Analysis of Survey Data. Edited by R. L. Chambers and C. J. Skinner
Copyright © 2003 John Wiley & Sons, Ltd.
ISBN: 0-471-89987-9

CHAPTER 11

Nonparametric Regression with Complex Survey Data

R. L. Chambers, A. H. Dorfman and
M. Yu. Sverchkov

11.1. INTRODUCTION

The problem considered here is one familiar to analysts carrying out exploratory data analysis (EDA) of data obtained via a complex sample survey design. How does one adjust for the effects, if any, induced by the method of sampling used in the survey when applying EDA methods to these data? In particular, are adjustments to standard EDA methods necessary when the analyst's objective is identification of 'interesting' population (rather than sample) structures?

A variety of methods for adjusting for complex sample design when carrying out parametric inference have been suggested. See, for example, Skinner, Holt and Smith (1989) (abbreviated to SHS), Pfeffermann (1993) and Breckling *et al.* (1994). However, comparatively little work has been done to date on extending these ideas to EDA, where a parametric formulation of the problem is typically inappropriate.

We focus on a popular EDA technique, nonparametric regression or scatterplot smoothing. The literature contains a limited number of applications of this type of analysis to survey data, usually based on some form of sample weighting. The design-based theory set out in Chapter 10, with associated references, provides an introduction to this work. See also Chesher (1997).

The approach taken here is somewhat different. In particular, it is model based, building on the sample distribution concept discussed in Section 2.3. Here we develop this idea further, using it to motivate a number of methods for adjusting for the effect of a complex sample design when estimating a population regression function. The chapter itself consists of seven sections. In Section 11.2 we describe the sample distribution-based approach to inference, and the different types of survey data configurations for which we develop estimation methods. In Section 11.3 we set out a number of key identities that

allow us to reexpress the population regression function of interest in terms of related sample regression quantities. In Section 11.4 we use these identities to suggest appropriate smoothers for the sample data configurations described in Section 11.2. The performances of these smoothers are compared in a small simulation study reported in Section 11.5. In Section 11.6 we digress to explore diagnostics for informative sampling. Section 11.7 provides a conclusion with a discussion of some extensions to the theory.

Before moving on, it should be noted that the development in this chapter is an extension of Smith (1988) and Skinner (1994), see also Pfeffermann, Krieger and Rinott (1998) and Pfeffermann and Sverchkov (1999). The notation we employ is largely based on Skinner (1994).

To keep the discussion focused, we assume throughout that nonsampling error, from whatever source (e.g. lack of coverage, nonresponse, interviewer bias, measurement error, processing error), is not a problem as far as the survey data are concerned. We are only interested in the impact of the uncertainty due to the sampling process on nonparametric smoothing of these data. We also assume a basic familiarity with nonparametric regression concepts, comparable to the level of discussion in Härdle (1990).

11.2. SETTING THE SCENE

Since we are interested in scatterplot smoothing we suppose that two (scalar) random variables Y and X can be defined for a target population U of size N and values of these variables are observed for a sample taken from U. We are interested in estimating the smooth function $g_U(x)$ equal to the expected value of Y given $X = x$ over the target population U. Sample selection is assumed to be probability based, with π denoting the value of the sample inclusion probability for a generic population unit. We assume that the sample selection process can be (at least partially) characterized in terms of the values of a multivariate sample design variable Z (not necessarily scalar and not necessarily continuous). For example, Z can contain measures of size, stratum indicators and cluster indicators. In the case of ignorable sampling, π is completely determined by the values in Z. In this chapter, however, we generalize this to allow π to depend on the population values of Y, X and Z. The value π is therefore itself a realization of a random variable, which we denote by Π. Define the sample inclusion indicator I, which, for every unit in U, takes the value 1 if that unit is in the sample and is zero otherwise. The distribution of I for any particular population unit is completely specified by the value of Π for that unit, and so

$$\Pr(I = 1 | Y = y, X = x, Z = z, \Pi = \pi) = \Pr(I = 1 | \Pi = \pi) = \pi.$$

11.2.1. A key assumption

In many cases it is possible to assume that the population values of the row vector (Y, X, Z) are jointly independent and identically distributed (*iid*).

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Unfortunately, the same is usually not true for the sample values of these variables. However, since the methods developed in this chapter depend, to a greater or lesser extent, on some form of exchangeability for the sample data we make the following assumption:

Random indexing: The population values of the random row vector (Y, X, Z, I, Π) are *iid*.

That is, the values of Y, X and Z for any two distinct population units are generated independently, and, furthermore, the subsequent values of I and Π for a particular population unit only depend on that unit's values of Y, X and Z. Note that in general this assumption does not hold, e.g. where the population values of Y and X are clustered. In any case the joint distribution of the bivariate random variable (I, Π) will depend on the *population* values of Z (and sometimes on those of Y and X as well), so an *iid* assumption for (Y, X, Z, I, Π) fails. However, in large populations the level of dependence between values of (I, Π) for different population units will be small given their respective values of Y, X and Z, and so this assumption will be a reasonable one. A similar assumption underpins the parametric estimation methods described in Chapter 12, and is, to some extent, justified by asymptotics described in Pfeffermann, Krieger and Rinott (1998).

11.2.2. What are the data?

The words 'complex survey data' mask the huge variety of forms in which survey data appear. A basic problem with any form of survey data analysis therefore is identification of the relevant data for the analysis.

The method used to select the sample will have typically involved a combination of complex sample design procedures, including multi-way stratification, multi-stage clustering and unequal probability sampling. In general, the information available to the survey data analyst about the sampling method can vary considerably and hence we consider below a number of alternative data scenarios. In many cases we are secondary analysts, unconnected with the organization that actually carried out the survey, and therefore denied access to sample design information on confidentiality grounds. Even if we are primary analysts, however, it is often the case that this information is not easily accessible because of the time that has elapsed since the survey data were collected.

What *is* generally available, however, is the value of the sample weight associated with each sample unit. That is, the weight that is typically applied to the value of a sample variable before summing these values in order to 'unbiasedly' estimate the population total of the variable. For the sake of simplicity, we shall assume that this sample weight is either the inverse of the sample inclusion probability π of the sample unit, or a close proxy. Our dataset therefore includes the sample values of these inclusion probabilities. This leads us to:

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Data scenario A: Sample values of Y, X and Π are known. No other information is available.

This scenario is our base scenario. We envisage that it represents the minimum information set where methods of data analysis which allow for complex sampling are possible. The methods described in Chapter 10 and Chapter 12 are essentially designed for sample data of this type.

The next situation we consider is where some extra information about how the sampled units were selected is also available. For example, if a stratified design was used, we know the strata to which the different sample units belong. Following standard practice, we characterize this information in terms of the values of a vector-valued design covariate Z known for all the sample units. Thus, in the case where only stratum membership is known, Z corresponds to a set of stratum indicators. In general Z will consist of a mix of such indicators and continuous size measures. This leads us to:

Data scenario B: Sample values of Y, X, Z and Π are known. No other information is available.

Note that Π will typically be related to Z. However, this probability need not be completely determined by Z.

We now turn to the situation where we not only have access to sample data, but also have information about the nonsampled units in the population. The extent of this information can vary considerably. The simplest case is where we have population summary information on Z, say the population average $\bar{z}_u$. Another type of summary information we may have relates to the sample inclusion probabilities Π. We may know that the method of sampling used corresponds to a fixed size design, in which case the population average of Π is n/N. Both these situations are combined in:

Data scenario C: Sample values of Y, X, Z and Π are known. The population average $\bar{z}_u$ of Z is known, as is the fact that the population average of Π is n/N.

Finally, we consider the situation where we have access to the values of both Z and Π for *all* units in the population, e.g. from a population frame. This leads to:

Data scenario D: Sample values of Y, X, Z and Π are known, as are the nonsample values of Z and Π.

