Nonparametric Kernel Regression using Complex Survey Data

JOB MARKET PAPER

Luc Clair*
z34lmc@mun.ca

November 14, 2016

Abstract

Applied econometric analysis is often performed using data collected from large-scale surveys. These surveys use complex sampling plans in order to reduce costs and increase the estimation efficiency for subgroups of the population. These sampling plans result in unequal inclusion probabilities across units in the population and dependent observations within the sample, violating the assumption that the data is independently and identically distributed. The purpose of this chapter is to derive the asymptotic properties of a probability weighted nonparametric regression estimator under a combined inference framework. The nonparametric regression estimator considered is the local constant estimator. This work contributes to the literature in three ways. First, it derives the asymptotic properties for the multivariate mixed data case, including the asymptotic normality of the estimator. Second, I consider settings where the data in the population are i.i.d. and weakly correlated within clusters and show that the leading terms for the MSE are the same in both cases. Finally, I use least squares cross-validation for selecting the bandwidth for both continuous and discrete variables. I run Monte Carlo simulations designed to assess the finite-sample performance of the probability weighted local constant estimator versus the traditional local constant estimator for three sampling methods: simple random sampling, exogenous stratification, and endogenous stratification. Simulation results show that the estimator is consistent and that efficiency gains can be achieved by weighting observations by the inverse of their inclusion probabilities if the sampling is endogenous.

Keywords: Nonparametric regression, complex surveys, combined inference.

*I am grateful for the input and guidance I received from Dr. Jeff Racine, Dr. Jerry Hurley, and Phil DeCicca throughout my coursework and research. I would also like to thank Dr. Arthur Sweetman and Dr. Katherine Cuff for their feedback and general support. Furthermore, I would like to thank Dr. Rick Audas and the researchers from the Community-Based Primary Health Care team (CIHR: http://www.cihr-irsc.gc.ca/e/47153.html) for their support during the course of this research.

1 Introduction

In the statistics literature, auxiliary variables refer to additional independent or predictor variables in a regression analysis. These variables offer additional information and may be used to improve estimation of population parameters. Microeconometric research is frequently performed using data collected by surveys, which contain auxiliary information for every unit of the population of interest (Breidt & Opsomer 2000). Many of these surveys use complex sampling plans in order to reduce costs and increase the estimation efficiency for subgroups of the population. Each individual i in the population has a probability π_i of being included in the sample and this probability of being surveyed depends on what sampling methods are implemented. Complex sampling designs lead to unequal sampling probabilities for the units in the sample and create data with correlations between observations, violating the assumption that the data are independently and identically distributed (i.i.d.).

Survey datasets include a weight variable, w_i, which is interpreted as the number of people the respondent represents in the total population, with larger weights given to individuals who belong to groups that are sampled less frequently. As an example, the Canadian Community Health Survey includes a master weight variable WTS_M in the last column of the dataset and provides detailed methodology for computing the weight, including sampling design. The survey weight for each unit in the sample is equal to the inverse of the inclusion probability, i.e. $w_i = \pi_i^{-1}$.

The decision to include these weights in one's analysis has been a source of confusion, even for accomplished researchers (Solon, Haider & Wooldridge 2013). Many researchers adopt a model-based approach where they assume the data (y, x) is generated according to a given model, which describes the relationship between y and the auxiliary variables, $x = \{x_1, ..., x_q\}$, for all units of the finite population. Asymptotic results are based on the assumption that the population is itself a subset of nested superpopulations that grow to infinity. Once a model has been selected, usual inferential methods can then be applied (Lohr 2010).

More recently, economists have begun to specify when the use of survey weights is appropriate (Solon et al. 2013, Lohr 2010, Cameron & Trivedi 2009). If one's goal is to generate descriptive

statistics that will be used for public policy, including population totals and averages, it is recommended that weights be used. Solon et al. (2013) described a sample of units with unequal probability of inclusion as viewing the reflection of a representative sample through a 'funhouse mirror,' where oversampled subgroups will be exaggerated. Using weights clarifies the image and returns more precise estimates.

For estimating causal effects, the need to incorporate sampling design depends on the sampling criterion, the variable by which the sampling scheme was designed. As an example, consider a survey in which the sample was drawn based on income levels in which low income individuals were sampled at a higher rate than other individuals. If income is an explanatory variable, the sampling is exogenous and inclusion of survey weights will not improve estimation, and instead may reduce efficiency (Cameron & Trivedi 2009, p. 820). If, however, income is the dependent variable, survey weights should be included to control for *endogenous sampling*. Endogenous sampling is present when the sampling criterion is related to the error term; including survey weights is needed to ensure consistent results (Magee, Robb & Burbidge 1998). In this case, the use of model-based estimators will result in inconsistent estimates. Furthermore, by adopting a superpopulation model, a researcher is making the strong assumption that the model applies to population units who are not in the sample. Surveys contain a limited amount of auxiliary information, making it possible that key variables related to y were unobserved by the survey, and therefore, it is unlikely a theoretical model that holds for all observations exists. The inclusion probabilities then hold additional information for estimating regression parameters (Lohr 2010, p. 450).

In situations that warrant the use of survey weights, researchers turn to design-based methods. Purely design-based estimators make no modelling assumptions about the finite population of interest and are therefore free of misspecification. Inferences are based on the probability distribution induced by repeated sampling from the finite population, and the probability structure used for inference is that defined by a specified random sampling plan (Buskirk & Lohr 2005). Purely design-based estimators use only the design information and ignore additional auxiliary information provided for the sample. In order to utilize the auxiliary information, model-assisted estimators have been developed that retain superpopulation model assumptions but estimation is based on

survey design. Model-assisted estimators often resemble model-based estimators, only offset by the sample weight. Since model-assisted estimators include survey weights, they fall under the category of design-based estimators.

The properties of model-assisted estimators can be derived in three seetings: with model-based inference, with design-based inference, and with a combined inference. Model-based inference ignores sample design and results are the same as the i.i.d. case. Design-based inference only considers the random sampling plan and is not influenced by any relationship between y and x. Design consistency implies that the estimator approaches the true value after repeated sampling (Breidt & Opsomer 2000). In the combined apporach, it is assumed that a finite population is generated based on a selected model, where the predictor variables and outcome variable are assumed to follow a joint probability distribution. Then, a sample is drawn from this population based on a probability sampling design (Pfeffermann 1993). Model-assisted estimators are constructed so that the finite population quantities are estimated using design-consistent estimators, which are then used to estimate superpopulation parameters (Buskirk & Lohr 2005). This type of estimation can be thought of having two stages: a model stage and a design stage. In the model stage, a model is selected based on the belief that it has generated the population. Nonparametric regression models are attractive in this case as they are consistent under minimal restrictions on the underlying function. The relationship between the outcome and predictor variables is estimated in the design stage; a sample is drawn according to a specified sampling plan and the corresponding weights are included in the model.

Design-based methods have been well-developed in a parametric framework and have received growing attention in the nonparametric framework. The study of kernel density estimation with complex survey data has offered important insights (Breunig 2008, Buskirk & Lohr 2005, Breunig 2001, Bellhouse & Stafford 1999). Breunig (2001) and (2008) developed methods for kernel density estimation using data from clustered and stratified samples, respectively. Breunig (2008)'s approach was to divide the sample into strata, estimate the within stratum density, and sum across strata. Similarly, Breunig (2001) summed within cluster densities across clusters, including the inverse of the inclusion probability as the weight. While the within cluster inclusion probabilities may not be

available in all surveys, Breunig's (2011) paper highlighted a common problem found in econometric analysis: dealing with clustered data adds a degree of complexity as units within clusters may be correlated. Buskirk and Lohr (2005) derived the properties for the kernel density estimator for stratified samples under the three modes of inference outlined above. For the combined approach, they allowed for clustering in the population, resulting in dependent sample data. Their approach towards introducing within-cluster dependence was to assume the finite population was generated by a one-way random effects model.

Traditionally, population totals and averages were the quantities of interest in survey sampling. The population total, t_y, is simply the sum of y_i across all N units of a population, U, for a given variable y: $t_y = \sum_i^N y_i$. After taking a sample s of size n from the population, t_y can be estimated by $\hat{t}_y = \sum_{i \in s} y_i$. If the sample was designed to target subpopulations and is systematically unrepresentative of the population, $\hat{t}_y$ is biased and inconsistent. The Horvitz-Thompson estimator, $\hat{t}_{y,HT}$, was developed to estimate population totals incorporating sampling design (Horvitz & Thompson 1952). The Horvitz-Thompson estimator simply weights each unit by the inverse of their inclusion probability:

$$\hat{t}_{y,HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{1.1}$$

While this estimator is design unbiased and design consistent, it is a pure design-based estimator and ignores any available auxiliary variables. This motivated Breidt and Opsomer (2000), who estimated population totals using nonparametric regression techniques. Their estimator had the form:

$$\hat{t}_{y,BO} = \sum_{i=1}^{n} \frac{y_i - g_i}{\pi_i} + \sum_{i=1}^{N} g_i \tag{1.2}$$

and

$$g_i = e_1^T (X_{Ui}^T W_{Ui} X_{Ui})^{-1} X_{Ui}^T W_{Ui} y_U \tag{1.3}$$

the local polynomial estimator for the population. X is univariate and continuous and the subscript U denotes a population value. When working with survey data; however, one only has data from the sample and not the population. In order to estimate g_i, Breidt and Opsomer used a modified

local polynomial estimator that included a survey weight component in the weighting matrix:

$$\hat{g}_i = e_1^T (X_{si}^T W_{si} X_{si})^{-1} X_{si}^T W_{si} y_s \tag{1.4}$$

where $W_{si} = \text{diag}\{(\pi_j h)^{-1} K((x_j - x_i)/h)\}$, $K(\cdot)$ is a kernel function, and h is the bandwidth. The authors proved that this estimator is asymptotically design unbiased and consistent. This estimator was later revisited by Opsomer and Miller (2005) who provided a data-driven method for selecting the bandwidth for the local polynomial regression component under a design-based inference framework. The authors tested this estimator only in i.i.d. settings, which is unlikely to hold for data collected in complex surveys. Harms and Duchesne (2009) derived the asymptotic properties of the model-assisted local linear estimator under the combined inference framework. They showed that the bias of $\hat{g}(\cdot)$ is the same as in the i.i.d. case but the variance equaled that from the i.i.d. case multiplied by a correction factor derived from the sampling scheme. Using a modified plug-in method for selecting the bandwidth based on the MSE criterion, they examined the performance of $\hat{g}$ under exogenous and endogenous stratification. Unfortunately, this method only applies to the one continuous predictor variable case and cannot be used in the mixed variable case. Sánchez-Borrego et al. (2014) extended $\hat{t}_{y,BO}$ for the mixed variable case, estimating g_i using a modified local constant estimator.[1] This estimator was only evaluated under i.i.d. data and the bandwidth was selected using plug-in values and a survey cross-validation criterion which selected the optimal set of bandwidths from the plug-in values.

This chapter develops the asymptotic properties of the estimator proposed by Sánchez-Borrego et al. (2014) under the combined inference framework and tests the performance of the estimator against the traditional model-based local constant estimator. My research is guided by the following question: Does the use of survey weights improve the mean squared error (MSE) compared to models that do not take into account sampling design?

A common assumption when adopting a combined inference framework is that the realizations of the predictor variables in the superpopulation are i.i.d. (Harms & Duchesne 2010, Bellhouse

[1] Local polynomial with degree zero.

& Stafford 1999). This assumption may not be appropriate for survey sampling as it rules out clustering in the superpopulation. Clusters are subgroups of the population in which member units are likely to share characteristics. It is reasonable to assume that there is dependence between observations within the same cluster. Therefore, I consider two cases for the superpopulation model: one where the realization of the predictor variables are i.i.d. and one where the data are weakly dependent within clusters. Weak dependence implies that the dependence between two observations i and $i+\tau$ goes to zero as τ grows to infinity (Li & Racine 2007). At the design-stage for both cases, I assume that a sample is drawn based on a complex sampling plan.

As in any nonparametric setting the properties of this regression estimator depends on the choice of smoothing parameters; i.e. bandwidths for kernel estimators. It is preferable to adopt data-driven approaches for the selection of smoothing parameters (rather than subjective or ad hoc ones); therefore, I propose using least-squares cross-validation for selecting the bandwidths of both continuous and discrete regressors. To test the finite-sample properties of the modified local constant estimator, I run Monte Carlo simulations for four data generating processes (DGP) under three sampling schemes: simple random sampling (SRS), exogenous stratification, and endogenous stratification.

The rest of this paper is divided into eight sections. Following this introduction, I provide an overview of sampling methods used in complex surveys and how to derive the overall inclusion probability and weight for each unit. Next, I summarize parametric methods for analyzing complex survey data, including weighted least squares (WLS) and other methods for dealing with endogenous stratification and clustering. Section four derives the properties of the estimator developed by Sánchez-Borrego et al. (2014) under the combined framework. I show that the bias under the combined inference framework has the same leading terms as the model-based inference framework for both i.i.d. and cluster populations. The variance is equal to that under model-based inference multiplied by a correction factor, with the leading terms being equal for both the i.i.d. and cluster population settings. Section five presents the cross-validation method used for selecting the bandwidth. Section six presents the simulation results for SRS, exogenous stratification, and endogenous stratification. Results show that efficiency gains can be made by including survey weights when

the sampling is endogenous. Furthermore, by varying the sample size, I provide evidence that the modified local constant is consistent. The application in section seven uses Survey of Labour and Income Dynamics data to look at the relationship between labour market duration and age, controlling for gender. Finally, I conclude with a brief summary.

2 Sampling methods

The three main categories of sampling methods are: SRS, stratified sampling, and clustered sampling. Survey designers utilize the latter two methods in order to reduce the costs of implementing the survey and increase the estimation efficiency for subgroups of the population of interest. By using these methods, however, the probability of each object being sampled may not be equal across observations. The most basic sampling technique is SRS without replacement. Without replacement implies that no unit can be selected twice or more, which is common in survey situations. The prefix "simple" is added because other sampling schemes include random design elements (Cameron & Trivedi 2009). The following describes different sampling techniques used in probability sampling and how inclusion probabilities are calculated for each method.

2.1 SRS without replacement

A SRS of size n from a population of size N is one where observations are drawn from the population at random and with equal probability. There are $P(N,n) = N!/(N-n)!$ samples of size n in the population, where "!" denotes the factorial operation. To calculate the inclusion probabilities, π_i, one simply divides the sample size by the size of the population:

$$\pi_i = \pi = \frac{n}{N}$$

for this reason, SRSs are called *self-weighting*. If SRS without replacement is used, then it is reasonable to assume that the data (y, x) are i.i.d.. In this case, we can use model-based estimators without loss of consistency.

Example 2.1 (Weighted Least Squares)**.** *The WLS estimator is of the form:*

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W Y \tag{2.1}$$

In the context of sampling, the weight matrix W is derived based on the sampling design, i.e. $W = diag(\pi_i^{-1})$. If SRS is used, $\pi_i = \pi$ and $\hat{\beta}_{WLS} = \hat{\beta}_{OLS}$.

SRS is rarely used in large surveys because it would be too costly to implement. Instead, researchers divide the population into strata and/or clusters and take SRSs within these subgroups.

