	8
White	51	12	45	13	59	68

Table 4
Normalized objective function for optimal DPQ designs under equal (d_i = 1) weighting of variances of the overall mean and the income effect in the two-way model, at four CanCORS sites

	Variance under SRS	Variance as percent of variance under SRS				
Site location		Race	Block	Race+Block	Income	Race+Income
Alabama	16.2	97	79	67	36	35
Los Angeles	11.8	98	85	76	49	47
North Carolina	10.2	97	89	86	59	55
Northern California	16.2	97	79	67	36	35

The methods described here for optimizing element sampling probabilities can be combined with stratification and cluster or multistage sampling. (Neither of these design features appeared in the CanCORS study which motivated our research. Stratification was inconvenient given the sequential identification of subjects and there was little prior information to guide construction of homogeneous strata. Telephone interviewing made it operationally unnecessary to cluster our subjects.) Because these design features can affect the sampling distributions of both the design and outcome variables, and the design objectives involve both the posited population model and the scientific model of interest, the number of possible combinations is even larger than in design for estimation of a population mean. We therefore limit ourselves to suggesting a few ideas to be followed up in future research.

Stratification can improve a design for a regression analysis in at least three ways: (1) to implement disproportionate sampling (using probabilities equal or close to those derived under our methodology), (2) to control the distribution of design variables to be closer to the optimal design than in an unstratified unequal-probability design, and (3) to reduce the within-stratum variation of the case influence statistics and thereby reduce the variance of coefficient estimates (Fuller 1975). Since the efficiency of the design is insensitive to small deviations around the optimum, some stratified designs with equal probabilities within strata might approach the efficiency of the optimal design. *Ad hoc* stratifications might have poorer efficiency, even with optimal allocation to strata. For example, stratifying blocks by the least prevalent race-income group represented yielded a design with about half the efficiency gain of our design using aggregated block composition.

With regard to the last point, note that designing homogeneous strata for estimation of regression coefficients is likely to be more difficult than for estimation of a mean. The influence of an observation depends on its residual from the regression model, not its raw value, so to reduce homogeneity the stratification would have to involve predictive variables not included in the model. Influence also depends on the observation's leverage for each coefficient, a possibly complex function of the covariates.

For cluster sampling, the equivalence of $E_\pi \mathrm{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}}$ and $E_\xi \mathrm{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}}$ might not hold except under restrictive assumptions such as independent residuals; thus the terms of the middle factor of (5) would take a more complex form. There are several possible cases for cluster sampling depending on the relationship between the cells and the clusters, which should be elaborated on further research.

Another natural extension is to nonlinear regression models and other estimands defined by estimating equations. The weighted least squares formulation of the Newton-Raphson step (McCullagh and Nelder 1989, sec. 2.5) for a generalized linear model can be applied by suitably defining σ^2_{bk} in (3) and hence in (4)-(6); a similar procedure can be applied for other estimating equations (Binder 1981; Binder 1983; Morel 1989). Because the variances are functions of the model predictions, implementing this modification requires design assumptions about the fitted model as well as about the distribution of the covariates.

Every optimization has its costs, which for our methods can be both practical and statistical.

In the CanCORS study, incident cases of the cancers under study were identified in real time through a field operation ("rapid case ascertainment"); patients then had to be contacted on a very tight schedule to start contacting them for interviews within the desired interval (3 months from their dates of diagnosis). Thus, the practical issues of survey implementation were exacerbated. Among the concerns that ultimately led us not to implement the DPQ design were (1) the difficulty of accurately geocoding patients within the time frame allowed; (2) incomplete and inaccurate race identification in the case ascertainment data, and (3) lower-than-expected participation rates, which made any sampling problematical.

Such issues are less problematic in surveys with a static sampling frame that can be processed on a less stringent timeline, particularly in large-scale and/or repeated surveys in which even modest variance reductions justify some added complexity. They could be used, for example, to evaluate the potential gains through geographically-based oversampling in surveys for which national estimates by race are required.

Statistical concerns about our design strategy arise because optimization for one set of predetermined statistical objectives is likely to reduce efficiency for others. It is difficult in any but the most tightly focused study to anticipate all potential analyses. Simultaneous optimization for a reasonably comprehensive collection of analyses, and investigation of sensitivity of the design to varying the relative weights of the various objectives, should give some protection against an overspecialized design. However, this approach can only be used with variables for which there are some data prior to the study. The results in Section 2.6 suggest that monitoring the effect of disproportionate sampling on the precision of the population mean gives some protection against designs that are excessively inefficient for unanticipated analyses and variables, although the bounds there are not very general.