11.2.3. Informative sampling and ignorable sample designs

A key concern of this chapter is where the sampling process somehow confounds standard methods for inference about the population characteristics of interest. It is a fundamental (and often unspoken) 'given' that such standard methods assume that the distribution of the sample data and the corresponding population distribution are the same, so inferential statements about the former

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

apply to the latter. However, with data collected via complex sample designs this situation no longer applies.

A sample design where the distribution of the sample values and population values for a variable Y differ is said to be *informative* about Y. Thus, if an unequal probability sample is taken, with inclusion probabilities proportional to a positive-valued size variable Z, then, provided Y and Z are positively correlated, the sample distribution of Y will be skewed to the right of its corresponding population distribution. That is, this type of unequal probability sampling is informative.

An extreme type of informative sampling discussed in Chapter 8 by Scott and Wild is case–control sampling. In its simplest form this is where the variable Y takes two values, 0 (a control) and 1 (a case), and sampling is such that all cases in the population (of which there are $n \ll N$) are selected, with a corresponding random sample of n of the controls also selected. Obviously the population proportion of cases is n/N. However, the corresponding sample proportion (0.5) is very different.

In some cases an informative sampling design may become uninformative given additional information. For example, data collected via a stratified design with nonproportional allocation will typically be distributed differently from the corresponding population distribution. This difference is more marked the stronger the relationship between the variable(s) of interest and the stratum indicator variables. Within a stratum, however, there may be no difference between the population and sample data distributions, and so the overall difference between these distributions is completely explained by the difference in the sample and population distributions of the stratum indicator variable.

It is standard to characterize this type of situation by saying a sampling method is *ignorable* for inference about the population distribution of a variable Y given the population values of another variable Z if Y is independent of the sample indicator I given the population values of Z. Thus, if Z denotes the stratum indicator referred to in the previous paragraph, and if sampling is carried out at random within each stratum, then it is easy to see that I and Y are independent within a stratum and so this method of sampling is ignorable given Z.

In the rest of this chapter we explore methods for fitting the population regression function $g_U(x)$ in situations where an informative sampling method has been used. In doing so, we consider both ignorable and nonignorable sampling situations.

11.3. RE-EXPRESSING THE REGRESSION FUNCTION

In this section we develop identities which allow us to re-express $g_U(x)$ in terms of sample-based quantities as well as quantities which depend on Z. These identities underpin the estimation methods defined in Section 11.4.

We use $f_U(w)$ to denote the value of the population density of a variable W at the value w, and $f_s(w)$ to denote the corresponding value of the sample density

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11, By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of this variable. This sample density is defined as the density of the conditional variable $W|I = 1$. See also Chapter 12. We write this (conditional) density as $f_s(w) = f_U(w|I = 1)$. To reduce notational clutter, conditional densities $f(w|V = v)$ will be denoted $f(w|v)$. We also use $E_U(W)$ to denote the expectation of W over the population (i.e. with respect to f_U) and $E_s(W)$ to denote the expectation of W over the sample (i.e. with respect to f_s). Since development of expressions for the regression of Y on one or more variables will be our focus, we introduce special notation for this case. Thus, the population and sample regressions of Y on another (possibly vector-valued) variable W will be denoted $g_U(w) = E_U(Y|W = w) = E_U(Y|w)$ and $g_s(w) = E_s(Y|W = w) = E_s(Y|w)$ respectively below.

We now state two identities. Their proofs are straightforward given the definitions of I and Π and the random indexing assumption of Section 11.2.1:

$$f_s(w|\pi) = f_U(w|\pi) \tag{11.1}$$

and

$$f_U(\pi) = f_s(\pi)E_U(\Pi)/\pi = f_s(\pi)/(\pi E_s(1/\Pi)). \tag{11.2}$$

Consequently

$$f_U(w) = \frac{\int \pi^{-1} f_s(w|\pi) f_s(\pi) \mathrm{d}\pi}{E_s\left[\Pi^{-1}\right]} \tag{11.3}$$

and so

$$E_U(W) = E_s\left[\Pi^{-1}E_s(W|\Pi)\right]/E_s\left[\Pi^{-1}\right].$$

Recollect that $g_U(x)$ is the regression of Y on X at $X = x$ in the population. Following an application of Bayes' theorem, one can then show

$$g_U(x) = \frac{E_s\left[\Pi^{-1} f_s(x|\Pi) g_s(x, \Pi)\right]}{E_s\left[\Pi^{-1} f_s(x|\Pi)\right]} \tag{11.4}.$$

From the right hand side of (11.4) we see that $g_U(x)$ can be expressed in terms of the ratio of two sample-based unconditional expectations. As we see later, these quantities can be estimated from the sample data, and a plug-in estimate of $g_U(x)$ obtained.

11.3.1. Incorporating a covariate

So far, no attempt has been made to incorporate information from the design covariate Z. However, since the development leading to (11.4) holds for arbitrary X, and in particular when X and Z are amalgamated, and since $g_U(x) = E_U(g_U(x, Z)|x)$, we can apply (11.4) twice to obtain

$$g_U(x) = \frac{E_s\left[\Pi^{-1} f_s(x|\Pi) E_s(g_U(x, Z)|x, \Pi)\right]}{E_s\left[\Pi^{-1} f_s(x|\Pi)\right]} \tag{11.5a}$$

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where

$$g_U(x,z) = \frac{E_s\left[\Pi^{-1} f_s(x,z|\Pi) g_s(x,z,\Pi)\right]}{E_s\left[\Pi^{-1} f_s(x,z|\Pi)\right]} \tag{11.5b}$$

An important special case is where the method of sampling is *ignorable* given Z; that is, the random variables Y and X are independent of the sample indicator I (and hence Π) given Z. This implies that $g_U(x,z) = g_s(x,z)$ and hence

$$g_U(x) = \frac{E_s\left[\Pi^{-1} f_s(x|\Pi) E_s(g_s(x,Z)|x,\Pi)\right]}{E_s\left[\Pi^{-1} f_s(x|\Pi)\right]}. \tag{11.6}$$

Under ignorability given Z, it can be seen that $E_s(g_s(x,Z)|x,\pi) = g_s(x,\pi)$, and hence (11.6) reduces to (11.4). Further simplification of (11.4) using this ignorability then leads to

$$g_U(x) = E_s\left[\Pi^{-1} f_s(x|Z) g_s(x,Z)\right] / E_s\left[\Pi^{-1} f_s(x|Z)\right], \tag{11.7}$$

which can be compared to (11.4).

11.3.2. Incorporating population information

The identities (11.4), (11.5) and (11.7) all express $g_U(x)$ in terms of sample moments. However, there are situations where we have access to population information, typically about Z and Π. In such cases we can weave this information into estimation of $g_U(x)$ by expressing this function in terms of estimable population and sample moments.