2.2 Stratification

Stratification uses supplementary information to help design samples. If characteristics differ between subpopulations, it is possible to obtain more precise estimates of population quantities by dividing the population into subgroups, called strata, and taking a probability sample from each subgroup. These strata are mutually exclusive, i.e. each sample unit can belong to only one stratum. By dividing the population based on similar characteristics, the units within each stratum are more homogeneous than the rest of the population. Common variables used for stratification include age, gender, geographical location, size of units, socio-economic status, education level, and occupational status.

The use of stratified sampling is important if it is possible that a particular subgroup could be excluded from a sample. Selecting a specified number from each strata ensures the representation of units across the entire population. Furthermore, an analyst may want to study subgroups of the population and compare results from both subsamples, e.g. males and females. By reducing sampling error, stratification offers greater precision in the estimates of underlying population parameters. Stratified samples may also be more convenient to administer, reducing the cost of implementing the survey (Lohr 2010, p. 73).

Stratification can occur at two stages: the design stage and the analysis stage. Stratification at the design stage is when analysts stratify when selecting the sample. In contrast, analysts may be presented with a SRS and wish to increase precision of populations of interest. Post-stratification

is the process of partitioning non-stratified data. While this method does increase precision, it is not as effective as stratification in the design stage.

To calculate the weights used in stratification, first divide the N population into H strata. Denote N_h as the number of sampling units in stratum h. The strata constitute the entire population, so we must have: $N_1 + N_2 + ... + N_H = N$. Survey designers then decide the number of units to be sampled from each strata. The total sample size is then: $n = n_1 + n_2 + ... + n_H$. If each n_h, $h = 1, ..., H$ is selected by SRS, the population total t_y can be estimated by:

$$\hat{t}_{y,str} = \sum_{h=1}^{H} \sum_{i \in S_h} \frac{N_h}{n_h} y_{hi} \tag{2.2}$$

where S_h is the set of n_h units in the SRS for stratum h and y_{hi} is the value of the ith unit in stratum h. The term N_h/n_h is the weight carried by the individual, as the probability of including unit i in stratum h is $\pi_{hi} = n_h/N_h$. If the size of strata differ and the number of units sampled from each strata differ, the probability of inclusion will differ across units from other strata. It is possible for the stratified sample to be self-weighting if the sampling fraction π_{hi} is the same for each stratum. In this case the sampling weights are the same under SRS; however, the bias and variance of $\hat{t}_{y,str}$ must still take the stratification into account (Lohr 2010, p. 78-79).

2.3 Clustering

In order to take a SRS or stratified sample from a population, all units must be known and clearly defined. In practice, creating a list of all available sampling units may be infeasible. A cluster sample is a probability sample in which the N population units are divided into several groups or clusters so that each cluster is representative of the entire population. The clusters are then sampled according to some sampling design. For example, one may want to sample punk rock fans in Ontario and be unable to identify all fans within the province. By identifying punk rock clubs, it would be possible to take a sample of these clubs and survey patrons to collect the desired data. In this example, the clubs are the clusters or primary sampling units (PSU) and patrons the observation units or secondary sampling units (SSU). While targeting the desired population,

patrons of the same club are likely to have more common characteristics than other clubs. From this example, two issues about using clustered sampling become clear: a cluster sample of size n may not be as informative as an SRS of size n and there may be correlation between observations within a cluster.

The widespread use of clustered sampling can be explained by it's ease of implementation. It saves time and other resources to focus on observation units within a group rather than sampling across all groups. This differs significantly from stratified samples; while both methods divide the population into groups, units from all strata are represented in stratified samples. Only units within sampled clusters are represented. Stratification helps to increase precision, cluster sampling tends to decrease it.

Two-stage cluster sampling is the process by which a sample of clusters is chosen, then observation units are sampled according to a sampling plan.[2] Consider a population that can be divided into C clusters, each with size N_s for $s = 1, ..., C$. If a sample of c clusters is taken in the first stage, with n_i units sampled from each of cluster $i = 1, ..., c$ in the second stage of sampling, the inclusion probabilities are calculated by:

$$P(j\text{th SSU in the } i\text{th PSU is selected}) = \frac{c}{C}\frac{n_i}{N_i}$$

2.4 Complex Surveys

Complex surveys use multistage sampling with the option of using different sampling methods at each stage. Each unit at each stage has an associated probability of being sampled. The probability of an object being sampled in the last stage is the product of probabilities from each stage of sampling. Consider a population divided into H strata in which a two-stage cluster design sample of c clusters and n_i SSUs is drawn from each stratum. The inclusion probability of individual j in cluster i from stratum h is:

$$\pi_{h|s|j} = \frac{N_h}{n_h}\frac{C}{c}\frac{N_i}{n_i}.$$

[2] One-stage sampling interviews all units within a sampled cluster

Note that with multi-stage sampling designs, there are multiple sampling criteria. Analysts base samples on multiple characteristics, which are themselves influenced by multiple variables.

In most econometric studies, the object in the final stage of sampling is usually a household or individual; sampling from a heterogeneous population of individuals may lead to another problem: non-response. Individuals who fail to respond or refuse to respond may share common characteristics which must be taken into account when calculating the overall probabilities of being sampled. When using data from complex surveys, one must be aware of the sampling plan and its criteria at every stage to ensure consistent estimation.

3 Parametric methods

Before examining the properties and performance of the proposed method, it would be beneficial to review previous estimators to gain knowledge as to when and why we would include survey weights. The reference point for most economists in estimating the relationship between X and y variables is ordinary least squares (OLS). OLS estimates the population parameter β in the equation $y_i = \mathbf{x}_i'\beta + u_i$ by $\hat{\beta} = (X^T X)^{-1} X^T y$.

When estimating descriptive statistics, the use of survey weights depends on the sampling scheme and whether the sample is representative of the population. If the sample perfectly represents the population, the population parameters can be consistently estimated using OLS (Magee et al. 1998). Example 2.1 showed that under SRS, $\hat{\beta}_{WLS}$ reduces to $\hat{\beta}_{OLS}$. As mentioned in the previous section, some sampling schemes are designed to obtain more precise information on subgroups that are of particular interest to analysts. By oversampling certain subpopulations, the sample presents a distorted view of the population as a whole and any estimation of descriptive statistics will be biased (Solon et al. 2013, Cameron & Trivedi 2009).

If one is trying to estimate causal effects between y and x, the decision to including sample weights is not as obvious. There may be times that it is preferable to ignore the sample design. Solon et al. (2013) described three settings where the inclusion of sample weights is necessary for consistent estimation: correction for heteroskedasticity, controlling for endogenous sampling and

identifying average partial effects. Here I focus on endogenous sampling.

It is helpful to define the indicator function as $I_i = \mathbf{1}(i \in S)$ where $I_i = 1$ if i is in the sample and $I_i = 0$ otherwise. The expectation of $E(I_i|\pi_i) = \pi_i$, $i \in U$. Assuming that $\pi_i > 0$ for all $i \in U$, the OLS estimator can then be written as:

$$\hat{\beta} = \left(\sum_{i \in S} x_i x_i^T\right)^{-1} \sum_{i \in S} x_i y_i = \left(\sum_{i \in U} I_i x_i x_i^T\right)^{-1} \sum_{i \in U} I_i x_i y_i. \tag{3.1}$$

OLS is only consistent if $E(E(I_i x_i u_i|\pi_i)|x) = E(\pi_i x_i u_i|x_i) = 0$. In a model with exogenous predictor variables, it must be that $E(x_i u_i|\pi_i) = E(x_i(y_i - x_i^T\beta|\pi_i)) = 0$. The very definition of endogenous sampling is that the sample was drawn based on values of y, i.e. $E(y_i|\pi_i) \neq 0 \Rightarrow E(x_i(y_i - x_i^T\beta|\pi_i)) \neq 0$. Therefore, under endogenous sampling, OLS estimates are inconsistent. The weighted least square estimator is given by $\hat{\beta}_{WLS} = \left(\sum_{i \in U} I_i w_i x_i x_i^T\right)^{-1} \sum_{i \in U} I_i w_i x_i y_i$, where the weights are equal to the inverse of the inclusion probability is consistent:

$$\hat{\beta}_{WLS} = \beta + \left(\sum_{i \in U} I_i w_i x_i x_i^T\right)^{-1} \sum_{i \in U} I_i w_i x_i u_i$$

and

$$\text{plim}(\hat{\beta}_{WLS}) - \beta = [E(I_i \pi_i^{-1} x_i x_i^T)]^{-1} E(I_i \pi_i x_i u_i) = [E(x_i x_i^T)]^{-1} E(x_i u_i) = 0$$

as long as $E(x_i x_i^T)$ is finite and nonsingular. When the sampling is exogenous, $E(\pi_i x_i u_i|x_i) = 0$ and OLS is consistent. In this case, WLS is still consistent; however, weighting variables may reduce efficiency (Greene 2012). It is at this point that we must be certain about our ability to control for endogenous sampling when using complex survey data. For multi-stage sampling, if the sample design is not based on the outcome variable, multiple sampling criteria from multiple sampling stages must be included as right-hand-side variables to ensure consistent estimation. Furthermore, misspecification because of omitted variables that were unavailable in the survey leads to inconsistent estimates. In this case, inclusion probabilities contain additional information for the estimation of the outcome variable (DuMouchel & Duncan 1983). It is often recommended to report both weighted and unweighted regression results. Next, I look at the properties of a

probability weighted nonparametric regression estimator.

4 Modified Nonparametric Regression Estimator

The purpose of this estimator is to estimate the relationship between the outcome variable and the auxiliary variables from a given sample by incorporating the sampling design. First, consider a finite population $U = \{1, ..., N\}$ of N units. For each $j \in U$ the outcome variable y_j and auxiliary variables $x_j = (x_j^d, x_j^c) = (x_{j1}^d, ..., x_{jr}^d, x_{j1}^c, ..., x_{jq}^c)$ are observed. x_j is a $(q + r) \times 1$ vector where the superscripts d and c denote that the variable is discrete or continuous, respectively. I use x_{jt}^c to denote the tth component of x_j^c and x_{jt}^d for the tth component of x_j^d and assume that x_j^d takes $c_t \geq 2$ different values in $\mathcal{D}_t = \{0, 1, ..., c_t - 1\}$, $t = 1, ..., r$. Next, a sample S of size n_s is drawn based on a complex sampling plan $p_N(\cdot)$, where $p_N(S)$ is the probability of drawing the sample S. The sampling rate is $Q = n_s/N$, with first order inclusion probabilities $\pi_j = Pr(j \in S) = \sum_{j \in S} p_N(S)$ and second order inclusion probabilities $\pi_{ji} = Pr(j, i \in S) = \sum_{j,i \in s} p_N(S)$. The variable n_s may be fixed (as in SRS) or random; however, I do not specify a sampling plan. The first and second order probabilities are the probabilities of obtaining the unit j and units j and i, respectively, while sampling from the population according to the complex sampling design. The model considered is a nonparametric regression model with additive errors of the form:

$$y_j = g(x_j) + u_j, \;\; j = 1, ..., n \tag{4.1}$$

where $E(u_j|x_j) = 0$ and $g(\cdot)$ is the unknown regression function and the object of interest. In this model, $g(\cdot)$ is the conditional expectation of y given x, i.e. $E(y|x) = g(x)$.

4.1 Local Constant Estimator

If data was available for every $i \in U$ then $g(x)$ in (4.1) could be estimated using the local-constant estimator. The local constant estimator was proposed by Nadaraya (1964) and Watson (1964) who wanted to estimate conditional mean functions as a locally weighted average, using a kernel

as a weighting function. The mathematical definition of $E(y|x)$ is:

$$E(y|x) = \int yf(y|x)dy = \int y\frac{f(y,x)}{f(x)}dy \tag{4.2}$$

where $f(y|x)$ is the conditional density of y given X, $f(y,x)$ is the joint density of y and x, and $f(x) = f(x^c, x^d)$ is the joint probability density function of (x^c, x^d). Nadaraya (1964) and Watson (1964) proposed substituting $f(y,x)$ and $f(x)$ by their kernel density estimates. For discrete regressors x_t^d, $t = 1, ..., r$, a variation on Aitchison and Aitken's (1976) kernel function can be used (scalar x) or embedded product kernel (multivariate x). This function is defined by

$$l(x_{it}^d, x_{jt}^d, \lambda) = \begin{cases} 1 & \text{if } x_{it}^d = x_{jt}^d \\ \lambda_t & \text{otherwise} \end{cases} \tag{4.3}$$

where $\lambda_t \in [0,1]$ is the smoothing parameter. When $\lambda_t = 0$ the above kernel function becomes an indicator function, and when $\lambda_t = 1$, it is a constant function and the (irrelevant) variable gets smoothed out. Here, a match between x_{it}^d and x_t^d determines the value of the discrete kernel function. The product kernel function for a vector of discrete variables is defined as

$$L(x_i^d, x^d, \lambda) = \prod_{t=1}^{r} \lambda_t^{1-\mathbf{1}(x_{it}^d = x_t)}.$$

Using k to denote a symmetric, univariate density function the product kernel for continuous variables is defined by:

$$W_h(x^c, x_i^c) = \prod_{t=1}^{q} \frac{1}{h_t} k\left(\frac{x_{ti}^c - x_i}{h_t}\right)$$

where $0 < h < \infty$ is the smoothing parameter. The shape of W depends on the choice of kernel function and the bandwidth and the distance between x_{ti}^c and x_i is the traditional Euclidean distance (Sánchez-Borrego, Opsomer, Rueda & Arcos 2014). A multivariate product kernel is given by

$$K_{h,ix} = W_h(x^c, x_i^c)L(x_i^d, x^d, \lambda).$$

The local constant estimator is then derived by substituting the kernel density estimators $\tilde{f}(x,y)$ and $\tilde{f}(x)$ for $f(x,y)$ and $f(x) = f(x^d, x^c)$, respectively, in equation (4.2):

$$\tilde{g}(x) = \int y \frac{\tilde{f}(y,x)}{\tilde{f}(x)} dy = \int y \frac{\frac{1}{h_y} \sum_{i \in U} k\left(\frac{y_i - y}{h_y}\right) K_{h,ix}}{\sum_{i \in U} K_{h,ix}} dy. \tag{4.4}$$

After analytic integration, with a bit of algebra the local constant estimators can then be written as:

$$\tilde{g}(x) = \frac{\sum_{i \in U} y_i K_{h,ix}}{\sum_{i \in U} K_{h,ix}}. \tag{4.5}$$

The benefit of using this method over parametric regression techniques is that it does not require the practitioner to specify the exact functional form of $E(y|x)$. Instead, $\tilde{g}(\cdot)$ is assumed to satisfy certain regularity conditions, including smoothness and moment conditions (Li & Racine 2007). Using the following assumptions, the asymptotic properties can be derived for the local constant.

