Nonparametric Estimation in Complex Surveys with Auxiliary Information

Jean D. Opsomer
Iowa State University, Department of Statistics
Ames, IA 50011, USA
jopsomer@iastate.edu

F. Jay Breidt
Colorado State University, Department of Statistics
Fort Collins, CO 80523, USA
jbreidt@stat.colostate.edu

Abstract

Auxiliary information is often used to improve the precision of survey estimators though regression techniques. We explain how regression estimation can be extended from the currently standard ratio and linear models to nonparametric models, under both simple and more complicated designs. While maintaining the good theoretical and practical properties of the linear regression estimators, nonparametric estimators are better able to take advantage of possibly complicated relationships between variables. The applicability of the estimator to complex designs and multiple types of auxiliary variables will be illustrated using practical examples.

Introduction

Regression models have a long history of being used to improve the precision of survey estimators. Typically, auxiliary information is incorporated into the survey inference through parametric linear models, leading to the familiar ratio and regression estimators, and more recently to generalized regression and nonlinear regression estimators. We refer to Fuller (2002) for an excellent review of these approaches.

We consider extensions of classical regression estimation in which the regression model is *nonparametric*, i.e. it is not specified to belong to a specific parametric family. In principle, a nonparametric approach allows the application of regression estimators to situations in which linear models are not successful in capturing the relationships between the survey variables and the auxiliary information. Model-based nonparametric estimators for surveys have been proposed by Chambers et al. (1993) and Firth and Bennett (1998), among others, while Breidt and Opsomer (2000) considered design-based estimation.

Properly constructed nonparametric regression estimators share most of the good practical and theoretical properties of linear regression estimators, while having the potential for significant improvements in efficiency. Here, we will discuss nonparametric regression estimation within the context of *model-assisted estimation* (Särndal et al. 1992), a flexible framework for incorporating auxiliary information in design-based estimation, and explain how nonparametric techniques can be applied in many situations with complex design and modelling structures. We will motivate this work with a description of an application in a multi-phase natural resource survey conducted in Utah (USA), and mention a number of other applications as well.

Defining the Nonparametric Estimator

We consider the classical finite population estimation problem, in which a population U of size N is surveyed according to sampling design $p(\cdot)$, with the goal of estimating unobserved population characteristics of interest. Let y_i represent a generic unobserved characteristic for the ith element

of the population. In this article, we focus on the estimation of $t_y = \sum_U y_i$, the total of the y_i for the population, based on observing the y_i for a sample of population elements $s \subset U$ drawn according to $p(\cdot)$, with $\pi_i = \Pr(i \in s), \pi_{ij} = \Pr(i, j \in s)$ the corresponding one-way and two-way inclusion probabilities for $i, j \in U$.

Let $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{qi})$ represent a vector of auxiliary characteristics for the ith element that are known for all $i \in U$. If the $\boldsymbol{x}_i$ and y_i are related to each other, an estimator that takes advantage of that fact is likely to be more efficient than one that does not. To introduce the estimator, consider the "superpopulation" regression model ξ with moments defined as

$$\begin{aligned} \mathrm{E}_\xi(y_i|\boldsymbol{x}_i) &= \mu(\boldsymbol{x}_i) \\ (1) \quad \mathrm{Var}_\xi(y_i|\boldsymbol{x}_i) &= v(\boldsymbol{x}_i). \end{aligned}$$

In this model, $m(\cdot)$ and $v(\cdot)$ have to satisfy some regularity conditions, but are otherwise left unspecified. In order to develop the model-assisted estimator, the finite population U is treated as a realization from this superpopulation model. If $\{(\boldsymbol{x}_i, y_i) : i \in U\}$ were observed, we could compute a "population fit" for the function $m(\cdot)$ using a nonparametric fitting technique such as local polynomial regression, as done in Breidt and Opsomer (2000). Other smoothing techniques such as *penalized splines regression* (Ruppert et al. 2003) can also be used. The mean function $m(\cdot)$ can also be further specified when the auxiliary information is multivariate, for instance as a *semiparametric model* combining both parametric and nonparametric components, or as a *generalized additive model* (Hastie and Tibshirani, 1990) that incorporates a link function. In those cases, the fitting techniques also need to be adapted to these special mean structures. These types of model extensions are currently being investigated by the author. Let μ_i represent the population fit obtained using one of these nonparametric fitting methods.