To start, note that

$$1/E_s\left[\Pi^{-1}|x\right] = E[\Pi f_s(x|\Pi)]/E[f_s(x|\Pi)]$$

and so we can rewrite (11.4) as

$$g_U(x) = \frac{E_s\left[\Pi^{-1} f_s(x|\Pi) g_s(x,\Pi)\right]}{E_s[f_s(x|\Pi)]} \frac{E_U[\Pi f_s(x|\Pi)]}{E_U[f_s(x|\Pi)]}. \tag{11.8}$$

The usefulness of this reexpression of (11.4) depends on whether the ratio of population moments on the right hand side of (11.8) can be evaluated from the available data. For example, suppose all we know is that $E_U(\Pi) = n/N$, and that $f_s(x|\Pi) = f_s(x)$. Here n is the sample size. Then (11.8) reduces to

$$g_U(x) = \frac{n}{N} E_s\left[\Pi^{-1} g_s(x,\Pi)\right].$$

Similarly, when population information on both Π and Z is available, we can replace (11.5) by

$$g_U(x) = \frac{E_s\left[\Pi^{-1} f_s(x|\Pi) E_s(g_U(x,Z)|x,\Pi)\right]}{E_s[f_s(x|\Pi)]} \frac{E_U[\Pi f_s(x|\Pi)]}{E_U[f_s(x|\Pi)]} \tag{11.9a}$$

where

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$g_U(x,z) = \frac{E_s\left[\Pi^{-1} f_s(x,z|\Pi) g_s(x,z,\Pi)\right]}{E_s[f_s(x,z|\Pi)]} \frac{E_U[\Pi f_s(x,z|\Pi)]}{E_U[f_s(x,z|\Pi)]}. \tag{11.9b}$$

The expressions above are rather complicated. Simplification does occur, however, when the sampling method is ignorable given Z. As noted earlier, in this case $g_U(x,z) = g_s(x,z)$, so $g_U(x) = E_U(g_s(x,Z)|x)$. However, since $f_U(x|z) = f_s(x|z)$ it immediately follows

$$g_U(x) = E_U[f_s(x|Z) g_s(x,Z)] / E_U[f_s(x|Z)]. \tag{11.10}$$

A method of sampling where $f_U(y|x) = f_s(y|x)$, and so $g_U(x) = g_s(x)$, is *noninformative*. Observe that ignorability given Z is not the same as being noninformative since it does not generally lead to $g_U(x) = g_s(x)$. For this we also require that the population and sample distributions of Z are the same, i.e. $f_U(z) = f_s(z)$.

We now combine these results on $g_U(x)$ obtained in the previous section with the data scenarios earlier described to develop estimators that capitalize on the extent of the survey data that are available.

11.4. DESIGN-ADJUSTED SMOOTHING

11.4.1. Plug-in methods based on sample data only

The basis of the plug-in approach is simple. We replace sample-based quantities in an appropriately chosen representation of $g_U(x)$ by corresponding sample estimates. Effectively this is a method of moments estimation of $g_U(x)$. Thus, in scenario A in Section 11.2.2 we only have sample data on Y, X and Π. The identity (11.4) seems most appropriate here since it depends only on the sample values of Y, X and Π. Our plug-in estimator of $g_U(x)$ is

$$\hat{g}_U(x) = \sum_s \pi_t^{-1} \hat{f}_s(x|\pi_t) \hat{g}_s(x,\pi_t) \Big/ \sum_s \pi_t^{-1} \hat{f}_s(x|\pi_t) \tag{11.11}$$

where $\hat{f}_s(x|\pi)$ denotes the value at x of a nonparametric estimate of the conditional density of the sample X-values given $\Pi = \pi$, and $\hat{g}_s(x,\pi)$ denotes the value at (x,π) of a nonparametric smooth of the sample Y-values against the sample X- and Π-values. Both these nonparametric estimates can be computed using standard kernel-based methods, see Silverman (1986) and Härdle (1990).

Under scenario B we have extra sample information, consisting of the sample values of Z. If these values explain a substantial part of the variability in Π, then it is reasonable to assume that the sampling method is ignorable given Z, and representation (11.7) applies. Our plug-in estimator of $g_U(x)$ is consequently

$$\hat{g}_U(x) = \sum_s \pi_t^{-1} f_s(x|z_t) \hat{g}_s(x,z_t) \Big/ \sum_s \pi_t^{-1} f_s(x|z_t). \tag{11.12}$$

If the information in Z is not sufficient to allow one to assume ignorability then one can fall back on the two-level representation (11.5). That is, one first computes an estimate of the population regression of Y on X and Z,

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\hat{g}_U(x,z) = \sum_s \pi_t^{-1} \hat{f}_s(x,z|\pi_t)\hat{g}_s(x,z,\pi_t) \Big/ \sum_s \pi_t^{-1}\hat{f}_s(x,z|\pi_t), \tag{11.13a}$$

and then smooths this estimate further (as a function of Z) against X and Π to obtain

$$\hat{g}_U(x) = \sum_s \pi_t^{-1} \hat{f}_s(x|\pi_t)\hat{E}_s(\hat{g}_U(x,Z)|x,\pi_t) \Big/ \sum_s \pi_t^{-1}\hat{f}_s(x|\pi_t) \tag{11.13b}$$

where $\hat{E}_s(\hat{g}_U(x,Z)|x,\pi_t)$ denotes the value at (x,π_t) of a sample smooth of the values $\hat{g}_U(x,z_t)$ against the sample X-and Π-values.

11.4.2. Examples

The precise form and properties of these estimators will depend on the nature of the relationship between Y, X, Z and Π. To illustrate, we consider two situations, corresponding to different sample designs.

Stratified sampling on Z

We assume a scenario B situation where Z is a mix of stratum indicators Z_1 and auxiliary covariates Z_2. We further suppose that sampling is ignorable within a stratum, so (11.12) applies. Let h index the overall stratification, with s_h denoting the sample units in stratum h. Then (11.12) leads to the estimator

$$\hat{g}_U(x) = \sum_h \sum_{t\in s_h} \pi_t^{-1}\hat{f}_{sh}(x|z_{2t})\hat{g}_{sh}(x,z_{2t}) \Big/ \sum_h \sum_{t\in s_h} \pi_t^{-1}\hat{f}_{sh}(x|z_{2t}) \tag{11.14}$$

where $\hat{f}_{sh}$ denotes a nonparametric density estimate based on the sample data from stratum h. In some circumstances, however, we will be unsure whether it is reasonable to assume ignorability given Z. For example, it could be the case that Π is actually a function of $Z = (Z_1, Z_2)$ and an unobserved third variable Z_3 that is correlated with Y and X. Here the two-stage estimator (11.13) is appropriate, leading to

$$\hat{g}_U(x, z_1 = h, z_2) = \hat{g}_h(x,z_2) = \frac{\sum_s \pi_t^{-1}\hat{f}_{sh}(x,z_2|\pi_t)\hat{f}_{sh}(\pi_t)\hat{g}_{sh}(x,z_2,\pi_t)}{\sum_s \pi_t^{-1}\hat{f}_{sh}(x,z_2|\pi_t)\hat{f}_{sh}(\pi_t)} \tag{11.15a}$$

and hence

$$\hat{g}_U(x) = \sum_s \pi_t^{-1}\hat{f}_s(x|\pi_t)\hat{E}_s(\hat{g}_{z_{1t}}(x,z_{2t})|x,\pi_t) \Big/ \sum_s \pi_t^{-1}\hat{f}_s(x|\pi_t) \tag{11.15b}$$

where $\hat{f}_{sh}(\pi)$ denotes an estimate of the probability that a sample unit with $\Pi = \pi$ is in stratum h, and $\hat{E}_s\big(\hat{g}_{z_1}(x,z_2)|x,\pi\big)$ denotes the value at (x,π) of a nonparametric smooth of the sample $\hat{g}_{z_1}(x,z_2)$-values defined by (11.15a) against the sample (X,Π)-values.