Assumption 4.1. *Denote $\mathcal{S}$ as the compact support of x. Then, $g(x)$, $f(x)$, and $\sigma^2(x) = E(u_i^2|x_i)$ are second order differentiable in $\mathcal{S}$. Letting $A_s(x)$ and $A_{ss}(x)$ denote the first and second order derivatives of any function A w.r.t. x_s, then $\int g_{ss}(x)^2 f(x) > 0$ for all $s = 1, ..., q$.*

Assumption 4.2 (Kernel function)**.** *The kernel function $k(\cdot)$:$\mathbb{R} \to \mathbb{R}$ is symmetric with $k(v) \geq 0$ with $v \in \mathbb{R}$, and bounded by finite constant z so that $k(v) \leq z$. $k(\cdot)$ is m times differentiable with $\int k(v) v^4 dv < \infty$. $k(\cdot)$ is a second order kernel and define $\kappa_2 = \int v^2 k(v) dv$ and $\kappa = \int k^2(v) dv$.*

Assumption 4.3. *$(h_1, ..., h_q, \lambda_1, ..., \lambda_r) \in [0, \eta]^{q+r}$ lies in a shrinking set and $\eta = \eta_N$ is a positive sequence that converges to zero at a rate slower than the inverse of any polynomial in N. $N h_1 ... h_q \geq t_N$ with $t_N \to \infty$ as $N \to \infty$.*

Assumption 4.3 is a common assumption in the literature, it requires that $h_s \to 0$ for all s and $N h_1, ..., h_q \to \infty$ as $N \to \infty$. Comparing a kernel function to a smooth histogram, the bandwidth h is the width of the histogram bars. As the bars become thinner and thinner, we require that the bins remain non-empty to ensure a smooth function. If Assumptions 4.1 to 4.3 hold and further

assuming that (x_i, y_i) are i.i.d., then the asymptotic pointwise MSE of $\tilde{g}(x)$ is given by:

$$\begin{aligned} & MSE(\tilde{g}(x)) \\ &= \left(\frac{\kappa_2}{2} \sum_{s=1}^{q} h_s^2 [g_{ss}(x) + 2g_s(x) f_s(x)] + \sum_{s=1}^{r} \sum_{t^d \in \mathcal{D}} \mathbf{1}(t^d, x^d)[g(x^c, t^d) - g(x^c, gx^d)] f(x^c, t^d) \lambda_s \right)^2 \\ &+ \frac{\kappa^q \sigma^2(x) f^{-1}(x)}{N h_1 ... h_q} + O\left((N h_1 ... h_q)^{-1} \left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s \right) + \left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s \right)^2 \right) \end{aligned} \tag{4.6}$$

and $\mathbf{1}(t^d, x^d) = \mathbf{1}(t^d \neq x^d) \Pi_{s \neq s'}^{r} \mathbf{1}(t^d = x^d)$ (Li & Racine 2007). The first term on the right hand side of equation (4.14) is the square of the pointwise bias of $\tilde{g}(x)$ and the second term is the pointwise variance. Using Liapunov's central limit theorem, the asymptotic normal distribution of $\tilde{g}(x)$ is

$$\begin{aligned} \sqrt{N h_1 ... h_q} \left[\tilde{g}(x) - g(x) - \sum_{s=1}^{q} B_s(x) h_s^2 - \sum_{s=1}^{r} D_s(x) \lambda_s \right] \\ \xrightarrow{d} N(0, \kappa^q \sigma^2(x) / f(x)). \end{aligned} \tag{4.7}$$

4.2 Modified Local Constant Estimator

In the case of survey sampling, only the y_i in $S \in U$ are known. In this context, Sánchez-Borrego et al. (2014) proposed replacing the population totals from (4.5) by their Horvitz-Thomspon estimators:

$$\hat{g}(x) = \frac{\sum_{i \in \mathcal{S}} \pi_i^{-1} y_i K_{h.ix}}{\sum_{i \in \mathcal{S}} \pi_i^{-1} K_{h.ix}} \tag{4.8}$$

Under SRS, the proposed estimator by Sánchez-Boreggo et al. (2014) becomes the traditional nonparametric estimator from (4.5). Before deriving the asymptotic properties of the estimator, the following assumption is needed for defining the sampling design.

Assumption 4.4 (Sample Design)**.** *The sampling plan $p_N(S)$ is such that as $i \to \infty$, the sampling rate $n_{s,i}/N_i$ converges with probability one to a finite constant $1 \geq Q > 0$. It is further assumed the design expectation of n_s is $E_P(n_s) = n$ For all N, the first order inclusion probabilities are such that for all N, $\min_{j \in U} \pi_j \geq \epsilon > 0$, with probability one. The second order inclusion probabilities*

satisfy $\min_{i,j \in U} \pi_{ij} \geq \epsilon^* > 0$ *and*

$$\lim_{i \to \infty} \sup n_{s,i} \max_{j, i \in U : j \neq i} |\pi_{ij} - \pi_i \pi_j| \leq \infty$$

with probability one.

4.2.1 Asymptotic Properties

As noted above, there are three methods of inference when deriving the asymptotic properties of estimators in survey sampling: model-based inference, pure design-based inference, and combined inference. Model-based inference assumes the data (y, x) are generated based on a given model and that the inclusion probabilities are uninformative. Using this mode of inference relies heavily on model specification as it is assumed that the model represents all units of the population. Naturally, this increases the appeal of nonparametric methods, such as the local constant described above, which makes no assumptions about the functional form of the model (other than smoothness and existence). However, if the sampling is endogenous, model based estimators will not take this into account and results will be inconsistent. If one takes a model-based approach, then the estimator in (4.5) is appropriate.

In pure design-based settings, inference depends on the probability distribution induced by the sampling design and not the probability distribution from an underlying model. Inferences drawn using a design based approach typically refer to a particular finite population of interest and usually ignore any model structure in the corresponding superpopulation. Expectations are taken with respect to the sampling scheme; therefore, asymptotic results depend on the sample size, the sampling design, and the bandwidths h and λ (Buskirk & Lohr 2005). Sánchez-Borrego et al. (2014) adopted a design-based setting for deriving the asymptotic properties for their modified local constant estimator (4.8). Under the assumption of i.i.d. data the authors show that the estimator is asymptotically design unbiased and design consistent with probability one.

The method of inference considered in this paper is the combined framework outlined in Pfeffermann (1993) and adopted by Harms and Duchesne (2010), and Buskirk and Lohr (2005). Using

this approach, superpopulation parameters are estimated using design-consistent estimators of finite population parameters. These quantities are then consistent estimators of the superpopulation parameters under the proposed model (Buskirk & Lohr 2005). This mode of inference follows two steps: first a finite population U is generated according to a superpopulation model, denoted as ξ, where elements in the finite population are presumed to be realizations of random variables with a joint probability distribution. As before, for each j in the population, the realization (x_j, y_j) is obtained, such that (x, y) follows the joint density $f(x, y)$. For the analysis that follows, it is assumed that the x_j's, $j \in U$, are i.i.d.. This is a popular assumption when working in the combined inference framework (Bellhouse & Stafford 1999). This represents the model stage. Later, I relax this assumption to account for clustered data. The main objective is to estimate the unknown object $g(x)$ specified by (4.1) between the predictor variables and the response. As before, to estimate $g(\cdot)$, a sample S of size n_s is drawn according to a complex sampling plan. For what follows, model-based inference, design-based inference, and combined inference are denoted by ξ, P, and C, respectively. The conditional mathematical expectation under the combined mode of inference is calculated as $E_C = E_\xi\{E_P\{\cdot|\pi\}|x\}$ where $x = \{x^c, x^d\}$. The following example shows how the combined framework is used to derive the asymptotic bias of the modified multivariate kernel density estimator $\hat{f}(x) = \sum_{i\in S} \pi^{-1} K_{h,ix}$.

Example 4.1 (Kernel Density Estimation under combined inference)**.** *If Assumptions* $4.1 - 4.4$ *are satisfied, the asymptotic bias of the modified kernel density estimator under the combined inference is equivalent to the bias of model-based estimator* $\tilde{f}(x)$*. Recall, the indicator function* $\mathbf{1}(i \in S)$*, which equals one if unit* i *is in sample* S*, zero otherwise, and that* $E_P(\mathbf{1}(i \in S)|\pi) = \pi_i$*. The modified density estimator can be written as:*

$$\hat{f}(x) = \sum_{i\in S} \pi^{-1} K_{h,ix} = N^{-1} \sum_{i=1}^{N} \pi^{-1} \mathbf{1}(i \in S) K_{h,ix} \tag{4.9}$$

Next, take the expectation of (4.9) using the combined inference method.

$$
\begin{aligned}
E_C(\hat{f}(x)|x) &= E_\xi\{E_P(\hat{f}(x))|\pi)|x\} = E_\xi\left\{E_P\left(\sum_{i\in S}\pi^{-1}K_{h,ix}\right)|x\right\} \\
&= E_\xi\left\{E_P\left(N^{-1}\sum_{i=1}^{N}\pi^{-1}\mathbf{1}(i\in\mathcal{S})K_{h,ix}\right)|x\right\} \\
&= E_\xi\left\{N^{-1}\sum_{i=1}^{N}\pi^{-1}E(\mathbf{1}(i\in\mathcal{S}))K_{h,ix}|x\right\} \\
&= E_\xi\left(N^{-1}\sum_{i=1}^{N}\pi_i^{-1}\pi_i K_{h,ix}\right) \\
&= E_\xi\left(N^{-1}\sum_{i=1}^{N}K_{h,ix}\right) \\
&= \sum_{t^d\in\mathcal{D}}\int_{\mathbb{R}^q}\prod_{s=1}^{q}h_s^{-1}w\left(\frac{t_s-x_s}{h_s}\right)\prod_{s=1}^{r}\lambda_s^{\mathbf{1}(t_s\neq x_s)}f(t^c,t^d)dt^c \\
&= \int_{\mathbb{R}^q}\prod_{s=1}^{q}w(v_s)f(x^c+hv_s,x^d)dv_s + \sum_{t^d\in\mathcal{D}}\mathbf{1}(t^d,x^d)\lambda_s\int_{\mathbb{R}^q}\prod_{s=1}^{q}w(v_s)f(x^c+hv_s,t^d)dv_s \\
&= \frac{\kappa_2}{2}\sum_{s=1}^{q}h_s^2 f_{ss}(x) + \sum_{s=1}^{r}\sum_{t^d\in\mathcal{D}}\mathbf{1}(t^d,x^d)f(x^c,t^d)\lambda_s + O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right) \qquad (4.10)
\end{aligned}
$$

Now consider the asymptotic properties of the proposed estimator $\hat{g}(x)$. Following Li and Racine (2007), I examine the numerator and denominator of $\hat{g}$(x) separately. First write:

$$\hat{g}(x) - g(x) = \frac{\hat{m}(x)}{\hat{f}(x)} \qquad (4.11)$$

where $\hat{m}(x) = (\hat{g}(x) - g(x))\hat{f}(x)$. Using the equation for the regression model with additive errors (4.1), $\hat{m}(x)$ can be written as:

$$
\begin{aligned}
\hat{m}(x) &= N^{-1}\sum_{i=1}^{N}\pi^{-1}\mathbf{1}(i\in S)[g(x_i)-g(x)]K_{h,ix} + N^{-1}\sum_{i=1}^{N}\pi^{-1}\mathbf{1}(i\in S)u_i K_{h,ix} \\
&= \hat{m}_1(x) + \hat{m}_2(x)
\end{aligned}
$$

where the definition of $\hat{m}_1(x)$ and $\hat{m}_2(x)$ should be evident. In Appendix **A**, I show that the leading

term of the expectation of $\hat{m}_1(x)$ under the combined framework is:

$$E_C[\hat{m}(x)|x] = \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2[g_{ss}(x)f(x) + 2g_s(x)f_s(x)] + \sum_{s=1}^{r}\sum_{t^d\in\mathcal{D}} \mathbf{1}(t^d, x^d)[g(x^c, t^d) - g(x^c, gx^d)]f(x^c, t^d)\lambda_s + O\left(\sum_{s=1}^{q} h_s + \sum_{s=1}^{r}\lambda_s\right) \tag{4.12}$$

Since $\hat{g}(x) - g(x) = \hat{m}(x)/\hat{f}(x) = \hat{m}(x)/(f(x) + o_p(1))$, the bias of $\hat{g}(x)$ is equivalent to the bias under the model-based estimator $\tilde{g}(x)$. This result is not surprising, Harms and Duchesne (2010) showed the bias of the model-assisted local linear in the scalar continuous variable case was equal to the bias of the traditional model-based local linear estimator with one continuous predictor variable. Also, in Appendix **A**, I show the variance of the model assisted estimator under the combined framework is equal to the variance under the model framework multiplied by a correction factor:

$$\begin{aligned}\text{var}_C\{[(\hat{g}(x) - g(x))|\pi]|x\} =& \frac{1}{nh_1...h_q}\left\{N^{-2}n\sum_{i=1}^{N}(w_i - 1) + \frac{n}{N}\right\}\frac{\kappa^q\sigma(x)}{f(x)} \\ &+ O\left[(Nh_1...h_q)^{-1}\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r}\lambda_s\right)\right]\end{aligned} \tag{4.13}$$

Note that under SRS, the correction factor equals one and equation (4.13) reduces to the variance of $\tilde{g}(x)$ evaluated over the sample data. Combining these two results proves Theorem 4.1.

Theorem 4.1. *If Assumptions* 4.1-4.4 *are satisfied, then the conditional pointwise MSE of the model-assisted local constant estimator* $\hat{g}(x)$ *under the combined inference mode is given by:*

$$\begin{aligned}&MSE(\hat{g}(x)) \\ &= \left[f^{-1}(x)\left(\frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2[g_{ss}(x) + 2g_s(x)f_s(x)] + \sum_{s=1}^{r}\sum_{t^d\in\mathcal{D}} \mathbf{1}(t^d, x^d)[g(x^c, t^d) - g(x^c, gx^d)]f(x^c, t^d)\lambda_s\right)\right]^2 \\ &+ (\Delta + Q)\frac{\kappa^q\sigma^2(x)f^{-1}(x)}{Nh_1...h_q} + o_p\left((Nh_1...h_q)^{-1}\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r}\lambda_s\right) + \left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r}\lambda_s\right)^2\right)\end{aligned} \tag{4.14}$$

where $\Delta = N^{-2}n\sum_{i=1}^{N}(w_i - 1)$ *and* $Q = n/N$.

The following theorem is proved in Appendix **A.2** and describes the asymptotic normality of $\hat{g}(x)$. The proof makes use of the fact that $\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s = o(1)$ and $(Nh_1...h_q)^{-1}$=o(1).

Theorem 4.2. *Under the assumption that* x *is an interior point and Assumptions* 4.1-4.4 *are satisfied, the asymptotic normality of* $\hat{g}(x)$ *is defined by:*

$$\sqrt{Nh_1...h_q}\left(\hat{g}(x) - g(x) - \sum_{s=1}^{q} h_s^2 - \sum_{s=1}^{q} \lambda_s\right) \xrightarrow{d} N(0, (\Delta + Q)\kappa^q \sigma^2(x)/f(x)) \tag{4.15}$$

4.2.2 Dependent data

Up to this point, it has been assumed that the data in the population U is i.i.d. from which a sample is drawn based on a complex sampling plan. While correlation of data is a well-known problem in time series data, it is also present in cross-sectional survey data, where it is often be ignored. As pointed out in Buskirk and Lohr (2005) and Breunig (2001), the i.i.d. assumption causes difficulty for estimation using survey data because it implies there is no clustering effects in the superpopulation. In survey data, it is reasonable to assume that individuals within the same cluster would be considered dependent in the superpopulation.

Assume now that the population $U = \{1, ..., N\}$ can be divided into C clusters. Denote N_c as the size of cluster c with $c = 1, ..., C$, so that $N = N_1 + ... + N_C$. Assume that each individual $i \in c$ lies on a straight line where individuals i and $i + \tau$ are positioned such that $\tau - 1$ individuals lie between them. For what follows, I assume that there is weak dependence between observations within the same cluster. Weak dependence means that the dependence between two observations x_{ci} and $x_{c'j}$ $(c = c')$ goes to zero as the number of observations between i and j goes to infinity, i.e. if $j = i + \tau$ then $cov(x_i, x_j) \to 0$ as $\tau \to \infty$. To model weak dependence between observations with clusters, I make use of the ρ mixing process outlined in Definition 4.1.