More broadly, we might ask when the DPQ analysis is the scientifically relevant estimand. Regression models are often used in analyses intended to be generalizable to broader populations, rather than to describe the finite

population at hand, just as the CanCORS sites were selected purposively to study patterns and variations in care that might reflect broader national patterns. While using sampling weights in enumerative studies is relatively uncontroversial, there has been a lively debate about the use of weights in analytic studies (Hansen, Madow and Tepping 1983 and discussion; DuMouchel and Duncan 1983; Bellhouse 1984; Pfeffermann 1993, Fuller 2002, sec. 5). A population-descriptive analysis offers some robustness against the possibility that the sample will be selected in way that distorts typical relationships. Thus, even where a pure DPQ analysis cannot be justified on grounds of enumerative representativeness, a sample drawn to optimize unweighted estimation of regression coefficients might have limited scientific value. For example, suppose that the CanCORS data would be analyzed with an *unweighted* regression to estimate a simple income effect (a contrast of means), using block level design information from the census. Optimally the sample would draw from a collection of blocks which, taken together, have about half their residents in poverty. Since poverty rates are rarely that high, this effectively requires sampling only from the blocks with the highest poverty rates. Such a sample would be unrepresentative of either of the income groups. Similarly, a sample that overrepresented Black residents by sampling from mostly Black blocks would (if analyzed without weights) be unrepresentative of the Black population in general, because the services available in highly segregated areas are likely to differ from those in more mixed areas.

More general formulations are needed, with clearly stated assumptions and objectives, that "consider[s] the model parameters as the ultimate target parameters but at the same time focuses on the DPQ's as a way to secure the robustness of the inference" (Pfeffermann 1993), taking into account the scientific objectives of the study. Previous proposals include testing the null hypothesis that the weights have no effect on the regression (DuMouchel and Duncan 1983; Fuller 1984), or including design variables (Nathan and Holt 1980; Little 1991) or the weights themselves (Rubin 1985) as control variables in the regression. These approaches are problematical, however, when the weights are functions of the covariates of primary scientific interest. We have attempted through flexible contrast weighting (Section 2.5) to take a step toward such a general formulation, extending the DPQ approach to allow a focus on a range of valid inferences for particular scientific objectives rather than exclusively on inference for finite populations. From this range, the investigator can select an inferential objective and sample design adapted to the structure of the population and the practicalities of study design.

Appendix

Equivalence of sampling and model variances

We show that $E_\xi \mathrm{Var}_\pi \hat{\boldsymbol{\beta}} \approx E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}}$ under the following conditions:

1. $1/N\ \mathbf{X}_U'\mathbf{X}_U \to \Sigma$ for some positive definite Σ. This minimal condition relates the hypothetical sequence of populations.
2. The design-based regression estimator can be written as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_U + R_{n,S}$ where $\mathrm{Var}_\xi E_\pi R_{n,S} = o(n^{-1})$ and $\mathrm{Var}_\pi E_\xi R_{n,S} = o(n^{-1})$. Note that $\hat{\boldsymbol{\beta}}$ cannot strictly be defined as in (2), because the matrix inverse is undefined when the sample values of x do not span the design space and hence its expectation and variance are also undefined. A scalar ratio estimator likewise might be undefined with nonzero but $o(n^{-1})$ probability because the sample might have only 0 values for the denominator variable. Assigning some arbitrary value in that event, the estimator nonetheless could have good asymptotic properties. A similar argument lets us assume that a suitable $\hat{\boldsymbol{\beta}}$ can be defined. We do not specify how (2) must be modified to technically satisfy the condition since this depends on the specifics of ξ and the sequence of designs.
3. $\max(\pi_i) = O(n/N)$ and $n = o(N)$, essentially our assumption that finite population corrections can be ignored.
4. Homoscedasticity, $\mathrm{Var}_\xi y_k = \sigma^2$; this is not restrictive since it can always be made true by a suitable transformation of x and y.