Calculation of (11.15) requires 'smoothing within smoothing' and so will be computer intensive. A further complication is that the sample $\hat{g}_{z_1}(x,z_2)$ values smoothed in (11.15b) will typically be discontinuous between strata, so that standard methods of smoothing may be inappropriate.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Array Sampling on X

Suppose the random variable X takes n distinct values $\{x_t; t = 1, 2, \ldots, n\}$ on the population U. Suppose furthermore that Z is univariate, taking m_t distinct (and strictly positive) values $\{z_{jt}; j = 1, 2, \ldots, m_t\}$ when $X = x_t$, and that we have $Y = X + Z$. The population values of Y and Z so formed can thus be thought of as defining an array, with each row corresponding to a distinct value of X. Finally suppose that the sampling method chooses one population unit for each value x_t of X (so the overall sample size is n) with probability

$$\pi_{jt} = z_{jt} \Big/ \sum_{j \in t} z_{jt} = z_{jt}/s_t.$$

Given this set-up, inspection of the sample data (which includes the values of Π) allows one to immediately observe that $g_s(x, z) = x + z$ and to calculate the realized value of s_t as z_{st}/π_{st} where z_{st} and π_{st} denote the sample values of Z and Π corresponding to $X = x_t$. Furthermore, the method of sampling is ignorable given Z, so $g_U(x, z) = g_s(x, z)$ and hence $g_U(x) = x + \mu$, where $\mu = E_U(Z)$. If the total population size N were known, an obvious (and efficient) estimator of μ would be

$$\hat{\mu} = \left[\sum_{t=1}^{n} m_t\right]^{-1} \left[\sum_{t=1}^{n} s_t\right] = N^{-1} \sum_{t=1}^{n} \pi_{st}^{-1} z_{st}$$

and so $g_U(x)$ could be estimated with considerable precision. However, we do not know N and so are in scenario B. The above method of sampling ensures that every distinct value of X in the sample is observed once and only once. Hence $\hat{f}_s(x|z) = 1/n$. Using (11.12), our estimator of $g_U(x)$ then becomes

$$\hat{g}_U(x) = \left[\sum_{t=1}^{n} \pi_{st}^{-1}\right]^{-1} \left[\sum_{t=1}^{n} \pi_{st}^{-1}(x + z_{st})\right] = x + \left[\sum_{t=1}^{n} \pi_{st}^{-1}\right]^{-1} \left[\sum_{t=1}^{n} \pi_{st}^{-1} z_{st}\right]. \tag{11.16}$$

This is an approximately unbiased estimator of $g_U(x)$. To see this we note that, by construction, each value of π_{st} represents an independent realization from a distribution defined on the values $\{\pi_{jt}\}$ with probabilities $\{\pi_{jt}\}$, and so $E_U(\pi_{st}^{-1}) = m_t$. Hence

$$E_U(\hat{g}_U(x)) \approx x + \left[\sum_{t=1}^{n} m_t\right]^{-1} \left[\mu \sum_{t=1}^{n} m_t\right] = x + \mu = g_U(x).$$

11.4.3. Plug-in methods which use population information

We now turn to the data scenarios where population information is available. To start, consider scenario C. This corresponds to having additional summary

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

population information, typically population average values, available for Z and Π. More formally, we know the value of the population size N and one or both of the values of the population averages $\bar{\pi}$ and $\bar{z}$.

How this information can generally be used to improve upon direct sample-based estimation of $g_U(x)$ is not entirely clear. The fact that this information *can* be useful, however, is evident from the array sampling example described in the previous section. There we see that, given the sample data, this function can be estimated very precisely once either the population mean of Z is known, or, if the method of sampling is known, we know the population size N. This represents a considerable improvement over the estimator (11.16) which is only approximately unbiased for this value.

However, it is not always the case that such dramatic improvement is possible. For example, suppose that in the array sampling situation we are not told (i) the sample values of Z and (ii) that sampling is proportional to Z within each row of the array. The only population information is the value of N. This is precisely the situation described following (11.8), and so we could use the estimator

$$\hat{g}_U(x) = N^{-1} \sum_{t=1}^{n} \pi_{st}^{-1} \hat{g}_s(x, \pi_{st}) \tag{11.17}$$

where $\hat{g}_s(x, \pi)$ is the estimated sample regression of Y on X and Π. It is not immediately clear why (11.17) should in general represent an improvement over (11.16). Note that the regression function $\hat{g}_s(x, \pi)$ in (11.17) is not 'exact' (unlike $g_s(x, z) = x + z$) and so (11.17) will include an error arising from this approximation.

Finally, we consider scenario D. Here we know the population values of Z and Π. In this case we can use the two-level representation (11.9) to define a plug-in estimator of $g_U(x)$ if we have reason to believe that sampling is not ignorable given Z. However, as noted earlier, this approach seems overly complex. Equation (11.10) represents a more promising alternative provided that the sampling is ignorable given Z, as will often be the case for this type of scenario (detailed population information available). This leads to the estimator

$$\hat{g}_U(x) = \sum_{t=1}^{N} \hat{f}_s(x|z_t) \hat{g}_s(x, z_t) \Big/ \sum_{t=1}^{N} \hat{f}_s(x|z_t). \tag{11.18}$$

Clearly, under stratified sampling on Z, (11.18) takes the form

$$\hat{g}_U(x) = \sum_{h} \sum_{t \in h} \hat{f}_{sh}(x|z_{2t}) \hat{g}_{sh}(x, z_{2t}) \Big/ \sum_{h} \sum_{t \in h} \hat{f}_{sh}(x|z_{2t}) \tag{11.19}$$

which can be compared to (11.14). In contrast, knowing the population values of Z under array sampling on X means we know the values m_t and s_t for each t and so we can compute a precise estimate of $g_U(x)$ from the sample data.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

11.4.4. The estimating equation approach

The starting point for this approach is Equation (11.4), which can be equivalently written

$$g_U(x) = E_s(\Pi^{-1} Y | X = x) / E_s(\Pi^{-1} | X = x) \tag{11.20}$$

providing $E_s(\Pi^{-1}|X = x) > 0$. That is, $g_U(x)$ can always be represented as the solution of the equation

$$E_s[\Pi^{-1}(Y - g_U(x))|X = x] = 0. \tag{11.21}$$

Replacing the left hand side of (11.21) by a kernel-based estimate of the regression function value leads to the estimating equation ($h_s(x)$ is the bandwidth)

$$\sum_s K\left(\frac{x - x_t}{h_s(x)}\right)\pi_t^{-1}(Y_t - \hat{g}_U(x)) = 0 \tag{11.22}$$

which has the solution

$$\hat{g}_U(x) = \sum_s K\left(\frac{x - x_t}{h_s(x)}\right)\pi_t^{-1} y_t \Big/ \sum_s K\left(\frac{x - x_t}{h_s(x)}\right)\pi_t^{-1}. \tag{11.23}$$

This is a standard Nadaraya–Watson-type estimator of the sample regression of Y on X, but with kernel weights modified by multiplying them by the inverses of the sample inclusion probabilities.

The Nadaraya–Watson nonparametric regression estimator is known to be inefficient compared to local polynomial alternatives when the sample regression function is reasonably smooth (Fan, 1992). Since this will typically be the case, a popular alternative solution to (11.22) is one that parameterizes $\hat{g}_U(x)$ as being *locally linear* in x. That is, we write

$$\hat{g}_U(x) = \hat{a}(x) + \hat{b}(x)x \tag{11.24}$$

and hence replace (11.22) by

$$\sum_s K\left(\frac{x - x_t}{h_s(x)}\right)\pi_t^{-1}(y_t - \hat{a}(x) - \hat{b}(x)x_t) = 0. \tag{11.25}$$

The parameters $\hat{a}(x)$ and $\hat{b}(x)$ are obtained by weighted local linear least squares estimation. That is, they are the solutions to

$$\sum_s \pi_t^{-1} K\left(\frac{x - x_t}{h_s(x)}\right)\left(y_t - \hat{a}(x) - \hat{b}(x)x_t\right)\begin{pmatrix}1\\ x_t\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}. \tag{11.26}$$

Following arguments outlined in Jones (1991) it can be shown that, for either (11.23) or (11.24), a suitable bandwidth $h_s(x)$, in terms of minimizing the mean squared error $E_s(\hat{g}_U(x) - g_U(x))^2$, must be of order $n^{-1/5}$.

One potential drawback with this approach is that there seems no straightforward way to incorporate population auxiliary information. These estimators

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

are essentially scenario A estimators. One ad hoc solution to this is to replace the inverse sample inclusion probabilities in (11.22) and (11.26) by sample weights which reflect this information (e.g. weights that are calibrated to known population totals). However, it is not obvious why this modification should lead to improved performance for either (11.23) or (11.24).