Definition 4.1 (ρ-mixing)**.** *Let* $\mathcal{M}_i^{i+\tau}$ *be a subset of* $\{x_s\}_{s=j}^{j+\tau}$ *that is closed under complementation and under countable union operations so that* $\mathcal{M}_j^{j+\tau}$ *is a* σ*-field of* $\{x_s\}_{s=j}^{j+\tau}$*. Then, the sequence*

$\{x_i\}_{i=-\infty}^{\infty}$ is said to be ρ-mixing if the coefficient $\rho_\tau \to 0$ as $\tau \to \infty$, with

$$\rho_\tau = \rho(\tau) = \sup_{i\in\mathbb{N}} \sup_{A\in\mathcal{M}_{i+\tau}^{\infty}, B\in\mathcal{M}_{-\infty}^{i}} \left| \frac{\text{cov}(A,B)}{\sqrt{\text{var}(A)\text{var}(B)}} \right|$$

Note, if the observations were independent of each other, $\rho_\tau = 0$ for $\tau \geq 1$.

In addition to Assumptions 4.1-4.4, I need to make the following assumption about the cluster sizes so that the estimator remains consistent.

Assumption 4.5. *(i)There are ρ-mixing observations $x_{c1}, ..., x_{cN_c}$ for all $c = 1, ..., C$, with $\rho_\tau = 0$ if $c \neq c'$. The ρ-mixing process satisfies $\rho(\tau) = O(\tau^{-(1+\epsilon)})$ for some (small) $\epsilon > 0$. (ii) It is also assumed that $\min N_c \to \infty$ as $N \to \infty$.*

Part (i) of Assumption 4.5 implies that $\sum_{\tau=1}^{\infty} \rho(\tau) < \infty$ and part (ii) ensures that the size of the smallest cluster grows as the population goes to infinity. The following theorem shows that the MSE convergence rate of $\hat{g}(x)$ with weakly dependent data in clusters is the same as the i.i.d. case under the combined framework. The proof in Appendix **A.3** shows that the leading terms for the bias and the variance with weakly dependent data are the same as for the i.i.d. case.

Theorem 4.3. *If Assumptions $4.1 - 4.5$ are satisfied, the order of the pointwise MSE of the estimator $\hat{g}(x)$ is*

$$E\left\{[\hat{g}(x) - g(x)]^2\right\} = o_p\left((Nh_1...h_q)^{-1}\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r}\lambda_s\right) + \left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r}\lambda_s\right)^2\right). \tag{4.16}$$

5 Bandwidth Selection

A critical component to any nonparametric regression technique is the choice of the smoothing parameters (h, λ). Selecting the smoothing parameters for the q continous variables creates a trade-off between the bias and variance of the estimator. Large values of h_s will oversmooth the underlying density and increase the bias while reducing the variance. Conversly, small h_s will undersmooth the underlying density shrinking the bias but increasing the variability of the estimator. For the

univariate continuous variable case, Harms and Duchesne (2010) used a modified plug-in method for selecting the bandwidth according to the MSE criterion in the combined inference mode. The optimal bandwidth in that case was equal to that of the i.i.d. case multiplied by a correction factor equal to $(\Delta + Q)$. This method is not applicable to the the multivariate case and therefore, not applicable for the present model. In their simulations, Sánchez-Borrego et al. (2014) used a plug-in method for the bandwidth in which they selected three values for h and five values for λ. In addition they used survey cross-validation to choose among the fifteen possible combinations of the fixed values for h and λ.

It is widely accepted that data-driven methods for selecting the bandwidths in a nonparametric kernel regression setting is required for proper inference and analysis. I propose using least-squares cross-validation (LSCV), a fully automatic data-driven method, for selecting (h, λ)= $(h_1, ..., h_q, \lambda_1, ..., \lambda_r)$. LSCV chooses (h, λ) to minimize the following cross-validation function:

$$CV(h, \lambda) = \sum_{i \in S} (y_i - \hat{g}_{-i}(x_i))^2 M(x_i) \tag{5.1}$$

where $g_{-i}(x_i) = \sum_{j \in S, j \neq i} \pi_j^{-1} y_j K_h, ij / (\sum_{j \in S, j \neq i} \pi_j^{-1} K_h, ij)$ is the leave-one-out kernel estimator of $g(x_i)$ and $K_{ij} = \Pi_{j \neq i}^{q} k((x_{is}^c - x_{js}^c)/h_s) \Pi_{s=1}^{r} \lambda_s^{\alpha}$ with $\alpha = \mathbf{1}(x_{is} \neq x_{js})$ is equal to 1 if $x_{is} \neq x_{js}$ and zero otherwise. $0 < M(x_i) < 1$ is a weight function which serves to avoid difficulties caused by dividing by zero. Using the leave-one-out kernel estimator helps to avoid a computational issue encountered when optimizing according to $\sum_{i \in S} (\hat{u}_i)^2 = \sum_{i \in S} (y_i - \hat{g}(x_i))^2$. By letting $h_s \to 0$, $\hat{g}(x_i)$ can be made close to y_i for any $i \in S$. Hence, $\sum_{i \in S} (\hat{u}_i)^2$ can be made very small as $h \to 0$. At the same time, MSE_C remains greater than zero for all values of h (Opsomer & Miller 2005). By replacing $\hat{g}(x_i)$ with the delete-one estimator, the difference $y_i - \hat{g}_{-i}(x_i)$ does not go to zero as $h \to 0$. The proceeding analysis requires the following assumption about the weight function $M(x_i)$:

Assumption 5.1. *$M(\cdot)$ is continuous, nonnegative and has a compact support $\mathcal{S}$.*

In Appendix **A.4**, I show that, if we ignore the the terms unrelated to (h,λ), the leading term

of $CV(h,\lambda)$ is given by $E[CV_0(h,\lambda)]$:

$$E[CV_0(h,\lambda)] = \sum_{x^d\in\mathcal{D}} \int \left(\left\{ \sum_{s=1}^{q} B_s(x) h_s^2 + \sum_{s=1}^{r} D_s \lambda_s \right\}^2 f(x) + \sum_{j\neq i} \pi_j^{-1} \frac{\kappa^q \sigma(x)}{N^2 h_1 ... h_q} \right) M(x) dx$$

$$= \sum_{x^d\in\mathcal{D}} \int \left(\left\{ \sum_{s=1}^{q} B_s(x) h_s^2 + \sum_{s=1}^{r} D_s \lambda_s \right\}^2 f(x) + (\Delta_{-i} + Q) \frac{\kappa^q \sigma(x)}{n h_1 ... h_q} \right) M(x) dx \quad (5.2)$$

with $\Delta_{-i} = n/N^2 \sum_{j\neq i}(w_j - 1)$. Therefore, the leading term for least squares cross-validation is the same for an i.i.d. sample except for the correction term on the second term on the right-hand side of Equation (5.2). Define $a_1, ..., a_q, b_1, ..., b_r$ as $h_s = N^{1/(4+q)} a_s$ $(s = 1, ..., q)$ and $\lambda_s = N^{2/(4+q)} b_s$ $(s = 1, ..., r)$. Then we obtain $E[CV_0(h,\lambda)] = \chi_r(a,b)$ with

$$\chi_r = \sum_{x^d\in\mathcal{D}} \int \left(\left\{ \sum_{s=1}^{q} B_s(x) a_s^2 + \sum_{s=1}^{r} D_s b_s \right\}^2 f(x) + \sum_{j\neq i} \pi_j^{-1} \frac{\kappa^q \sigma(x)}{a_1 ... a_q} \right) M(x) dx. \quad (5.3)$$

6 Simulations

In this section, I estimate the performance of the modified nonparametric regression estimator $\hat{g}$ and compare it against that of the traditional Nadaraya-Watson estimator $\tilde{g}$, which ignores sampling weights, under different population DGPs. The bandwidths are computed using the least squares cross-validation method described in Section 5. I wish to assess the finite-sample properties of each estimator; performance is measured based on MSE defined as:

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{g}(x_i) - g(x_i))^2.$$

For each Monte Carlo replication, I generate three predictor variables: $x = \{x_1^c, x_1^d, x_2^d\}$. x_1^c is a uniform variable with support within the interval [0,1], and x_1^d and x_2^d are independent binary factor variables. The populations are then generated using the following regression model:

$$y = g(x) + u, \; u \sim N(0, \sigma^2). \quad (6.1)$$

I consider four DGPs for $g(x)$, which are outlined in Table 1 in Appendix **B**. The first population, $g_1(X)$, was considered in Sánchez-Borrego et al. (2014) and is simply linear in the continuous variable, x_1^c. In population two, the relationship between y and x_1^c is quadratic introducing a further degree of smoothness. Population three, also known as *bump*, was considered by Harms and Duchesne (2009) and Sánchez-Borrego et al. (2014). This function produces a noticeable bump at $x_1^c = 0.5$. Population four is the most complex function considered. The *Härdle* function is characterized by a peak at $x_1^c = 0.63$, a valley at $x_1^c = 0.91$, and a saddle point at $x_1^c = 0.79$. The relationship between all DGPs and x_1^c are displayed in Figure 1 in Appendix **C**. The error term for each population is assumed to be normally distributed with mean of zero and standard deviation σ.

Each run of the Monte Carlo generates a population of size $N = 10000$, then draws a sample of size n based on three different sampling methods: SRS without replacement, stratification based on x_1^c, and stratification based on y_i, $i = 1, 2, 3, 4$. The number of replications is 800. For stratified samples, the population was divided into three strata of varying size, with unequally sized samples being drawn from each strata. Within-strata sampling was performed by SRS without replacement. Table 2 in Appendix **B** displays the strata borders and the sample size drawn from each strata. Sensitivity analysis is performed by varying the sample size ($n = 200, 400, 800$) and standard deviation of the error term in the regression model ($\sigma = 0.25, 0.50, 1.00, 2.00$).

Both the local constant and modified local constant estimators are computed using the Gaussian kernel for the continuous variable x_1^c and the variant of Aitchison and Aiken (1976) kernel in (4.3) for discrete variables x_1^d and x_2^d. Note, I am only considering the relevant data case: I expect the bandwidths $\lambda_r < 1$, $r = 1, 2$ for x_1^d and x_2^d. I let h denote the bandwidth for x_1^c.

Tables **??-??** in Appendix **B** present the results from the Monte Carlo simulation for all DGPs and combinations of n and σ. In each table, columns four and five report the MSEs for $\hat{g}(x)$ and $\tilde{g}(x)$, denoted by MSE_W and MSE_U, respectively. The values in brackets below the reported MSEs are the *median absolute deviations* (MAD) of the MSEs; a robust measure of the variability of the MSE (Andersen 2008). The MAD is calculated by $\text{median}(|MSE_q^{(m)} - \text{median}(MSE_q)|)$ with $m = 1, ..., 800$ and $q = W, U$. Compared to the standard deviation, the MAD is more resilient to

outliers. The superscripts W and U in columns six to eleven denote values from the probability weighted estimator and unweighted estimator, respectively. All values presented are the median values of their respective measure.

Table **??** displays the results from SRS. Not surprisingly, the MSE for both estimators is equal up to the fourth decimal for all combinations of n and σ. The differences are simply due to bootstrap sampling error. In this case, the inclusion probabilities are equal for all individuals and $\hat{g}(x)$ reduces to $\tilde{g}(x)$. As Harms and Duchesne (2010) point out in their simulations, the results from SRS act as a benchmark for other sampling plans. Keeping n constant, as σ increases, so too does the MSE of both estimators. In order for $\hat{g}(x)$ to be a consistent estimator, the MSE needs to decrease as the sample size increases. Keeping σ constant and increasing n, the MSE decreases for all DGPs. This provides evidence that the estimator is consistent. As the functions increase in the degree of complexity, the MSE also increases. The bandwidths selected for both models are equal for the three predictor variables. As the DGPs are all functions of x_1^d and x_2^d, the median bandwidths λ_1 and λ_2 are less than 1. Looking at the MADs for both the weighted and unweighted nonparametric estimator, they too are equal for all combinations of n and σ. This is reflected in Figure 2, which shows the boxplots for the MSEs of the estimators $\hat{g}(x)$ and $\tilde{g}(x)$. The values labelled "WLC" and "LC" are the median MSE values for $\hat{g}(x)$, the probability weighted local constant, and $\tilde{g}(x)$, the traditional local constant, respectively. The plots show that the distribution of MSEs are identical, with an increasing MSE as the DGP increases in complexity.[3] I included boxplots from OLS and WLS denoted by OLS and WLS, respectively. While the MSE for OLS and WLS are smaller for the linear DGP, they are much higher and more variable for the quadratic, bump, and Härdle DGP.

The results from stratification on the outcome variable are presented in Table 4. Here, weighting by inclusion probabilities clearly shows an improvement as the median MSE is smaller for $\hat{g}(x)$ than $\tilde{g}(x)$. By not accounting for endogenous sampling and unequal inclusion probabilities, the traditional local constant performs worse than the weighted local constant. Again, increasing the level of noise in the model reduces the efficiency of each estimator. $\hat{g}(x)$ remains consistent as the MSE decreases as n increases. The MADs suggest that MSE($\hat{g}(x)$) is more variable than

[3]Boxplots for other combinations of n and σ showed similar results.

MSE($\tilde{g}(x)$) for small sample sizes. When $n = 200$, $MAD(MSE_W)$ is greater than $MAD(MSE_U)$; this is true for all DGPs. As the sample size increases, however, MSE($\hat{g}(x)$) becomes less variable than MSE($\tilde{g}(x)$). This can be seen in Figure 3 in Appendix **C**. Figures 3a, 3c, 3e, and 3g display the boxplots of the MSE of $\hat{g}(x)$ and $\tilde{g}(x)$ for the linear, quadratic, bump, and Härdle DGPs, respectively, with $n = 200$ and $\sigma = 0.50$.[4] These plots show there is greater variability in the MSE for $\hat{g}(x)$. For linear, quadratic, and Härdle models, the first three quantiles of MSE($\hat{g}(x)$) lie below the median of MSE($\tilde{g}(x)$). For the bump DGP in Figure 3e, the median of MSE($\hat{g}(x)$) is below the first quantile of of MSE($\tilde{g}(x)$). Figures 3b, 3d, 3f, and 3h show the boxplots of the MSE of $\hat{g}(x)$ and $\tilde{g}(x)$ for linear, quadratic, bump, and Härdle DGPs, respectively, with $n = 800$ and $\sigma = 0.50$. In these figures, MSE($\hat{g}(x)$) clearly shows a stochastic dominance over MSE($\tilde{g}(x)$). There is less variablility in MSE($\hat{g}(x)$) and the maximum value lies below the minimum of MSE($\tilde{g}(x)$) for all DGPs. Again, looking at Figure 4 in Appendix **C**, WLS regression performs best under linear DGP. However, as the DGP becomes more complex, $\hat{g}(x)$ outperforms both parametric estimators.

Looking now at the bandwidths in columns six to eleven in Table 4, the least squares cross-validation estimates for $\hat{g}$ are smaller than those for $\tilde{g}(x)$. Again, since x_1^d and x_2^d are both relevant variables, the median values for their bandwidths lie between zero and one. Note, that for linear and quadratic models with $n = 400$, median values for λ_1^W and λ_2^W are close to zero, meaning the function $l(x_i^d, x^d, \lambda)$ is approximately an indicator function.