Under these conditions,

$$\begin{aligned}\mathrm{Var}_{\pi\xi}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}_\pi E_\xi \hat{\boldsymbol{\beta}} + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}} \\ &= \mathrm{Var}_\pi \boldsymbol{\beta} + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}} \\ &= o(n^{-1}) + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}}\end{aligned}$$

On the other hand

$$\mathrm{Var}_{\xi\pi} \hat{\boldsymbol{\beta}} = \mathrm{Var}_\xi E_\pi \hat{\boldsymbol{\beta}} + E_\xi \mathrm{Var}_\pi \hat{\boldsymbol{\beta}}$$

The first term in the above equation is

$$\begin{aligned}\mathrm{Var}_\xi E_\pi(\boldsymbol{\beta}_U + R_{n,S}) &= \mathrm{Var}_\xi(\boldsymbol{\beta}_U + E_\pi R_{n,S}) \\ &= (\mathbf{X}_U'\mathbf{X}_U)^{-1}\sigma^2 + o(n^{-1}) + o(N^{-1/2}n^{-1/2}) \\ &= O(N^{-1}) + o(n^{-1}) + o(n^{-1/2}N^{-1/2}) = o(n^{-1})\end{aligned}$$

This proof is an elaboration of one by Isaki and Fuller (1982), summarized in Pfeffermann (1993, page 321).

Acknowledgements

This research was funded by grants U01-CA93344 (Zaslavsky and Zheng), U01-CA93324 (Zaslavsky), and U01-CA093348 (Adams) from the National Cancer Institute. The authors thank Nat Schenker and Van Parsons for useful comments on an earlier draft, and the associate editor and two referees for thoughtful comments.

References

Ayanian, J.Z., Chrischilles, E.A., Wallace, R.B., Fletcher, R.H., Fouad, M., Kiefe, C.I., Harrington, D.P., Weeks, J.C., Kahn, K.L., Malin, J.L., Lipscomb, J., Potosky, A.L., Provenzale, D.T., Sandler, R.S., Vanryn, M. and West, D.W. (2004). Understanding cancer treatment and outcomes: The Cancer Care Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*, 2, 2992-2996.

Bellhouse, D.R. (1984). A review of optimal designs in survey sampling. *The Canadian Journal of Statistics*, 12, 53-65.

Binder, D.A. (1981). On the variances of asymptotically normal estimators from complex surveys. *Survey Methodology*, 7, 157-170.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Botman, S.L., Moore, T.F., Moriarity, C.N. and Parsons, V.L. (2000). *Design and Estimation for the National Health Interview Survey, 1995-2004*. Vital and Health Statistics, 2(130). Washington, DC: National Center for Health Statistics.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chromy, J.R. (1987). Design optimization with multiple objectives. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA. 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Denton, N.A., and Massey, D.S. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard University Press.

Dumouchel, W.H., and Duncan, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

Eaton, D.K., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Harris, W.A., Lowry, R., Mcmanus, T., Chyen, D., Shanklin, S., Lim, C., Grunbaum, J.A. and Wechsler, H. (2006). Youth risk behavior surveillance - United States, 2005. *Morbidity and Mortality Weekly Report*, 55(SS-5), 1-108.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C37, 117-132.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-97.

Kish, L. (1974). Optimal and proximal multipurpose allocation. Alexandria, VA. In *Proceedings of the Social Statistics Section*, American Statistical Association, 111-118.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (Second Edition). London: Chapman & Hall Ltd.

Morel, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, 15, 203-223.

Morris, C., Newhouse, J.P. and Archibald, R. (1979). On the theory and practice of obtaining unbiased and efficient samples in social surveys and experiments. *Research in Experimental Economics*, 1, 199-220.

Nathan, G., and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, Series B, *Methodological*, 42, 377-386.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, (Eds., J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith). Amsterdam: Elsevier/North-Holland, 463-472.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Sastry, N., Ghosh-Dastidar, B., Adams, J. and Pebley, A.R. (2005). The design of a multilevel survey of children, families, and communities: The Los Angeles Family and Neighborhood Survey. *Social Science Research*, 35, 1000-1024.

Thomsen, I., Tesfu, D. and Binder, D.A. (1986). Estimation of design effects and intraclass correlations when using outdated measures of size. *International Statistical Review*, 54, 343-349.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, New York: John Wiley & Sons, Inc.

Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. In *ASA Proceedings of the Social Statistics Section*, American Statistical Association. Alexandria, VA.

Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. In *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA, 497-502.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.