11.4.5. The bias calibration approach

Suppose $\hat{g}_s(x)$ is a standard (i.e. unweighted) nonparametric estimate of the sample regression of Y on X at x. The theory outlined in Section 11.3 shows that this estimate will generally be biased for the value of the population regression of Y on X at x if the sample and population regression functions differ. One way around this problem therefore is to nonparametrically bias-calibrate this estimate. That is, we compute the sample residuals, $r_t = y_t - \hat{g}_s(x_t)$, and re-smooth these against X using a methodology that gives a consistent estimator of their population regression on X. This smooth is then added to $\hat{g}_s(x)$. For example, if (11.11) is used to estimate this residual population regression, then our final estimate of $g_U(x)$ is

$$\hat{g}_U(x) = \hat{g}_s(x) + \frac{\sum_s \pi_t^{-1} \hat{f}_s(x|\pi_t)\hat{g}_{sR}(x, \pi_t)}{\sum_s \pi_t^{-1} \hat{f}_s(x|\pi_t)} \tag{11.27}$$

where $\hat{g}_{sR}(x, \pi)$ denotes the value at (x, π) of a sample smooth of the residuals r_t against the sample X and Π values. Other forms of (11.27) can be easily written down, based on alternative methods of consistently estimating the population regression of the residuals at x.

This approach is closely connected to the concept of 'twicing' or double smoothing for bias correction (Tukey, 1977; Chambers, Dorfman and Wehrly, 1993).

11.5. SIMULATION RESULTS

Some appreciation for the behaviour of the different nonparametric regression methods described in the previous section can be obtained by simulation. We therefore simulated two types of populations, both of size $N = 1000$, and for each we considered two methods of sampling, one noninformative and the other informative. Both methods had $n = 100$, and were based on application of the procedure described in Rao, Hartley and Cochran (1962), with inclusion probabilities as defined below.

The first population simulated was defined by the equations:

$$Y = 1 + X + XZ + \varepsilon_Y \tag{11.28a}$$

$$X = 4 + 0.5Z + \varepsilon_X \tag{11.28b}$$

$$Z = 4 + 2\varepsilon_Z \tag{11.28c}$$

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $\varepsilon_Y, \varepsilon_X$ and ε_Z are independent standard normal variates. It is straightforward to see that $g_U(x) = E_U(Y|X = x) = 1 - x + x^2$. Figure 11.1 shows a realization of this population. The two sampling procedures we used for this population were based on inclusion probabilities that were approximately proportional to the population values of Z (*PPZ*: an informative sampling method) and X (*PPX*: a noninformative sampling method). These probabilities were defined by

$$PPZ\colon \pi_t = 100(z_t + \min_U(z) + 0.1)\Big/\sum_{u=1}^{N}(z_u + \min_U(z) + 0.1)$$

and

$$PPX\colon \pi_t = 100(x_t + \min_U(X) + 0.1)\Big/\sum_{u=1}^{N}(x_u + \min_U(X) + 0.1)$$

respectively.

The second population type we simulated reflected the heterogeneity that is typical of many real populations and was defined by the equations

$$Y = 30 + X + 0.0005ZX^2 + 3\sqrt{X\varepsilon_Y} \tag{11.29a}$$

$$X = 20 + 100\eta_X \tag{11.29b}$$

where ε_Y is a standard normal variate, η_X is distributed as Gamma(2) independently of ε_Y, and Z is a binary variable that takes the values 1 and 0 with

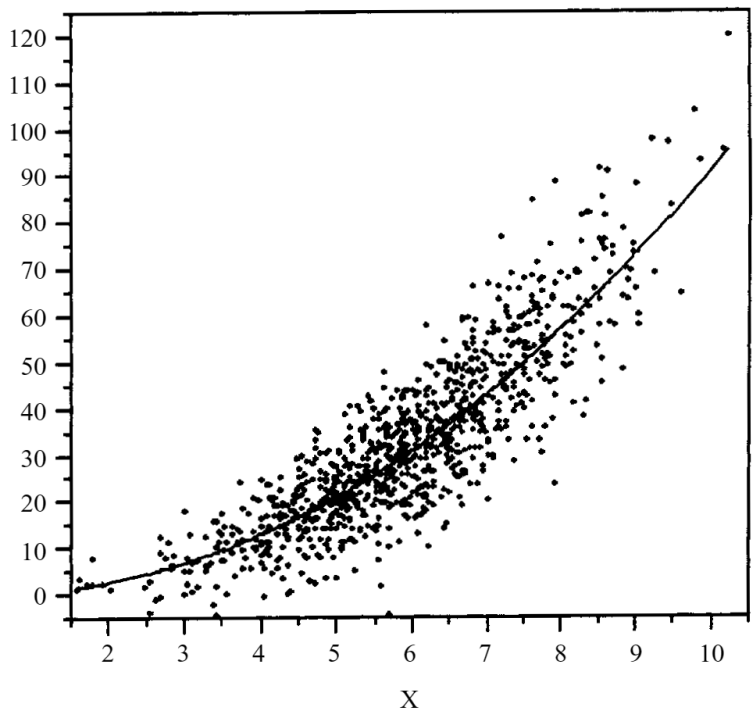

Figure 11.1 Scatterplot of Y vs. X for a population of type 1. Solid line is $g_U(x)$.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

probabilities 0.4 and 0.6 respectively, independently of both ε_Y and η_X. Figure 11.2 shows a realization of this population. Again, straightforward calculation shows that for this case $g_U(x) = E_U(Y|X = x) = 30 + x + 0.0002x^2$.

As with the first population type (11.28), we used two sampling methods with this population, corresponding to inclusion probabilities proportional to Z (*PPZ*: an informative sampling method) and proportional to X (*PPX*: a non-informative sampling method). These probabilities were defined by

$$PPZ: \pi_t = 100(z_t + 0.5) \Big/ \sum_{u=1}^{N} (z_u + 0.5)$$

and

$$PPX: \pi_t = 100x_t \Big/ \sum_{u=1}^{N} x_u$$

respectively. Note that *PPZ* above corresponds to a form of stratified sampling, in that all population units with $Z = 1$ have a sample inclusion probability that is three times greater than that of population units with $Z = 0$.

Each population type was independently simulated 200 times and for each simulation two samples were independently drawn using *PPZ* and *PPX* inclusion probabilities. Data from these samples were then used to fit a selection of the nonparametric regression smooths described in the previous section. These methods are identified in Table 11.1. The naming convention used in this table

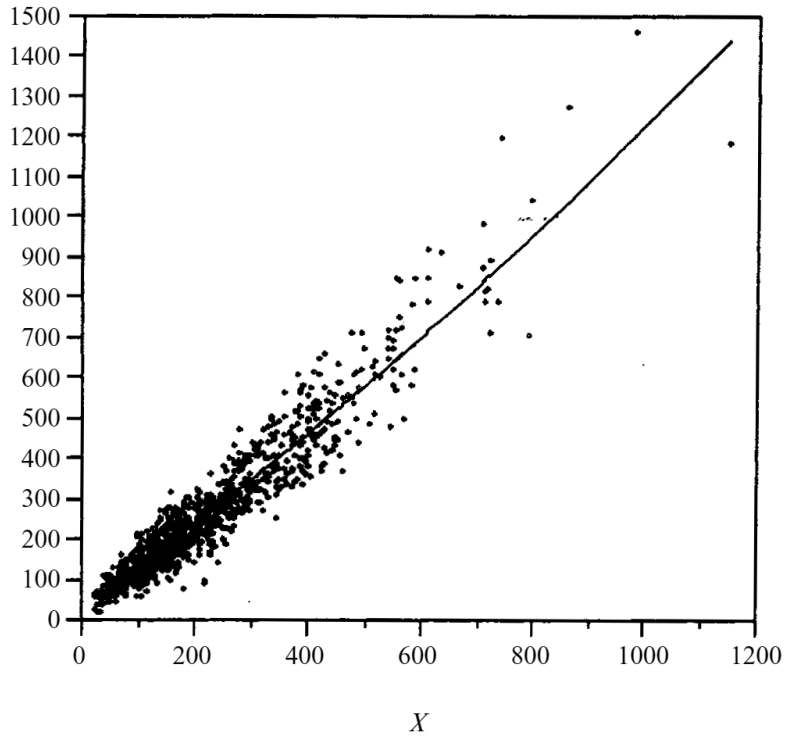

Figure 11.2 Scatterplot of Y vs. X for a population of type 2. Solid line is $g_U(x)$.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 11.1 Nonparametric smoothing methods evaluated in the simulation study.