Table **??** displays the results from stratification on the continous predictor variable x_1^c. Since x_1^c is included in the model, the sampling scheme is exogenous and accounted for. The MSE results differ by functional form, with the MSEs from more complex functions being higher for $\hat{g}(x)$ than $\tilde{g}(x)$. For the linear DGP, $g_1(x)$, $\hat{g}_1(x)$ and $\tilde{g}_1(x)$ are equal when $n = 200$ and $\sigma = 0.25$; median MSE($\hat{g}_1(x)$) is the same as median of MSE($\tilde{g}_1(x)$) and MADs are equal. This can be seen in Figure 5a in Appendix **C**. As σ increases, both MSE($\hat{g}_1(x)$) and MSE($\tilde{g}_1(x)$) become more variable, with MSE($\hat{g}_1(x)$) more variable than MSE($\tilde{g}_1(x)$) (Figure 5b). As n increases, MSE($\hat{g}_1(x)$) and MSE($\tilde{g}_1(x)$) continue to be equal, however, MSE($\hat{g}_1(x)$) becomes less variable than MSE($\tilde{g}_1(x)$) (Figures 5c-5f). For the quadratic DGP, $g_2(x)$, the median MSE($\hat{g}_2(x)$) is approximately equal to

[4] Results were similar for different values of σ

the median MSE($\tilde{g}_2(x)$) for all n. For $\sigma = 0.25, 0.50$, the MADs for MSE($\hat{g}_2(x)$) and MSE($\tilde{g}_2(x)$) are equal. As σ increases, MSE($\hat{g}_2(x)$) becomes more variable than MSE($\tilde{g}_2(x)$) for all n (see Figure 6in Appendix **C**). For bump and Härdle DGPs, both the median and the MAD are equal for MSE($\hat{g}_i(x)$) and MSE($\tilde{g}_i(x)$) for all n and $\sigma = 0.25, 0.50$, $i = 3, 4$. As σ increases, both the median and the variability of MSE($\hat{g}_i(x)$) are greater than for that for MSE($\tilde{g}_i(x)$), $i = 3, 4$. This suggests that the probability weighted nonparametric regression estimator performs worse than traditional nonparametric regression estimators for complex functional forms when sampling is exogenous.

This Monte Carlo simulation has provided evidence that the need to include survey weights depends on the sampling design. Under SRS, the weighted estimator reduces to the traditional nonparametric regression estimator and results are equivalent regardless of underlying DGP. Comparing across sampling schemes, $\hat{g}(x)$ is most efficient under SRS and least efficient under endogenous stratification. However, when endogenous stratification is present in the model, efficiency gains can be realized by including sample weights equal to the inverse of the inclusion probability relative to $\tilde{g}(x)$. If one is able to control for endogenous sampling by including the sampling criterion as a predictor variable, it may reduce efficiency to include survey weights in the model. By varying the sample size, I was able to provide evidence that the estimator $\hat{g}(x)$ is a consistent estimator as the MSE decreased as the sample increased. Furthermore, if the relationship between y and x is non-linearnear, linear parametric models will be inconsistent regardless of weighting.

7 Application

The application considered here is an extension of the example in Harms and Duchesne (2010). The authors used data from the 2000 cycle of the Survey of Labour and Income Dynamics (SLID) to estimate the relationship between age and labour market duration (LMD) for 38,941 individuals. The purpose of the SLID is to understand the economic well-being of Canadians, collecting data on the primary source of income, education, and demographic backgrounds of its participants. The sampling scheme is based on a stratified, multi-stage design that uses probability sampling. The result is unequal sampling weights for individuals in the sample. The weights not only represent the

sampling plan but also account for nonresponse and are calibrated for to meet certain benchmark criteria.

The application presented in this paper differs in two ways from Harms and Duschesne (2010). First, in order to help reduce heteroskedasticity in the model, the outcome variable I consider is LMD as a percent of age over 18. Figure 9a displays the relationship between LMD, measured in months, and Age, measured in years. To improve readability, only 200 points are plotted. The size of each point is determined by the weight for each observation, with larger points representing a larger weight. Looking only at LMD versus age, the variance of LMD for older individuals is likely to be higher compared to younger individuals. Second, I extend the model to include a discrete variable Gender, which takes on two values, male or female.

I estimate the model using both the probability weighted local constant estimator and the traditional local constant estimator. In both cases, the Gaussian kernel was used for the continuous variable Age and the variation of the Aitchison and Aiken (1974) kernel was used for the discrete variable Gender. The bandwidths for the weighted local constant were computed using the cross validation method described in section 5. The bandwidths for the traditional local constant were computed using method outlined in Li and Racine (2007, Chaper 2). The solid blue and and green lines in Figure 9b represent the weighted regression results for males and females respectively. It is clear that these two curves are pulled closer to observations which represent a greater number of individuals in the population compared to the unweighted estimates (the dashed lines in Figure 9b). Results also show that females spend a smaller percentage of time in the labour force compared to men as the black and green lines lie below the red and blue lines. Table 6 shows the bandwidths for both estimators; the bandwidths for gender are under 1 for both estimators, indicating it is a significant predictor of LMD.

8 Conclusion

This chapter took an extensive look at nonparametric regression estimation for the multivariate mixed data types using complex survey data. The purpose of this overview was to derive the

asymptotic properties of the probability weighted nonparametric regression estimator under the combined inference framework. Populations were studied under both i.i.d. and weakly dependent frameworks. For both settings, the bias of the modified nonparametric regression estimator had the same leading terms and order of probability as under the model based framework.

A secondary purpose of this paper was to assess settings under which incorporating sampling weights is appropriate. Using a data-driven method for selecting the bandwidth, I showed that under SRS, the MSE of the traditional and modified local constant estimators were equal. For exogenous stratification, boxplots of the MSE of both estimators showed that for less complex functions, the modified nonparametric regression estimator held a slight advantage over or performed equally well as the traditional nonparametric regression estimator. For more complex functions, the unweighted local constant showed less variability and smaller mean and median MSE. When sampling is endogenous, efficiency gains can be made by including survey weights. If one is sure it is possible to control for endogenous sampling by including the sampling criterion as a predictor variable, it may be best to use model-based estimators. The issue is that complex surveys use multi-stage sampling and contain only limited auxiliary variables. Therefore, it may not be possible to include all sampling criteria and produce a model that applies to all members of the population. To be safe, it is recommended that both weighted and unweighted results be reported.

References

Aitchison, J. & Aitken, C. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrica* **63**, 413–420.

Andersen, R. (2008), *Modern Methods for Robust Regression*, SAGE Publications, Inc.

Bellhouse, D. & Stafford, J. (1999), 'Density estimation from complex surveys', *Statistica Sinica* **9**, 407–424.

Bellhouse, D. & Stafford, J. (2001), 'Local polynomial regression estimators in complex surveys', *Survey Methods* **27**, 197–203.

Breidt, F. & Opsomer, J. (2000), 'Local polynomial regression estimators in survey sampling', *The annals of Statistics* (28), 1026–1053.

Breunig, R. V. (2001), 'Density estimation for clustered data', *Econometric Reviews* **20**(3), 353–367.

Breunig, R. V. (2008), 'Nonparametric density estimation for stratified samples', *Statistics and Probability Letters* **78**, 2194–2200.

Buskirk, T. D. & Lohr, S. L. (2005), 'Asymptotic properties of kernel density estimation with complex survey data', *Journal of Statistical Planning and Inference* (128), 165.

Cameron, C. & Trivedi, P. (2009), *Microeconometrics: Methods and applications*, Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA.

DuMouchel, W. & Duncan, G. (1983), 'Using sample survey weights in multiple regression analyses of stratifed samples', *Journal of the American Statistical Association* **78**, 535–543.

Greene, W. H. (2012), *Econometric Analysis: Seventh Edition*, Prentice Hall, Saddle River, NJ.

Harms, T. & Duchesne, P. (2010), 'On kernel nonparametric regression designed for complex survey data', *Metrika* (72), 111–138.

Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* (47), 663–685.

Li, Q. & Racine, J. S. (2007), *Nonparametric Econometrics*, Princeton University Press, Princeton, NJ.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, second edn, Brooks/Cole, 20 Channel Center Street, Boston, MA 02210.

Magee, L., Robb, A. & Burbidge, J. (1998), 'On the use of sampling weights when estimating regression models with survey data', *Journal of Econometrics* **84**, 251–271.

Nadarya, E. (1964), 'On estimating regression', *Theory of Probability and its Applications* **9**, 141–142.

Opsomer, J. & Miller, C. (2005), 'Selecting the amount of smoothing in nonparametric regression estimation for complex surveys', *Journal of Nonparametric Statistics* **17**(5), 593–611.

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

Sánchez-Borrego, I., Opsomer, J., Rueda, M. & Arcos, A. (2014), 'Nonparametric estimation with mixed data types in survey sampling', *Rev Mat Complut* (27), 685–700.

Särdinal, C., Swensson, B. & Wretman, J. (1992), *Model-assisted survey sampling*, Springer, New York.

Solon, G., Haider, S. J. & Wooldridge, J. (2013), What are we weighting for?, Working Paper 18859, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138.

StatsCan (2013), 'Canadian community health survey - mental health component'.

Waston, G. (1964), 'Smooth regression analysis', *Sankhya, Series A* **26**, 359–372.

A Proofs

A.1 Proof of Theorem 4.1

Using a combined framework, first find the expectation of $\hat{m}(x)$, $E_C(\hat{m}(x)|x) = E_C(\hat{m}_1(x)|x)$. If Assumptions 4.1-4.4 are satisfied:

$$\begin{aligned}
E_C(\hat{m_1}(x)|x) =& E_\xi\left[E_P(\hat{m}_1(x)|\pi)|x\right] = E_\xi\left[E_P\left(N^{-1}\sum_{i=1}^{N}\pi_i^{-1}\mathbf{1}(i\in\mathcal{S})[g(x_i)-g(x)]K_{h,ix}\right)\right] \\
=& E_\xi\left(N^{-1}\sum_{i=1}^{N}[g(x_i)-g(x)]K_{h,ix}\right) \\
=& \sum_{t^d\in\mathcal{D}}\int_{\mathbb{R}^q}[g(t)-g(x)]f(t)K_{h,tx}dt^c \\
=& \frac{\kappa_2}{2}\sum_{s=1}^{q}h_s^2[g_s s(x)f(x)+2g_s(x)f_s(x)] + \sum_{s=1}^{r}\sum_{t^d\in\mathcal{D}}\mathbf{1}(t^d,x^d)[g(x^c,t^d)-g(x^c,gx^d)]f(x^c,t^d)\lambda_s \\
&+ o_p\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right) \\
=& \sum_{s=1}^{q}h_s^2 B_s(x)f(x) + \sum_{s=1}^{r}D_s\lambda_s + o_p\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right) \qquad \text{(A.1)}
\end{aligned}$$

where $B_s = [g_s s(x)f(x)+2g_s(x)f_s(x)]\kappa_2/2$, $D_s = (t^d,x^d)[g(x^c,t^d)-g(x^c,gx^d)]f(x^c,t^d)$, and $K_{h,tx} = \prod_{s=1}^{q}h_s^{-1}k((t_s-x_s)/h_s)\prod_{s=1}^{r}\lambda_s^{\mathbf{1}(t_s^d\neq x_s^d)}$. By using the Taylor expansion method from Särdinal et al. (1992), Harms and Duchesne (2009) derived the following result:

$$\hat{g}(x)-g(x) \approx \frac{1}{N}\hat{f}^{-1}\left[\sum_{i=1}^{N}\pi_i^{-1}\mathbf{1}(i\in\mathcal{S})K_{h,ix}u_i\right] \qquad \text{(A.2)}$$

where u_i is the population residual. The asymptotic design-based variance of $\hat{g}(x)-g(x)$ is therefore:

$$\begin{aligned}
Avar_P\{\hat{g}(x)-g(x)|\pi\} &= \mathrm{var}_P\left(\frac{1}{N}\hat{f}^{-1}(x)\sum_{i=1}^{N}\pi^{-1}\mathbf{1}(i\in S)K_{h,ix}\right) \\
&= \frac{1}{N^2}\hat{f}^{-2}(x)\sum_i\sum_j\frac{1}{\pi_i\pi_j}\mathrm{var}_P(\mathbf{1}(i\in S))K_{h,ix}K_{h,jx}u_iu_j \\
&= \frac{1}{N^2}\hat{f}^{-2}(x)\sum_i\sum_j\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}K_{h,ix}K_{h,jx}u_iu_j \qquad \text{(A.3)}
\end{aligned}$$

Replacing $\hat{f}(x)$ with the expression $\hat{f}(x) = f(x) + o_p(1)$, the expression for $A\text{var}_P$ becomes:

$$A\text{var}_P\{\hat{g}(x) - g(x)|\pi\} = \frac{1}{N^2} f^{-2}(x) \left[\sum_i \frac{1-\pi_i}{\pi_i} K_{h,ix}^2 u_i^2 + \sum_i \sum_{j \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} K_{h,ix} K_{h,jx} u_i u_j \right] \tag{A.4}$$

The variance under the combined framework is then derived using the following expression:

$$\text{var}_C\{(\hat{g}(x) - g(x))|x\} = \text{var}_\xi\{E_P[\hat{g}(x) - g(x)|\pi]|x\} + E_\xi\{\text{var}_P[\hat{g}(x) - g(x)|\pi]|x\} \tag{A.5}$$

Using a similar derivation for the bias of $\hat{f}(x)$ and $\hat{m}(x)$, the first term in $(A.5)$ becomes the traditional equation for the variance of the local constant estimator:

$$\text{var}_\xi\{E_P[(\hat{g}(x) - g(x))|\pi]|x\} = \text{var}_\xi\{[\tilde{g}(x) - g(x)]|x\}$$

Looking first at the variance of $\hat{m}_1$:

$$\begin{aligned} \frac{1}{N}\text{var}_\xi([g(t) - g(x)]K_{h,tx}) &= \frac{1}{N}\left[\sum_{t^d \in \mathcal{D}} \int_{\mathbb{R}^q} [g(t) - g(x)]^2 f(t^c, t^d) K_{h,tx}^2 dt^c - O\left(\sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s \right) \right] \\ &= O\left((Nh_1 ... h_q)^{-1} \left(\sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s \right) \right). \end{aligned} \tag{A.6}$$

Next, derive the expression for $E_\xi((\hat{m}_2(x))^2|x)$:

$$\begin{aligned} & E_\xi\left[\left(\frac{1}{N} \sum_{i=1}^N u_i K_{h,ix} \right)^2 \right] \\ &= \frac{1}{N} E[\sigma(t)^2 K_{h,tx}^2] \\ &= \frac{1}{N} \sum_{t^d \in \mathcal{D}} \int_{\mathbb{R}^q} \sigma(t)^2 f(t^c, t^d) K_{h,tx}^2 dt^c \\ &= \frac{1}{N}\left[\int_{\mathbb{R}^q} \sigma(x^c + hv, x^d)^2 f(x^c + hv, x^d) \prod_{s=1}^q h_s^{-1} w^2(v_s) dv_s + O\left(\sum_{s=1}^r \lambda_s \right) \right] \\ &= \frac{\kappa^q \sigma^2(x) f(x)}{N h_1 ... h_q} + O\left((Nh_1 ... h_q)^{-1} \left(\sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s \right) \right) \end{aligned} \tag{A.7}$$