The *difference estimator* defined as

$$(2) \quad \widehat{t}_{y,\mathrm{diff}} = \sum_U \mu_i + \sum_s \frac{y_i - \mu_i}{\pi_i},$$

is design unbiased, with corresponding design variance

$$(3) \quad \mathrm{Var}_p(\widehat{t}_{y,\mathrm{diff}}) = \sum\sum_U (\pi_{ij} - \pi_i\pi_j)\frac{y_i - \mu_i}{\pi_i}\frac{y_j - \mu_j}{\pi_j},$$

which depends on the size of the residuals $(y_i - \mu_i)$. Intuitively, if the superpopulation model (1) provides a good fit to the data, then the variance (3) will be smaller than that of estimators that either do not use the auxiliary information, or that use it in a poorly fitting model. An important advantage of the nonparametric approach is that, by not restricting the unknown function $\mu(\cdot)$ in (1) to be linear or parametric, nonparametric estimators have the potential to achieve smaller residuals in a wider range of practical applications.

The estimator (2) is not feasible, since the y_i are only observed for the sample. Therefore, the μ_i are estimated by design consistent estimators $\widehat{\mu}_i$, obtained by applying design-weighted versions of existing smoothing techniques, and the resulting nonparametric regression estimator is

$$(4) \quad \widehat{t}_y = \sum_U \widehat{\mu}_i + \sum_s \frac{y_i - \widehat{\mu}_i}{\pi_i}.$$

For sufficiently large samples, this estimator behaves like the difference estimator in (2). In particular, it is design consistent and it will share the efficiency properties of (2). An estimator for the asymptotic design variance of (4) is given by

$$(5) \quad \widehat{V}(\widehat{t}_y) = \sum\sum_s \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{y_i - \widehat{\mu}_i}{\pi_i}\frac{y_j - \widehat{\mu}_j}{\pi_j}.$$

When the $\widehat{\mu}_i$ are obtained by design-weighted local polynomial regression or penalized splines, $\widehat{t_y}$ has some additional desirable properties. It can be written as a linear combination of the sample observations, $\sum_s w_i(s)y_i$, with weights $w_i(s)$ independent of the y_i. These regression weights can be used for any variables collected in the same survey, and to the extent that such variables also follow model (1), they will also benefit from the efficiency gain. The resulting regression weights $w_i(s)$ are *calibrated* for the auxiliary variables, so that $\sum_s w_i(s)\boldsymbol{x}_i = \sum_U \boldsymbol{x}_i$. Finally, the estimators are location and scale invariant. In short, the weights obtained with the help of these nonparametric techniques behave as expected by analysts and other end-users of survey data, who therefore do not need to be familiar with nonparametric regression methods in order to apply these weights in their own estimation tasks.

Extensions to Complex Designs and an Application

The general framework of model-assisted estimation makes it easy the extend the nonparametric regression estimator to surveys in which the sampling design includes several stages or phases, and different auxiliary information is available at the different stages or phases. As a simple example, consider two-stage sampling of individuals, with the households as clusters. In many such situations, auxiliary information might be available for the clusters (households), but not for the elements within the households. In that case, the estimator (4) can be adjusted to include a cluster-level regression model.

Another example, which we describe in more detail, is two-phase sampling, with some auxiliary variables available at the population level, additional auxiliary variables at the phase one level, and the study variables of interest at the phase two level. Following Särndal et al. (1992), we can fit two (nonparametric, parametric or semiparametric) regression models, one for each "level" of auxiliary information, and combine them to produce a model-assisted estimator.