Name	Estimator type	Definition
M(Π_s)	Scenario A plug-in	(11.11)
M($Z_s\Pi_s$)	Scenario B plug-in	(11.12)
M($Z_{strs}\Pi_{strs}$)	Stratified scenario B plug-in	(11.14)
M(Z)	Scenario D plug-in	(11.18)
M(Z_{str})	Stratified scenario D plug-in	(11.19)
Elin(Π_s)	Weighted local linear smooth	(11.24)
Elin + Elin(Π_s)	Unweighted+weighted smooth	(11.27[b])
Elin	Unweighted local linear smooth	(11.24[a])
Elin+Elin	Unweighted+unweighted smooth	(11.27[b])

[a] Based on (11.25) with $\pi_t = 1$.

[b] These are bias-calibrated or 'twiced' methods, representing two variations on the form defined by (11.27). In each case the first component of the method's name is the name of the initial (potentially biased) smooth and the second component (after '+') is the name of the second smooth that is applied to the residuals from the first. Note that the 'twiced' standard smooth Elin+Elin was only investigated in the case of *PPZ* (i.e. informative) sampling.

denotes a plug-in (moment) estimator by 'M' and one based on solution of an estimating equation by 'E'. The amount of sample information required in order to compute an estimate is denoted by the symbols that appear within the parentheses associated with its name. Thus, M(Π_s) denotes the plug-in estimator defined by (11.11) that requires one to know the sample values of Π and nothing more (the sample values of Y and X are assumed to be always available). In contrast, M(Z) is the plug-in estimator defined by (11.25) that requires knowledge of the population values of Z before it can be computed. Note the twiced estimator Elin + Elin. This uses an unweighted locally linear smooth at each stage. In contrast, the twiced estimator Elin + Elin(Π_s) uses a weighted local linear smoother at the second stage. We included Elin + Elin in our study for the *PPZ* sampling scenarios only to see how much bias from informative sampling would be soaked up just by reapplication of the smoother. We also investigated the performance of the Nadaraya–Watson weighted smooth in our simulations. However, since its performance was uniformly worse than that of the locally linear weighted smooth Elin(Π_s), we have not included it in the results we discuss below.

Each smooth was evaluated at the 5th, 6th, ..., 95th percentiles of the population distribution of X and two measures of goodness of fit were calculated. These were the mean error, defined by the average difference between the smooth and the actual values of $g_U(x)$ at these percentiles, and the root mean squared error (RMSE), defined by the square root of the average of the squares of these differences. These summary measures were then averaged over the 200 simulations. Tables 11.2 to 11.5 show these average values.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11, By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 11.2 Simulation results for population type 1 (11.28) under *PPZ* sampling.

	Bandwidth coefficient							
Method	0.5	1	2	3	4	5	6	7
	Average mean error							
M(Π_s)	0.20	−0.09	−0.31	−0.35	−0.17	0.09	0.29	0.39
M($Z_s\Pi_s$)	1.51	1.79	2.09	2.23	2.33	2.39	2.43	2.44
M(Z)	0.22	0.02	−0.36	−0.38	−0.18	0.08	0.26	0.35
Elin(Π_s)	0.28	0.24	0.40	0.64	0.84	0.93	0.95	0.91
Elin + Elin(Π_s)	0.27	0.17	0.13	0.17	0.25	0.36	0.45	0.52
Elin	1.85	1.91	2.13	2.38	2.56	2.64	2.64	2.59
Elin+Elin	1.82	1.83	1.86	1.91	1.99	2.09	2.18	2.24
	Average root mean squared error							
M(Π_s)	3.80	2.48	2.21	3.54	4.81	5.66	6.13	6.36
M($Z_s\Pi_s$)	14.08	5.03	2.93	2.65	2.65	2.71	2.77	2.83
M(Z)	3.36	2.28	2.16	3.54	4.79	5.60	6.03	6.24
Elin(Π_s)	2.50	1.94	1.48	1.40	1.43	1.48	1.54	1.61
Elin + Elin(Π_s)	2.95	2.27	1.67	1.44	1.33	1.26	1.25	1.29
Elin	2.93	2.57	2.45	2.60	2.74	2.84	2.88	2.90
Elin+Elin	3.30	2.77	2.36	2.26	2.67	2.32	2.40	2.48

Table 11.3 Simulation results for population type 1 (11.28) under *PPX* sampling.

	Bandwidth coefficient							
Method	0.5	1	2	3	4	5	6	7
	Average mean error							
M(Π_s)	0.01	0.04	0.05	−2.25	0.05	0.10	0.18	0.25
M($Z_s\Pi_s$)	−0.17	0.29	0.31	0.44	0.53	0.57	0.58	0.56
M(Z)	−1.13	−1.19	−1.24	−0.95	−0.52	−0.12	0.14	0.26
Elin(Π_s)	0.09	0.14	0.36	0.62	0.82	0.92	0.93	0.89
Elin + Elin(Π_s)	0.05	0.06	0.11	0.18	0.26	0.35	0.43	0.50
Elin	0.09	0.14	0.35	0.58	0.73	0.78	0.75	0.68
	Average root mean squared error							
M(Π_s)	3.79	2.93	1.92	4.34	1.41	1.37	1.44	1.55
M($Z_s\Pi_s$)	12.71	5.47	2.26	1.69	1.60	1.61	1.66	1.74
M(Z)	4.05	2.49	2.59	3.81	4.98	5.77	6.21	6.42
Elin(Π_s)	2.46	1.92	1.47	1.38	1.42	1.49	1.57	1.64
Elin + Elin(Π_s)	2.96	2.29	1.70	1.45	1.33	1.27	1.27	1.32
Elin	2.46	1.93	1.48	1.37	1.39	1.45	1.53	1.63

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 11.4 Simulation results for population type 2 (11.29) under *PPZ* sampling.

	Bandwidth coefficient							
Method	0.5	1	2	3	4	5	6	7
	Average mean error							
M(Π_s)	1.25	1.01	1.05	1.17	1.35	1.45	1.48	1.52
M($Z_s\Pi_s$)	7.90	9.49	8.97	9.05	9.40	9.74	10.11	10.46
M($Z_{strs}\Pi_{strs}$)	1.72	1.11	1.10	1.28	1.38	1.42	1.39	1.36
M(Z)	1.12	0.95	0.99	1.11	1.29	1.39	1.42	1.45
M(Z_{str})	1.67	1.06	1.06	1.25	1.35	1.39	1.37	1.33
Elin(Π_s)	1.52	1.29	1.44	1.58	1.70	1.80	1.86	1.93
Elin + Elin(Π_s)	−5.28	−5.57	−5.58	−5.99	−1.07	−6.02	−5.93	−5.81
Elin	8.06	8.28	8.90	9.41	9.75	9.99	10.15	10.30
Elin+Elin	0.20	0.21	0.30	0.41	0.43	0.51	0.64	0.78
	Average root mean squared error							
M(Π_s)	28.19	17.11	10.40	8.98	8.18	7.68	7.44	7.38
M($Z_s\Pi_s$)	75.21	41.54	23.44	16.65	15.03	14.92	15.21	15.79
M($Z_{strs}\Pi_{strs}$)	15.78	12.20	9.66	8.61	8.15	7.90	7.73	7.67
M(Z)	28.47	17.51	10.34	8.90	8.07	7.55	7.27	7.20
M(Z_{str})	15.75	12.12	9.56	8.54	8.07	7.79	7.61	7.52
Elin(Π_s)	15.41	12.77	10.58	9.49	8.87	8.54	8.39	8.41
Elin + Elin(Π_s)	17.60	14.86	12.38	11.43	10.82	10.40	10.13	9.90
Elin	17.15	15.48	14.50	14.36	14.46	14.72	15.05	15.47
Elin+Elin	16.05	12.94	9.91	8.72	8.00	7.58	7.33	7.21