Combining (A.6), (A.7), and $\text{var}_\xi\{[\tilde{g}(x) - g(x)]|x] = [f(x)]^{-2}\text{var}_\xi(\tilde{m}(x))$, the first term in equation (A.5) is:

$$\text{var}_\xi\{E_P[(\hat{g}(x) - g(x)]\} = \frac{\kappa^q \sigma^2(x)}{f(x) N h_1 ... h_q} + o_p\left((Nh_1 ... h_q)^{-1} \left(\sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s \right) \right). \tag{A.8}$$

Plugging the result from (A.4) into the second term in equation (A.5) becomes:

$$\begin{aligned}
E_\xi\{\text{var}_P\{\hat{g}(x)-g(x)|\pi\}\} =&(Nf(x))^{-2}E_\xi\big[\sum_i \frac{1-\pi_i}{\pi_i}K_{h,ix}^2u_i^2 \\
&+\sum_i\sum_{j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}K_{h,ix}K_{h,jx}u_iu_j\big] \qquad \text{(A.9)}\\
=&(Nf(x))^{-2}\sum_i\frac{1-\pi_i}{\pi_i}E_\xi(K_{h,ix}^2\sigma^2)\\
=&(Nf(x))^{-2}\sum(x_i)\left(\frac{1-\pi_i}{\pi_i}\right)\sum_{t^d\in\mathcal{D}}\int_{\mathbb{R}^q}\sigma(t)^2f(t^c,t^d)K_{h,tx}^2dt^c\\
=&\sum(x_i)\left(\frac{1-\pi_i}{\pi_i}\right)\frac{\kappa^q\sigma(x)[f(x)]^{-1}}{N^2h_1...h_q}\\
&+o_p\left((Nh_1...h_q)^{-1}\left(\sum_{s=1}^q h_s^2+\sum_{s=1}^r\lambda_s\right)\right) \qquad \text{(A.10)}
\end{aligned}$$

where the second term in (A.9) is a zero mean function. To get the expression for $\text{var}_C\{(\hat{g}(x)-g(x))|x\}$, simply sum (A.8) and (A.10):

$$\begin{aligned}
\text{var}_C\{[(\hat{g}(x)-g(x))|\pi]|x\} =&\frac{1}{nh_1...h_q}\left\{N^{-2}n\sum_{i=1}^N\left(\frac{1-\pi_i}{\pi_i}\right)\frac{\kappa^q\sigma(x)}{f(x)}\right\}\\
&+\frac{\kappa^q\sigma^2(x)f(x)}{Nh_1...h_q}+o_p\left((Nh_1...h_q)^{-1}\left(\sum_{s=1}^q h_s^2+\sum_{s=1}^r\lambda_s\right)\right)\\
=&\frac{1}{nh_1...h_q}\left\{N^{-2}n\sum_{i=1}^N(w_i-1)+\frac{n}{N}\right\}\frac{\kappa^q\sigma(x)}{f(x)}\\
&+o_p\left((Nh_1...h_q)^{-1}\left(\sum_{s=1}^q h_s^2+\sum_{s=1}^r\lambda_s\right)\right). \qquad \text{(A.11)}
\end{aligned}$$

A.2 Proof of Theorem 4.2

In order to prove the aysymptotic normality of $\hat{g}(x)$, I make use of the following theorem taken from the statistical appendix in Li and Racine (2007).

Theorem A.1 (Liapunov Double Array Central Limit Theorem)**.** *Let $\{Z_{n,i}\}$ be a sequence of independent (double array) random variables with $E|Z_{n,i}|^{2+\delta}<\infty$ for some $\delta>0$. Let $S_n=\sum_{i=1}^n Z_{n,i}$, and $\sigma_n^2=var(S_n)=\sum_{i=1}^n\sigma_{n,i}$. If $\sigma_n^2=\sigma^2+o(1)$, and*

$$\lim_{n\to\infty}\sum_{i=1}^n E|(Z_{n,i}-E(Z_{n,i}))|^{2+q}=0 \text{ for some } \delta>0. \qquad \text{(A.12)}$$

then

$$\sigma_n^{-1}(S_n - E(S_n)) = \sigma_n^{-1}\sum_{i-1}^{n}[(Z_{n,i} - E(Z_{n,i})] \xrightarrow{d} N(0,1). \tag{A.13}$$

Turning now to the modified local constant,

$$\begin{aligned}
&\sqrt{Nh_1...h_q}\left(\hat{g}(x) - g(x) - \sum_{s=1}^{q} B_s(x)h_s^2 - \sum_{s=1}^{q} D_s(x)\lambda_s\right) \\
&\quad\equiv \sqrt{Nh_1...h_q}\frac{\left(\hat{g}(x) - g(x) - \sum_{s=1}^{q} B_s(x)h_s^2 - \sum_{s=1}^{q} D_s(x)\lambda_s\right)\hat{f}(x)}{\hat{f}(x)} \\
&\quad= \sqrt{Nh_1...h_q}\frac{\left(\hat{m}(x) - \sum_{s=1}^{q} B_s(x)h_s^2\hat{f}(x) - \sum_{s=1}^{q} D_s(x)\lambda_s\hat{f}(x)\right)}{\hat{f}(x)} \\
&\quad= \sqrt{Nh_1...h_q}\frac{[\hat{m}(x) - E(\hat{m}(x))]}{\hat{f}(x)} + O\left(\sqrt{Nh_1...h_q}\left(\sum_{s=1}^{q} h_s^2 - \sum_{s=1}^{q}\lambda_s\right)\right) \\
&\quad= \sqrt{Nh_1...h_q}\frac{[\hat{m}(x) - E(\hat{m}(x))]}{\hat{f}(x)} + o(1) \\
&\quad= \frac{1}{f(x)}\sum_{i=1}^{N} Z_{N,i} + o(1)
\end{aligned} \tag{A.14}$$

where $Z_{N,i} = (\sqrt{Nh_1...h_q})^{-1}[\pi^{-1}\mathbf{1}(i \in S)(y_i - g(x))K_{h,ix} - E(\pi^{-1}\mathbf{1}(i \in S)(y_i - g(x))K_{h,ix})]$ and $\hat{f}(x) = f(x) + o_p(1)$. Next, take the expectation of the absolute value of $Z_{N,i}$ raised to the power of $2 + \delta$, where δ is some constant and $\delta > 0$:

$$E|Z_{N,i}|^{2+q} = (\sqrt{Nh_1...h_q})^{-(2+q)}E[\pi^{-1}\mathbf{1}(i \in S)(y_i - g(x))K_{h,ix} - E(\pi^{-1}\mathbf{1}(i \in S)(y_i - g(x))K_{h,ix})]^{2+q}$$

Using the C_r inequality and Liapunov's central limit theorem we get (Li & Racine 2007):

$$\begin{aligned}
E|Z_{N,i}|^{2+q} &\leq \frac{2^{1+q}E[\pi^{-1}\mathbf{1}(i \in S)(y_i - g(x))K_{h,ix}]^{2+q}}{(\sqrt{Nh_1...h_q})^{2+q}} \\
&= o(1)
\end{aligned} \tag{A.15}$$

and

$$\frac{1}{f(x)}\sum_{i=1}^{N} Z_{N,i} \xrightarrow{d} N(0, (\Delta + Q)\kappa^q\sigma^2(x)/f(x))$$

A.3 Proof of Theorem 4.3

The proof for the bias of $\hat{g}(x)$ with weakly dependent data is the same as in section $A.1$ of this appendix. Looking at the variance of $\hat{g}(x) - g(x)$ under the combined framework, we get:

$$\text{var}_C\{(\hat{g}(x) - g(x))|x\} = \text{var}_\xi\{E_P[\hat{g}(x) - g(x)|\pi]|x\} + E_\xi\{\text{var}_P[\hat{g}(x) - g(x)|\pi]|x\}$$

As above, the first term reduces to $\text{var}_\xi\{\hat{g}(x)-g(x)|x\}$. Again, defining $\hat{m}_1(x)$ and $\hat{m}_2(x)$ as above, I first look at the variance of $\hat{m}_1(x)$:

$$\begin{aligned}
\text{var}_\xi\{\hat{m}_1(x)|x\} =& \text{var}_\xi\left[\frac{1}{N}\sum_{i=1}^{N}(g(x_i)-g(x))K_{h,ix}\right] \\
=& \frac{1}{N^2}\text{var}_\xi\left(\sum_{i=1}^{N}(g(x_i)-g(x))K_{h,ix}\right) \\
=& \frac{1}{N^2}\text{var}_\xi\left(\sum_{i=1}^{N}(g(x_i)-g(x))K_{h,ix}\right) \\
=& \frac{1}{N^2}\left(\sum_{i=1}^{N}\text{var}_\xi[(g(x_i)-g(x))K_{h,ix}]\right) \\
&+ \frac{1}{N^2}\left(\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j\neq i}^{N_c}\text{cov}\left\{[(g(x_{ci})-g(x))K_{h,ix}],[(g(x_{cj})-g(x))K_{h,jx}]\right\}\right) \qquad \text{(A.16)}
\end{aligned}$$

In appendix **A.1**, $\text{var}_\xi[(g(x_i)-g(x))K_{h,ix}]$ was shown to be $O\left((h_1...h_q)^{-1}\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)\right)$. The second term in$(A.16)$ is present because the data is no longer independent within clusters. Because the data is cross-sectional, it is stationary, and assuming ρ-mixing we can write:

$$\begin{aligned}
|\text{cov}\{(g(x_{ci})-g(x))K_{h,ix},(g(x_{cj})-g(x))K_{h,jx}\}| \leq& \rho(j-i)\text{var}([(g(x_i)-g(x))K_{h,ix}]) \\
=& O(h_1...h_q) \qquad \text{(A.17)}
\end{aligned}$$

Therefore, $\text{var}_\xi(\hat{m}_1(x))$ can be written as:

$$\begin{aligned}
\text{var}_\xi(\hat{m}_1(x)) =& \frac{1}{N^2}\left(\sum_{i=1}^{N}\text{var}_\xi[(g(x_i)-g(x))K_{h,ix}]\right) \\
&+ \frac{1}{N^2}\left(\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j\neq i}^{N_c}\text{cov}\left\{[(g(x_{ci})-g(x))K_{h,ix}],[(g(x_{cj})-g(x))K_{h,jx}]\right\}\right) \\
\leq& \frac{1}{N^2}\left(N\text{var}_\xi[(g(x_i)-g(x))K_{h,ix}]+N\text{var}_\xi[(g(x_i)-g(x))K_{h,ix}]\sum_{t=1}^{\infty}\rho(t)\right) \\
=& O\left((Nh_1...h_q)^{-1}\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)\right) \\
=& o((Nh_1...h_q)^{-1}) \qquad \text{(A.18)}
\end{aligned}$$

Combining this result with the equation for the bias of $\hat{m}_1(x)$ from appendix **A.1** implies that

$$\hat{m}_1(x)=\sum_{s=1}^{q}B_s(x)h_s^2+\sum_{s=1}^{r}D_s\lambda_s+o\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s+(Nh_1...h_q)^{-1}\right)$$

Furthermore, the expression for $E((\hat{m}_2(x))^2)$ is the same as under the i.i.d. case, i.e.

$$E((\hat{m}_2(x))^2) = \frac{\kappa^q \sigma^2(x) f(x)}{N h_1 ... h_q} + o_p\left((N h_1 ... h_q)^{-1}\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s\right)\right)$$

$$\text{MSE}(\hat{g}(x)) = o_p((N h_1 ... h_q)^{-1} \sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s$$

A.4 Cross-Validation

Write $g(x_j) = g(x_j) + g(x_i) - g(x_i) = g(x_i) + R_{ij}$. Plug into the regression model $y_j = g(x_j) + u_j$:

$$y_j = g(x_i) + R_{ij} + u_j$$

Then, we can re-write the leave-one-out kernel estimator for $\hat{g}(x)$ as:

$$\begin{aligned}
\hat{g}_{-i}(x_i) &= \left[\sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij}\right]^{-1} \sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij} y_j \\
&= \left[\sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij}\right]^{-1} \sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij} (g(x_i) + R_{ij} + u_j) \\
&= g(x_i) + \left[\sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij}\right]^{-1} \sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij} (R_{ij} + u_j) \qquad \text{(A.19)}
\end{aligned}$$

Using the definition for the modified kernel density estimator $\hat{f}(x^c, x^d)$ we can re-write equation $(A.19)$ as:

$$\hat{g}_{-1}(x_i) = g(x_i) + \frac{1}{N} \hat{f}_i^{-1} \sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij} (R_{ij} + u_j) \qquad \text{(A.20)}$$

where $\hat{f}_i = N^{-1} \sum_{i \neq j} \pi_j^{-1} \mathbf{1}(j \in \mathcal{S}) K_{h,ij}$. The definition for $CV(h)$ can now be written as:

$$\begin{aligned}
CV(h) =& N^{-1} \sum_{i,j \in U} (y_i - \hat{g}_{-i}(x_i))^2 M(x_i) \\
=& N^{-1} \sum_{i,j \in U} (g(x_i) + u_i - \hat{g}_{-i}(x_i))^2 M(x_i) \\
=& N^{-1} \sum_{i,j \in U} (g(x_i) - \hat{g}_{-i}(x_i))^2 - 2N^{-1} \sum_{i,j \in U} [u_i (g(x_i) - \hat{g}_{-i}(x_i))] M(x_i) + N^{-1} \sum_{i,j \in U} u_i^2 M(x_i)
\end{aligned} \qquad \text{(A.21)}$$

The third term in equation $(A.21)$ does not depend on (h, λ) and the second term has an order smaller than the first term. So asymptotically, minimizing $CV_{lc}(h, \lambda)$ is equivalent to minimizing

$\sum_{i,j\in U}[g(x_i)-g_{-i}(X_i)]^2M(x_i)$.

$$\begin{aligned}
CV_0(h,\lambda) =& N^{-1}\sum_{i,j\in U}[g(x_i)-g(x_i)-\frac{1}{N}\sum_{i\neq j}\pi_j^{-1}K_{h,ij}(R_{ij}+u_j)\hat{f}_i^{-1}]^2M(X_i)\\
=&\frac{1}{N}\sum_{i,j\in U}\left[\frac{1}{N}\sum_{i\neq j}\pi_j^{-1}\mathbf{1}(j\in\mathcal{S})K_{h,ij}(R_{ij}+u_j)\hat{f}_i^{-1}\right]^2M(x_i)\\
=&\frac{1}{N}\sum_{i,j\in U}\left[\frac{1}{N}\sum_{i\neq j}\pi_j^{-1}K_{h,ij}(g(x_j)-g(x_i)+u_j)\hat{f}_i^{-1}\right]^2M(x_i)\\
=&\frac{1}{N}\sum_{i,j\in U}\left\{\left[\frac{1}{N}\sum_{i\neq j}\pi_j^{-1}K_{h,ij}(g(x_j)-g(x_i))+\frac{1}{N}\sum_{i\neq j}\pi_j^{-1}K_{h,ij}u_j\right]\hat{f}_i^{-1}\right\}^2M(x_i)
\end{aligned}\tag{A.22}$$