Let s_a represent the phase one sample with inclusion probabilities $\pi_{ai}, i \in U$, and let s represent the phase two sample selected from s_a with conditional inclusion probabilities $\pi_{i|s_a}, i \in s_a$. Let $\boldsymbol{x}_{Ui}$ represent ancillary variables available for $i \in U$, $\boldsymbol{x}_{ai}$ variables available for $i \in s_a$, and y_i a variable of interest observed for $i \in s$. Note that $\boldsymbol{x}_{ai}$ can include the variables in $\boldsymbol{x}_{Ui}$. We define two superpopulation mean models,

$$\begin{aligned} \mathrm{E}_\xi(y_i|\boldsymbol{x}_{Ui}) &= \mu_U(\boldsymbol{x}_{Ui}) \\ \mathrm{E}_\xi(y_i|\boldsymbol{x}_{ai}) &= \mu_a(\boldsymbol{x}_{ai}), \end{aligned}$$

both with corresponding variance functions $v_U(\cdot), v_a(\cdot)$. Both models are fitted on the observations in s using survey-weighted nonparametric techniques, and the resulting model-assisted estimator is

$$t_y = \sum_U \widehat{\mu}_{Ui} + \sum_{s_a} \frac{\widehat{\mu}_{ai} - \widehat{\mu}_{Ui}}{\pi_{ai}} + \sum_s \frac{y_i - \widehat{\mu}_{ai}}{\pi_{ai}\,\pi_{i|s_a}}. \quad (6)$$

The design properties and the asymptotic variance of this estimator can be derived within the model-assisted inference framework.

The estimator in (6) was applied to data from the Utah Forest Health Monitoring survey conducted by the U.S. Forest Service. Remote sensing data were collected on a dense grid of 67,216 plots (phase one), a subset of 3,107 of these plots was field visited (phase two), and a set of ecological indicators of forest health was collected on a further subset of 71 plots (phase three). For this application, the phase one was treated as the population of interest. Two semiparametric additive models were fitted, one relating the variables of interest to the remote sensing data and another to both the remote sensing and the field data. The semiparametric additive structure was

chosen so that both categorical and continuous covariates could be accommodated, and the model fits were computed using the `gam()` procedures in *S-Plus*. The resulting estimator was much more efficient than an estimator that completely ignores the auxiliary information, and also outperformed a model-assisted estimator that only used parametric models.

Conclusion

Nonparametric techniques can extend the range of applications in which regression estimation, and model-assisted estimation particular, improves the efficiency of survey estimators. The authors are currently working on extending the theoretical investigation of this approach to more complex models and designs, as described here, and to nonparametric small area estimation. We are also collaborating with several U.S. government agencies, including the Forest Service and the Bureau of Labor Statistics, on the study of the practical behavior of the nonparametric estimators in a large-scale survey environment. The eventual goal of the research is to make available to survey practitioners a new set of estimation tools with good theoretical and practical properties.

References

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.

Chambers, R. L., A. H. Dorfman, and T. E. Wehrly (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association 88*, 268–277.

Firth, D. and K. E. Bennett (1998). Robust models in probability sampling. *Journal of the Royal Statistical Association, Series B 60*, 3–21.

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

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Washington, D. C.: Chapman and Hall.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. To appear: Cambridge University Press.

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

RÉSUMÉ

L'estimation à partir de données d'enquêtes est souvent améliorée par l'utilisation de la régression. Nous expliquons comment l'estimation par régression peut être étendue aux modèles nonparamétriques. Les estimateurs par régression nonparamétrique maintiennent les bonnes propriétés théoriques et pratiques des modèles lineaires traditionels, mais ils parviennent à capturer des relations plus compliquées entre les variables. L'utilisation de ces estimateurs dans des enquêtes avec des plans d'échantillonnage compliqués est illustrée par des examples pratiques.