All kernel-based methods used an Epanechnikov kernel. In order to examine the impact of bandwidth choice on the different methods, results are presented for a range of bandwidths. These values are defined by a bandwidth coefficient, corresponding to the value of b in the bandwidth formula

$$\text{bandwidth} = b \times \text{sample range of } X \times n^{-1/5}.$$

Examination of the results shown in Tables 11.2–11.5 shows a number of features:

1. RMSE-optimal bandwidths differ considerably between estimation methods, population types and methods of sampling. In particular, twicing-based estimators tend to perform better at longer bandwidths, while plug-in methods seem to be more sensitive to bandwidth choice than methods based on estimating equations.
2. For population type 1 with *PPZ* (informative) sampling we see a substantial bias for the unweighted methods Elin and Elin + Elin, and consequently large RMSE values. Somewhat surprisingly the plug-in method M($Z_s\Pi_s$) also displays a substantial bias even though its RMSE values are

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11, By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 11.5 Simulation results for population type 2 (11.29) under *PPX* sampling.

Method	Bandwidth coefficient 0.5	1	2	3	4	5	6	7
	Average mean error							
M(Π_s)	0.09	0.07	0.03	−0.09	−0.11	−0.09	−0.13	−0.19
M($Z_s\Pi_s$)	16.62	0.58	0.50	0.51	0.51	0.59	0.59	0.54
M($Z_{strs}\Pi_{strs}$)	0.36	0.47	0.72	0.80	0.82	0.75	0.60	0.34
M(Z)	−0.09	0.46	0.66	0.71	0.77	0.70	0.50	0.28
M(Z_{str})	0.31	0.42	0.66	0.74	0.76	0.71	0.56	0.29
Elin(Π_s)	0.27	0.49	0.83	1.06	1.19	1.27	1.32	1.37
Elin + Elin(Π_s)	−0.43	−0.37	−0.32	−0.18	−0.02	0.08	0.18	0.31
Elin	0.27	0.47	0.61	0.71	0.73	0.65	0.48	0.30
	Average root mean squared error							
M(Π_s)	20.38	14.44	10.03	8.26	7.37	7.01	6.97	7.10
M($Z_s\Pi_s$)	217.7	29.72	12.04	10.13	9.43	8.94	8.74	8.79
M($Z_{strs}\Pi_{strs}$)	12.06	9.67	7.89	7.18	6.92	6.97	7.19	7.51
M(Z)	15.94	10.23	7.88	7.01	6.62	6.52	6.66	6.99
M(Z_{str})	11.92	9.47	7.70	6.97	6.71	6.75	6.96	7.26
Elin(Π_s)	12.79	9.90	7.82	7.08	6.69	6.49	6.42	6.44
Elin + Elin(Π_s)	15.76	12.49	9.52	8.40	7.61	7.08	6.75	6.58
Elin	12.79	9.94	8.03	7.42	7.16	7.10	7.23	7.51

not excessive. Another surprise is the comparatively poor RMSE performance of the plug-in method M(Z) that incorporates population information. Clearly the bias-calibrated method Elin + Elin(Π_s) is the best overall performer in terms of RMSE, followed by Elin(Π_s).

3. For population type 1 with *PPX* (noninformative) sampling there is little to choose between any of the methods as far as bias is concerned. Again, we see that the plug-in method M(Z) that incorporates population information is rather unstable and the overall best performer is bias-calibrated method Elin + Elin(Π_s). Not unsurprisingly, for this situation there is virtually no difference between Elin and Elin(Π_s).
4. For population type 2 and *PPZ* (informative) sampling there are further surprises. The plug-in method M($Z_s\Pi_s$), the standard method Elin and the bias-calibrated method Elin + Elin(Π_s) all display bias. In the case of M($Z_s\Pi_s$) this reflects behaviour already observed for population type 1. Similarly, it is not surprising that Elin is biased under informative sampling. However, the bias behaviour of Elin + Elin(Π_s) is hard to understand, especially when compared to the good performance of the unweighted twiced method Elin + Elin. In contrast, for this population all the plug-in methods (with the exception of M($Z_s\Pi_s$)) work well. Finally

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

we see that in this population longer bandwidths are definitely better, with the optimal bandwidth coefficients for all methods except Elin and M($Z_s\Pi_s$) most likely greater than $b = 7$.

5. Finally, for population type 2 and *PPX* (noninformative) sampling, all methods (with the exception of M($Z_s\Pi_s$)) have basically similar bias and RMSE performance. Although the bias performance of M($Z_s\Pi_s$) is unremarkable we see that in terms of RMSE, M($Z_s\Pi_s$) is again a relatively poor performer. Given the *PPX* sampling method is noninformative, it is a little surprising that the unweighted smoother Elin performs worse than the weighted smoother Elin(Π_s). In part this may be due to the heteroskedasticity inherent in population type 2 (see Figure 11.2), with the latter smoother giving less weight to the more variable sample values.
6. The preceding analysis has attempted to present an overview of the performance of the different methods across a variety of bandwidths. In practice, however, users may well adopt a default bandwidth that seems to work well in a variety of situations. In the case of local linear smoothing (the underlying smoothing method for all the methods for which we present results), this corresponds to using a bandwidth coefficient of $b = 3$. Consequently it is also of some interest to just look at the performances of the different methods at this bandwidth. Here we see a much clearer picture. For population type 1, Elin(Π_s) is the method of choice, while M(Z_{str}) is the method of choice for population type 2.

It is not possible to come up with a single recommendation for what design-adjusted smoothing method to use on the basis of these (very limited) simulation results. Certainly they indicate that the bias-calibrated method Elin + Elin(Π_s), preferably with a larger bandwidth than would be natural, seems an acceptable general purpose method of nonparametrically estimating a population regression function. However, it can be inefficient in cases where plug-in estimators like M($Z_{strs}\Pi_{strs}$) and M(Z_{str}) can take advantage of nonlinearities in population structure caused by stratum shifts in the relationship between Y and X. In contrast, the plug-in method M($Z_s\Pi_s$) that combines sample information on both Π and Z seems too unstable to be seriously considered, while both the basic plug-in method M(Π_s) and the more complex population Z-based M(Z) are rather patchy in their performance – both are reasonable with population type 2, but are rather unstable with population type 1. In the latter case this seems to indicate that it is not always beneficial to include auxiliary information into estimation of population regression functions, even when the sampling method (like *PPZ*) is ignorable given this auxiliary information.

11.6. TO WEIGHT OR NOT TO WEIGHT? (WITH APOLOGIES TO SMITH, 1988)

The results obtained in the previous section indicate that some form of design adjustment is advisable when the sampling method is informative. However,

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11, By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

adopting a blanket rule to always use a design-adjusted method may lead to loss of efficiency when the sampling method is actually noninformative. On the other hand it is not always the case that inappropriate design adjustment leads to efficiency loss, compare Elin and Elin(Π_s) in Table 11.5.

Is there a way of deciding whether one should use design-adjusted (e.g. weighting) methods? Equivalently, can one, on the basis of the sample data, check whether the sample is likely to have been drawn via an informative sampling method?