Again, using the definition $\hat{f}(x)=f(x)+o_p(1)$, write $CV_0(h,\lambda)$ as:

$$\begin{aligned}
CV_0(h,\lambda) =&\frac{1}{N}\sum_{i,j\in U}(m_{1i}+m_{2i})^2f_i^{-2}M(x_i)+(s.o)\\
=&\frac{1}{N}\left(\sum_{i,j\in U}m_{1i}^2f_i^{-2}M(x_i)+\sum_{i,j\in U}m_{2i}^2f_i^{-2}M(x_i)+\sum_{i,j\in U}m_{1i}m_{2i}f_i^{-2}M(x_i)\right)
\end{aligned}\tag{A.23}$$

where $m_{1i}=1/N\sum_{i\neq j}\pi_j^{-1}K_{h,ij}(g(x_j)-g(x_i))$, $m_{2i}=1/N\sum_{i\neq j}\pi_j^{-1}K_{h,ij}u_j$, and $s.o$ denotes smaller order terms. The leading term of $CV(h,\lambda)$ is $CV_0(h,\lambda)=E[CV_0(h,\lambda)]+(s.o.)$.

$$\begin{aligned}
E_C[CV_0(h,\lambda)] &= E_C\left\{\frac{1}{N}\left(\sum_{i,j\in U}m_{1i}^2f_i^{-2}M(x_i)+\sum_{i,j\in U}m_{2i}^2f_i^{-2}M(x_i)+\sum_{i,j\in U}m_{1i}m_{2i}f_i^{-2}M(x_i)\right)\right\}\\
&= E_C[m_{1i}^2f_i^{-2}M(x_i)]+E_C[m_{2i}^2f_i^{-2}M(x_i)]
\end{aligned}\tag{A.24}$$

because $E_C(m_{1i}m_{2i}f_i^{-2}M(x_i))=0$. Looking at the first term in equation $(A.24)$:

$$\begin{aligned}
E_C[m_{1i}^2f_i^{-2}|x_i] =&E_C[\frac{1}{N}\sum_{i\neq j}\frac{1}{N}\sum_{i\neq l}\pi_j^{-1}\mathbf{1}(j\in\mathcal{S})K_{h,ij}(g(x_j)-g(x_i))\\
&\times\pi_l^{-1}\mathbf{1}(l\in\mathcal{S})K_{h,il}(g(x_j)-g(x_i))f_i^{-2}M(x_i)]\\
=&\frac{1}{N^2}E_C\big[\sum_{j\neq i}\pi_j^{-1}\mathbf{1}(j\in\mathcal{S})K_{h,ij}(g(x_j)-g(x_i))\sum_{l\neq i}\pi_l^{-1}\mathbf{1}(l\in\mathcal{S})K_{h,il}(g(x_j)\\
&-g(x_i))f_i^{-2}M(x_i)\big]+\frac{1}{N^2}E_C[\sum_{j\neq i}\pi_j^{-2}\mathbf{1}(j\in\mathcal{S})K_{h,ij}^2(g(x_j)-g(x_i))^2]
\end{aligned}\tag{A.25}$$

Following Li and Racine (2004), compute $E(R_{ij}K_{h,ij}f_i^{-1}|x_i)$:

$$\begin{aligned}
E_C(R_{ij}K_{h,ij}f_i^{-1}|x_i) =& E_\xi[E_D[\sum_{j\neq i}\pi_j^{-1}\mathbf{1}(j\in\mathcal{S})K_{h,ij}(g(x_j)-g(x_i))f_i^{-1}|\pi]|x] \\
=& E_\xi[K_{h,ij}(g(x_j)-g(x_i))f_i^{-1}|x] \\
=& \frac{\kappa_2}{2}\sum_{s=1}^{q}[g_{ss}(X_i)f(X_i)+2g_s(X_i)f_s(X_i)]f^{-1}(X_i) \\
&+\sum_{s=1}^{r}\sum_{v^d\in\mathcal{D}}\mathbf{1}(x^d,v^d)[g(x^c,v^d)-g(x)]f(x^c,x^d)\lambda_s f(X_i)^{-1} \\
&+O\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right). \qquad \text{(A.26)}
\end{aligned}$$

v^d is a placeholder term where it is assumed that the data are identically distributed. Then,

$$\begin{aligned}
&E_\xi[K_{h,ij}(g(x_j)-g(x_i))K_{h,il}(g(x_j)-g(x_i))f_i^{-2}] \\
&= E_\xi[K_{h,ij}(g(x_j)-g(x_i))f_i^{-1}]E_\xi[K_{h,il}(g(x_j)-g(x_i))f_i^{-1})] \\
&= \{E_\xi[K_{h,ij}(g(x_j)-g(x_i))]f_i^{-1}\}^2 \\
&= \left\{\sum_{s=1}^{q}B_s(x)h_s^2+\sum_{s=1}^{r}D_s\lambda_s\right\}^2+O\left(\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)^3\right) \qquad \text{(A.27)}
\end{aligned}$$

The second term in $(A.25)$ is $O((Nh_1...h_q)^{-1}(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s))$:

$$\begin{aligned}
&N^{-1}E_C[\sum_{j\neq i}\pi_j^{-2}\mathbf{1}(j\in\mathcal{S})K_{h,ij}^2(g(x_j)-g(x_i))^2] \\
&= N^{-1}E_\xi\left[\sum_{j\neq i}\pi_j^{-1}K_{h,ij}^2(g(x_j)-g(x_i))^2f_i^2\right] \\
&= N^{-1}f_i^{-2}\sum_{j\neq i}\pi_j^{-1}\sum_{x^d\in\mathcal{D}}\int K_{h,ij}^2(g(x_j)-g(x_i))^2dx_j^c \\
&= O((Nh_1...h_q)^{-1}(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s)) \qquad \text{(A.28)}
\end{aligned}$$

Using equations (A.27) and (A.28):

$$\begin{aligned}
E[\hat{m}_{1i}^2f_i^{-2}M(x_i)] =& \sum_{x^d\in\mathcal{D}}\int\left\{\sum_{s=1}^{q}B_s(x)h_s^2+\sum_{s=1}^{r}D_s\lambda_s\right\}^2f(x)M(x)dx^c \\
&+O\left(\left(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s\right)^3+(Nh_1...h_q)^{-1}(\sum_{s=1}^{q}h_s^2+\sum_{s=1}^{r}\lambda_s)\right). \qquad \text{(A.29)}
\end{aligned}$$

Next, solve for the second term on the right hand side of (A.23).

$$\begin{aligned} E_C[\hat{m}_{2i}^2 f_i^{-2} M(x_i)] =& E_C\{f_i^{-2} M(x_i) E_C[\hat{m}_{1i}^2 | x_i]\} \\ =& \frac{1}{N^2} E_C\{f_i^{-2} M(x_i) E_C[\sum_{j \neq i} \pi_j^{-2} \mathbf{1}(i \in \mathcal{S}) u_j^2 K_{h,ij}^2 | x_i]\} \\ =& \left(N^2 \prod_{s=1}^{q} h_s \right)^{-1} \sum_{j \neq i} \pi_j^{-1} \kappa^q \sum_{x^d \in \mathcal{D}} \int \sigma^2(x) M(x) dx^c \\ & + O\left(\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s \right)^3 + (N^2 h_1 ... h_q)^{-1} (\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s) \right) \end{aligned} \tag{A.30}$$

$$E[CV_0(h, \lambda)] = \sum_{x^d \in \mathcal{D}} \int \left(\left\{ \sum_{s=1}^{q} B_s(x) h_s^2 + \sum_{s=1}^{r} D_s \lambda_s \right\}^2 f(x) + \sum_{j \neq i} \pi_j^{-1} \frac{\kappa^q \sigma(x)}{N^2 h_1 ... h_q} \right) M(x) dx \tag{A.31}$$

Minimizing $CV(h)$ is equivalent to minimizing $CV_1(h)$ because $n^{-1} \sum_{i,j \in U} u_i$ is not related to h$h_1, ...h_q$.

B Tables

Table 1: Four population conditional mean functions

Name	Expression
Linear	$g_1(X) = x_1^c + \beta x_1^d + \beta x_2^d$
Quadratic	$g_2(X) = 1 + 2(x_1^c - 0.5)^2 + \beta x_1 + \beta x_2$
Bump	$g_3(X) = 1 + x_1 + 2(x_1^c - 0.5)^2 + e^{-200(x_1^c - 0.5)^2} + \beta x_1^d + \beta x_2^d$
Harlde	$g_4(X) = \sin^3(2\pi_0 x_1^c) + \beta x_1^d + \beta x_2^d$, $\pi_0 = 3.1415$

Table 2: Strata borders

Strata Variable	Strata Borders	Sample size
x_1^c	$x_1^c \leq 0.40$	$n/2$
	$0.40 < x_1^c \leq 0.8$	$n/5$
	$x_1^c > 0.8$	$3 * n/10$
y	$y \leq$15% quantile	$n/2$
	15% quantile$< y \leq$85% quantile	$n/5$
	$y >$ 85% quantile	$3 * n/10$

Table 3: Simulation results for median of MSE and bandwidths from simple random sampling

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Linear	200	0.25	0.0085	0.0085	0.0630	0.0630	0.0125	0.0125	0.0128	0.0128
			(0.0015)	(0.0015)						
Linear	200	0.50	0.0228	0.0228	0.0915	0.0914	0.0323	0.0324	0.0324	0.0324
			(0.0048)	(0.0048)						
Linear	200	1.00	0.0645	0.0647	0.1254	0.1253	0.0802	0.0802	0.0812	0.0812
			(0.0176)	(0.0176)						
Linear	200	2.00	0.1985	0.1946	0.1563	0.1559	0.1912	0.1956	0.1840	0.1881
			(0.0699)	(0.0606)						
Linear	400	0.25	0.0049	0.0049	0.0541	0.0541	0.0075	0.0076	0.0075	0.0076
			(0.0008)	(0.0008)						
Linear	400	0.50	0.0133	0.0133	0.0784	0.0784	0.0189	0.0189	0.0193	0.0193
			(0.0024)	(0.0024)						
Linear	400	1.00	0.0380	0.0380	0.1089	0.1089	0.0534	0.0534	0.0521	0.0521
			(0.0082)	(0.0082)						
Linear	400	2.00	0.1136	0.1140	0.1445	0.1445	0.1291	0.1296	0.1298	0.1304
			(0.0337)	(0.0339)						
Linear	800	0.25	0.0028	0.0028	0.0467	0.0467	0.0042	0.0042	0.0040	0.0040
			(0.0004)	(0.0004)						
Linear	800	0.50	0.0078	0.0078	0.0668	0.0668	0.0112	0.0112	0.0111	0.0111
			(0.0015)	(0.0015)						
Linear	800	1.00	0.0216	0.0216	0.0935	0.0935	0.0302	0.0302	0.0307	0.0307
			(0.0047)	(0.0047)						
Linear	800	2.00	0.0672	0.0672	0.1280	0.1279	0.0832	0.0832	0.0826	0.0826
			(0.0175)	(0.0176)						
Quadratic	200	0.25	0.0130	0.0130	0.0403	0.0402	0.0079	0.0079	0.0083	0.0083
			(0.0018)	(0.0018)						
Quadratic	200	0.50	0.0340	0.0341	0.0581	0.0581	0.0226	0.0226	0.0216	0.0216
			(0.0059)	(0.0059)						
Quadratic	200	1.00	0.0943	0.0941	0.0836	0.0836	0.0570	0.0570	0.0572	0.0572
			(0.0188)	(0.0188)						
Quadratic	200	2.00	0.2781	0.2781	0.1169	0.1166	0.1269	0.1278	0.1234	0.1240
			(0.0746)	(0.0747)						
Quadratic	400	0.25	0.0077	0.0077	0.0335	0.0335	0.0053	0.0053	0.0053	0.0053
			(0.0010)	(0.0010)						
Quadratic	400	0.50	0.0208	0.0208	0.0489	0.0488	0.0135	0.0135	0.0129	0.0129
			(0.0033)	(0.0033)						

*Values in brackets are the median absolue deviation. *Continued on next page*

Table 3 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Quadratic	400	1.00	0.0575	0.0575	0.0705	0.0705	0.0358	0.0358	0.0373	0.0373
			(0.0106)	(0.0106)						
Quadratic	400	2.00	0.1597	0.1596	0.1008	0.1008	0.0913	0.0913	0.0909	0.0911
			(0.0397)	(0.0397)						
Quadratic	800	0.25	0.0045	0.0045	0.0282	0.0282	0.0032	0.0032	0.0031	0.0031
			(0.0005)	(0.0005)						
Quadratic	800	0.50	0.0124	0.0124	0.0408	0.0408	0.0079	0.0079	0.0082	0.0082
			(0.0017)	(0.0017)						
Quadratic	800	1.00	0.0341	0.0342	0.0595	0.0595	0.0215	0.0215	0.0216	0.0216
			(0.0058)	(0.0058)						
Quadratic	800	2.00	0.0963	0.0963	0.0838	0.0838	0.0553	0.0553	0.0565	0.0565
			(0.0189)	(0.0189)						
Bump	200	0.25	0.0129	0.0129	0.0307	0.0307	0.1222	0.1222	0.1237	0.1237
			(0.0019)	(0.0019)						
Bump	200	0.50	0.0323	0.0321	0.0418	0.0417	0.2520	0.2521	0.2295	0.2313
			(0.0054)	(0.0051)						
Bump	200	1.00	0.0836	0.0837	0.0576	0.0574	0.4658	0.4640	0.4543	0.4521
			(0.0194)	(0.0195)						
Bump	200	2.00	0.2059	0.2062	0.0929	0.0921	0.7424	0.7255	0.8107	0.8178
			(0.0588)	(0.0590)						
Bump	400	0.25	0.0078	0.0078	0.0271	0.0270	0.0795	0.0795	0.0802	0.0802
			(0.0009)	(0.0009)						
Bump	400	0.50	0.0197	0.0197	0.0360	0.0360	0.1729	0.1730	0.1755	0.1755
			(0.0028)	(0.0028)						
Bump	400	1.00	0.0519	0.0520	0.0466	0.0465	0.3427	0.3432	0.3583	0.3591
			(0.0109)	(0.0109)						
Bump	400	2.00	0.1301	0.1303	0.0742	0.0743	0.5742	0.5758	0.6549	0.6481
			(0.0314)	(0.0000)						
Bump	800	0.25	0.0046	0.0046	0.0235	0.0235	0.0492	0.0493	0.0497	0.0497
			(0.0006)	(0.0006)						
Bump	800	0.50	0.0124	0.0124	0.0315	0.0316	0.1162	0.1162	0.1186	0.1186
			(0.0017)	(0.0017)						
Bump	800	1.00	0.0321	0.0321	0.0412	0.0412	0.2499	0.2499	0.2507	0.2507
			(0.0056)	(0.0055)						
Bump	800	2.00	0.0821	0.0821	0.0571	0.0570	0.4758	0.4764	0.4584	0.4584
			(0.0188)	(0.0189)						
Hardle	200	0.25	0.0231	0.0231	0.0178	0.0178	0.0618	0.0617	0.0614	0.0614

*Values in brackets are the median absolue deviation. *Continued on next page*

Table 3 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
			(0.0025)	(0.0025)						
Hardle	200	0.50	0.0592	0.0592	0.0234	0.0234	0.1392	0.1392	0.1431	0.1432
			(0.0080)	(0.0080)						
Hardle	200	1.00	0.1486	0.1480	0.0324	0.0322	0.2895	0.2895	0.2750	0.2763
			(0.0234)	(0.0224)						
Hardle	200	2.00	0.3665	0.3672	0.0565	0.0567	0.4995	0.5002	0.4642	0.4604
			(0.0786)	(0.0783)						
Hardle	400	0.25	0.0145	0.0145	0.0157	0.0157	0.0401	0.0401	0.0408	0.0408
			(0.0015)	(0.0015)						
Hardle	400	0.50	0.0372	0.0372	0.0206	0.0206	0.0981	0.0981	0.0940	0.0941
			(0.0040)	(0.0040)						
Hardle	400	1.00	0.0948	0.0949	0.0274	0.0274	0.2139	0.2140	0.2145	0.2146
			(0.0136)	(0.0137)						
Hardle	400	2.00	0.2363	0.2358	0.0407	0.0408	0.3937	0.3924	0.4158	0.4159
			(0.0413)	(0.0410)						
Hardle	800	0.25	0.0087	0.0087	0.0136	0.0136	0.0242	0.0242	0.0247	0.0247
			(0.0007)	(0.0007)						
Hardle	800	0.50	0.0230	0.0230	0.0182	0.0182	0.0623	0.0624	0.0599	0.0599
			(0.0023)	(0.0023)						
Hardle	800	1.00	0.0595	0.0595	0.0239	0.0239	0.1457	0.1457	0.1475	0.1475
			(0.0074)	(0.0074)						
Hardle	800	2.00	0.1490	0.1491	0.0321	0.0321	0.2873	0.2870	0.2899	0.2899
			(0.0214)	(0.0214)						

*Values in brackets are the median absolue deviation.