A definitive answer to this question is unavailable at present. However, we describe two easily implementable procedures that can provide some guidance.

To start, we observe that if a sampling method is noninformative, i.e. if $f_U(y|x) = f_s(y|x)$, then design adjustment is unnecessary and we can estimate $g_U(x)$ by simply carrying out a standard smooth of the sample data. As was demonstrated in Pfeffermann, Krieger and Rinott (1998) and Pfeffermann and Sverchkov (1999), noninformativeness is equivalent to conditional independence of the sample inclusion indicator I and Y given X in the population, which in turn is equivalent to the identity

$$\begin{aligned} E_s(\Pi^{-1}|X = x, Y = y) &= E_U(\Pi|X = x, Y = y) \\ &= E_U(\Pi|X = x) = E_s(\Pi^{-1}|X = x). \end{aligned}$$

This identity holds if the sample values of Π and Y are independent given X. Hence, one way of checking whether a sample design is noninformative is to test this conditional independence hypothesis using the sample data. If the sample size is very large, this can be accomplished by partitioning the sample distributions of Π, Y and X, and performing chi-square tests of independence based on the categorized values of Π and Y within each category of X. Unfortunately this approach depends on the overall sample size and the degree of categorization, and it did not perform well when we applied it to the sample data obtained in our simulation study.

Another approach, and one that reflects practice in this area, is to only use a design-adjusted method if it leads to significantly different results compared to a corresponding standard (e.g. unweighted) method. Thus, if we let $\hat{g}_s(x)$ denote the standard estimate of $g_U(x)$, with $\hat{g}_U(x)$ denoting the corresponding design-adjusted estimate, we can compute a jackknife estimate $v_J(x)$ of $\text{var}(\hat{g}_s(x) - \hat{g}_U(x))$ and hence calculate the standardized difference

$$W(x) = [\hat{g}_s(x) - \hat{g}_U(x)] \Big/ \sqrt{v_J(x)}. \tag{11.30}$$

We would then only use $\hat{g}_U(x)$ if $W(x) > 2$. A multivariate version of this test, based on a Wald statistic version of (11.30), is easily defined. In this case we would be testing for a significant difference between these two estimates at k different x-values of interest. This would require calculation of a jackknife estimate of the variance–covariance matrix of the vector of differences between the g-values for the standard method and those for the design-adjusted method at these k different x-values. The Wald statistic value could then be compared to, say, the 95th percentile of a chi-squared distribution with k degrees of

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

freedom. Table 11.6 shows how this approach performed with the sample data obtained in our simulations. We see that it works well at identifying the fact that the *PPX* schemes are noninformative. However, its ability to detect the informativeness of the *PPZ* sampling schemes is not as good, particularly in the case of population type 2. We hypothesize that this poor performance is due to the heteroskedasticity implicit in this population's generation mechanism.

The problem with the Wald statistic approach is that it does not test the hypothesis that is really of interest to us in this situation. This is the hypothesis that $g_U(x) = g_s(x)$. In addition, nonrejection of the noninformative sampling 'hypothesis' may result purely because the variance of the design-adjusted estimator is large compared to the bias of the standard estimator.

A testing approach which focuses directly on whether $g_U(x) = g_s(x)$ can be motivated by extending the work of Pfeffermann and Sverchkov (1999). In particular, from (11.21) we see that this equality is equivalent to

$$E_s[\Pi^{-1}(Y - g_s(x))|X = x] = 0. \tag{11.31}$$

The twicing approach of Section 11.4.5 estimates the left hand side of (11.31), replacing $g_s(x)$ by the standard sample smooth $\hat{g}_s(x)$. Consequently, testing $g_U(x) = g_s(x)$ is equivalent to testing whether the twicing adjustment is significantly different from zero.

A smooth function can be approximated by a polynomial, so we can always write

$$E_s(\Pi^{-1}(Y - g_s(x))|X = x) = \sum_{j \geq 0} a_j x^j. \tag{11.32}$$

Table 11.6 Results for tests of informativeness. Proportion of samples where null hypothesis of noninformativeness is rejected at the (approximate) 5% level of significance.

	PPX sampling (noninformative)		*PPZ* sampling (informative)	
Test method	Population type 1	Population type 2	Population type 1	Population type 2
Wald statistic test[a]	0.015	0.050	1.000	0.525
Correlation test[b]	0.010	0.000	0.995	0.910

[a] The Wald statistic test was carried out by setting $\hat{g}_s = \text{Elin}$ and $\hat{g}_U = \text{Elin}(\Pi_s)$, both with bandwidth coefficient $b = 3$. The x-values where these two estimators were compared were the deciles of the sample distribution of X.

[b] The correlation test was carried out using $\hat{g}_s = \text{Elin}$ with bandwidth coefficient $b = 3$. A sample was rejected as being 'informative' if any *one* of the subhypotheses H_0, H_1 and H_2 was rejected at the 5% level.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11. By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We can test H: $g_U(x) = g_s(x)$ by testing whether the coefficients a_j of the regression model (11.32) are identically zero. This is equivalent to testing the sequence of subhypotheses

$$H_0: cor_s(\Pi^{-1}, (Y - g_s(X))) = 0$$
$$H_1: cor_s(\Pi^{-1}(Y - g_s(X)), X) = 0$$
$$H_2: cor_s(\Pi^{-1}(Y - g_s(X)), X^2) = 0$$
$$\vdots$$

In practice $g_s(x)$ in these subhypotheses is replaced by $\hat{g}_s(x)$. Assuming normality of all quantities, a suitable test statistic for each of these subhypotheses is then the corresponding t-value generated by the *cor.test* function within S-Plus (Statistical Sciences, 1990). We therefore propose that the hypothesis H be rejected at a '95% level' if any *one* of the absolute values of these t-statistics exceeds 2. Table 11.6 shows the results of applying this procedure to the sample data in our simulations, with the testing restricted to the subhypotheses H_0, H_1 and H_2. Clearly the test performs rather well across all sampling methods and population types considered in our simulations. However, even in this case we see that there are still problems with identifying the informativeness of *PPZ* sampling for population type 2.

11.7. DISCUSSION

The preceding development has outlined a framework for incorporating information about sample design, sample selection and auxiliary population information into nonparametric regression applied to sample survey data. Our results provide some guidance on choosing between the different estimators that are suggested by this framework.

An important conclusion that can be drawn from our simulations is that it pays to use some form of design-adjusted method when the sample data have been drawn using an informative sampling scheme, and there is little loss of efficiency when the sampling scheme is noninformative. However, we see that there is no single design-adjusted method that stands out. Local linear smoothing incorporating inverse selection probability weights seems to be an approach that provides reasonable efficiency at a wide range of bandwidths. But locally linear plug-in methods that capitalize on 'stratum structure' in the data can improve considerably on such a weighting approach. For nonparametric regression, as with any other type of analysis of complex survey data, it helps to find out as much as possible about the population structure the sample is supposed to represent. Unfortunately there appears to be no general advantage from incorporating information on an auxiliary variable into estimation, and in fact there can be a considerable loss of efficiency probably due to the use of higher dimensional smoothers with such data.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

An important ancillary question we have also considered is identifying situations where in fact the sampling method is informative. Here an examination of the correlations between weighted residuals from a standard (unweighted) nonparametric smooth and powers of the X variable is a promising diagnostic tool.

Our results indicate there is a need for much more research in this area. The most obvious is development of algorithms for choosing an optimal bandwidth (or bandwidths) to use with the different design-adjusted methods described in this chapter. In this context Breunig (1999, 2001) has investigated optimal bandwidth choice for nonparametric density estimation from stratified and clustered sample survey data, and it should be possible to apply these ideas to the nonparametric regression methods investigated here. The extension to clustered sample data is an obvious next step.

10.1002/0470867205.ch11, Downloaded from https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/0470867205.ch11 By Harvard University, Wiley Online Library on [12/11/2023]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License