Table 4: Simulation results for median of MSE and bandwidths from stratification on y

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Linear	200	0.25	0.0143	0.0230	0.0446	0.0568	0.0020	0.0036	0.0019	0.0035
			(0.0029)	(0.0030)						
Linear	200	0.50	0.0638	0.1087	0.0461	0.1309	0.0056	0.0580	0.0053	0.0571
			(0.0208)	(0.0100)						
Linear	200	1.00	0.2066	0.3635	0.0581	0.1393	0.0081	0.0864	0.0120	0.0885
			(0.0879)	(0.0582)						
Linear	200	2.00	0.4287	0.9973	0.0789	0.1476	0.0404	0.1568	0.0352	0.1451
			(0.1727)	(0.2234)						

*Values in brackets are the median absolue deviation. *Continued on next page*

Table 4 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Linear	400	0.25	0.0088	0.0190	0.0403	0.0481	0.0005	0.0027	0.0005	0.0031
			(0.0019)	(0.0019)						
Linear	400	0.50	0.0344	0.1052	0.0467	0.0903	0.0005	0.0280	0.0005	0.0275
			(0.0099)	(0.0081)						
Linear	400	1.00	0.1161	0.3593	0.0560	0.1234	0.0005	0.0611	0.0005	0.0602
			(0.0381)	(0.0399)						
Linear	400	2.00	0.2520	0.9135	0.0733	0.1365	0.0005	0.0980	0.0005	0.1008
			(0.0974)	(0.1458)						
Linear	800	0.25	0.0051	0.0159	0.0345	0.0411	0.0005	0.0017	0.0005	0.0019
			(0.0009)	(0.0013)						
Linear	800	0.50	0.0196	0.0981	0.0397	0.0730	0.0005	0.0158	0.0005	0.0158
			(0.0044)	(0.0055)						
Linear	800	1.00	0.0681	0.3668	0.0446	0.1055	0.0005	0.0396	0.0005	0.0400
			(0.0180)	(0.0273)						
Linear	800	2.00	0.1513	0.9300	0.0623	0.1192	0.0005	0.0639	0.0005	0.0660
			(0.0463)	(0.1031)						
Quadratic	200	0.25	0.0174	0.0243	0.0326	0.0423	0.0018	0.0061	0.0021	0.0070
			(0.0028)	(0.0031)						
Quadratic	200	0.50	0.0793	0.1043	0.0343	0.0715	0.0029	0.0529	0.0028	0.0520
			(0.0187)	(0.0103)						
Quadratic	200	1.00	0.2427	0.3692	0.0499	0.0805	0.0022	0.0664	0.0005	0.0649
			(0.0776)	(0.0585)						
Quadratic	200	2.00	0.5112	1.0755	0.0737	0.0995	0.0005	0.1043	0.0036	0.1062
			(0.1975)	(0.1998)						
Quadratic	400	0.25	0.0112	0.0188	0.0272	0.0335	0.0005	0.0043	0.0005	0.0042
			(0.0018)	(0.0021)						
Quadratic	400	0.50	0.0447	0.0976	0.0341	0.0583	0.0005	0.0375	0.0005	0.0374
			(0.0104)	(0.0077)						
Quadratic	400	1.00	0.1345	0.3618	0.0451	0.0698	0.0005	0.0484	0.0005	0.0479
			(0.0391)	(0.0397)						
Quadratic	400	2.00	0.2936	0.9694	0.0657	0.0856	0.0005	0.0735	0.0005	0.0752
			(0.0911)	(0.1377)						
Quadratic	800	0.25	0.0068	0.0152	0.0236	0.0282	0.0005	0.0031	0.0005	0.0030
			(0.0009)	(0.0012)						
Quadratic	800	0.50	0.0254	0.0929	0.0301	0.0478	0.0005	0.0252	0.0005	0.0253
			(0.0048)	(0.0052)						
Quadratic	800	1.00	0.0845	0.3606	0.0353	0.0594	0.0005	0.0347	0.0005	0.0340

*Values in brackets are the median absolue deviation. *Continued on next page*

Table 4 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
			(0.0201)	(0.0274)						
Quadratic	800	2.00	0.1834	0.9575	0.0498	0.0733	0.0005	0.0495	0.0005	0.0487
			(0.0471)	(0.0947)						
Bump	200	0.25	0.0213	0.0281	0.0219	0.0323	0.0479	0.1201	0.0526	0.1186
			(0.0045)	(0.0030)						
Bump	200	0.50	0.0985	0.1274	0.0199	0.0443	0.0679	0.2620	0.0762	0.2755
			(0.0353)	(0.0128)						
Bump	200	1.00	0.3187	0.4853	0.0251	0.0481	0.1662	0.4613	0.1914	0.5034
			(0.1245)	(0.0634)						
Bump	200	2.00	0.5859	1.1416	0.0423	0.0758	0.3663	0.5664	0.4686	0.6306
			(0.2520)	(0.2232)						
Bump	400	0.25	0.0136	0.0240	0.0206	0.0289	0.0468	0.0780	0.0485	0.0809
			(0.0026)	(0.0021)						
Bump	400	0.50	0.0538	0.1196	0.0213	0.0396	0.0880	0.2012	0.0791	0.1934
			(0.0163)	(0.0087)						
Bump	400	1.00	0.1899	0.4618	0.0214	0.0408	0.1464	0.3323	0.1570	0.3505
			(0.0777)	(0.0398)						
Bump	400	2.00	0.3524	1.1102	0.0328	0.0566	0.3569	0.5048	0.3540	0.4561
			(0.1430)	(0.1561)						
Bump	800	0.25	0.0081	0.0206	0.0194	0.0248	0.0363	0.0514	0.0361	0.0510
			(0.0013)	(0.0015)						
Bump	800	0.50	0.0309	0.1132	0.0205	0.0349	0.0718	0.1314	0.0736	0.1360
			(0.0071)	(0.0063)						
Bump	800	1.00	0.1049	0.4512	0.0205	0.0357	0.1153	0.2488	0.1259	0.2567
			(0.0352)	(0.0285)						
Bump	800	2.00	0.1946	1.0934	0.0292	0.0458	0.2843	0.3927	0.2859	0.3869
			(0.0672)	(0.1130)						
Hardle	200	0.25	0.0352	0.0392	0.0124	0.0124	0.0062	0.0019	0.0057	0.0028
			(0.0038)	(0.0035)						
Hardle	200	1.00	0.1243	0.1680	0.0165	0.0220	0.0325	0.0910	0.0296	0.0919
			(0.0195)	(0.0151)						
Hardle	200	1.00	0.3320	0.4723	0.0233	0.0459	0.0994	0.1831	0.0949	0.1769
			(0.0815)	(0.0632)						
Hardle	200	2.00	0.7148	1.1880	0.0312	0.0548	0.2420	0.3532	0.2383	0.3551
			(0.2107)	(0.2186)						
Hardle	400	0.25	0.0246	0.0316	0.0109	0.0108	0.0064	0.0016	0.0067	0.0014
			(0.0025)	(0.0023)						

*Values in brackets are the median absolue deviation. *Continued on next page*

Table 4 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Hardle	400	0.50	0.0822	0.1393	0.0152	0.0186	0.0281	0.0737	0.0273	0.0727
			(0.0132)	(0.0095)						
Hardle	400	1.00	0.2137	0.4353	0.0200	0.0329	0.0725	0.1293	0.0694	0.1325
			(0.0480)	(0.0403)						
Hardle	400	2.00	0.4427	1.0907	0.0263	0.0428	0.1814	0.2661	0.1709	0.2561
			(0.1169)	(0.1346)						
Hardle	800	0.25	0.0156	0.0252	0.0102	0.0096	0.0052	0.0018	0.0057	0.0024
			(0.0016)	(0.0016)						
Hardle	800	0.50	0.0505	0.1182	0.0150	0.0157	0.0133	0.0540	0.0144	0.0534
			(0.0071)	(0.0064)						
Hardle	800	1.00	0.1340	0.3967	0.0189	0.0266	0.0331	0.0974	0.0318	0.0971
			(0.0256)	(0.0250)						
Hardle	800	2.00	0.2860	1.0267	0.0216	0.0340	0.1113	0.1959	0.1064	0.1945
			(0.0606)	(0.0944)						

*Values in brackets are the median absolue deviation.

Table 5: Simulation results for median of MSE and bandwidths from stratification on x_1^c

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Linear	200	0.25	0.0092	0.0092	0.0594	0.0557	0.0130	0.0126	0.0125	0.0124
			(0.0015)	(0.0014)						
Linear	200	0.50	0.0251	0.0251	0.0836	0.0811	0.0351	0.0331	0.0340	0.0327
			(0.0052)	(0.0050)						
Linear	200	1.00	0.0712	0.0712	0.1121	0.1251	0.0921	0.0794	0.0878	0.0794
			(0.0178)	(0.0166)						
Linear	200	2.00	0.2329	0.2329	0.1420	0.1800	0.2137	0.1992	0.2229	0.1961
			(0.0758)	(0.0586)						
Linear	400	0.25	0.0054	0.0054	0.0511	0.0468	0.0082	0.0080	0.0081	0.0080
			(0.0008)	(0.0008)						
Linear	400	0.50	0.0146	0.0146	0.0720	0.0681	0.0207	0.0204	0.0212	0.0205
			(0.0026)	(0.0025)						
Linear	400	1.00	0.0417	0.0417	0.0994	0.1029	0.0577	0.0508	0.0571	0.0525
			(0.0100)	(0.0097)						
Linear	400	2.00	0.1249	0.1249	0.1285	0.1585	0.1438	0.1263	0.1383	0.1250
			(0.0334)	(0.0323)						

*Values in brackets are the median absolute deviation. *Continued on next page*

Table 5 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Linear	800	0.25	0.0030	0.0030	0.0432	0.0394	0.0044	0.0046	0.0045	0.0049
			(0.0004)	(0.0005)						
Linear	800	0.50	0.0084	0.0084	0.0617	0.0569	0.0119	0.0117	0.0122	0.0120
			(0.0014)	(0.0015)						
Linear	800	1.00	0.0239	0.0239	0.0856	0.0840	0.0334	0.0318	0.0342	0.0323
			(0.0046)	(0.0048)						
Linear	800	2.00	0.0702	0.0702	0.1161	0.1310	0.0916	0.0795	0.0901	0.0784
			(0.0187)	(0.0170)						
Quadratic	200	0.25	0.0142	0.0141	0.0364	0.0348	0.0058	0.0058	0.0058	0.0057
			(0.0018)	(0.0018)						
Quadratic	200	0.50	0.0380	0.0383	0.0537	0.0511	0.0226	0.0221	0.0232	0.0236
			(0.0060)	(0.0061)						
Quadratic	200	1.00	0.1060	0.1064	0.0764	0.0741	0.0609	0.0589	0.0612	0.0594
			(0.0217)	(0.0207)						
Quadratic	200	2.00	0.2976	0.2967	0.1092	0.1092	0.1464	0.1380	0.1369	0.1293
			(0.0764)	(0.0738)						
Quadratic	400	0.25	0.0083	0.0084	0.0311	0.0297	0.0050	0.0052	0.0055	0.0055
			(0.0010)	(0.0010)						
Quadratic	400	0.50	0.0224	0.0228	0.0454	0.0430	0.0145	0.0144	0.0147	0.0150
			(0.0034)	(0.0034)						
Quadratic	400	1.00	0.0639	0.0635	0.0654	0.0618	0.0390	0.0389	0.0390	0.0392
			(0.0123)	(0.0114)						
Quadratic	400	2.00	0.1758	0.1749	0.0912	0.0890	0.0955	0.0928	0.0978	0.0954
			(0.0393)	(0.0381)						
Quadratic	800	0.25	0.0048	0.0049	0.0264	0.0253	0.0031	0.0031	0.0033	0.0033
			(0.0005)	(0.0005)						
Quadratic	800	0.50	0.0134	0.0135	0.0383	0.0362	0.0090	0.0093	0.0086	0.0087
			(0.0018)	(0.0019)						
Quadratic	800	1.00	0.0374	0.0383	0.0551	0.0516	0.0231	0.0228	0.0232	0.0235
			(0.0061)	(0.0060)						
Quadratic	800	2.00	0.1028	0.1043	0.0776	0.0749	0.0625	0.0596	0.0639	0.0609
			(0.0202)	(0.0186)						
Bump	200	0.25	0.0117	0.0115	0.0344	0.0338	0.1157	0.1142	0.1198	0.1195
			(0.0018)	(0.0017)						
Bump	200	0.50	0.0300	0.0291	0.0463	0.0434	0.2390	0.2397	0.2426	0.2484
			(0.0056)	(0.0052)						
Bump	200	1.00	0.0792	0.0699	0.0672	0.0614	0.4404	0.4787	0.4203	0.4478
			(0.0195)	(0.0152)						

*Values in brackets are the median absolute deviation. *Continued on next page*

Table 5 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
Bump	200	2.00	0.2085	0.1720	0.0967	0.1038	0.7204	0.7687	0.6967	0.7873
			(0.0626)	(0.0535)						
Bump	400	0.25	0.0073	0.0071	0.0301	0.0300	0.0771	0.0769	0.0778	0.0756
			(0.0009)	(0.0009)						
Bump	400	0.50	0.0190	0.0188	0.0408	0.0390	0.1620	0.1629	0.1669	0.1679
			(0.0030)	(0.0028)						
Bump	400	1.00	0.0485	0.0461	0.0553	0.0509	0.3219	0.3313	0.3348	0.3404
			(0.0105)	(0.0089)						
Bump	400	2.00	0.1248	0.1119	0.0852	0.0788	0.6133	0.6383	0.5458	0.5712
			(0.0341)	(0.0286)						
Bump	800	0.25	0.0044	0.0043	0.0263	0.0263	0.0476	0.0461	0.0482	0.0471
			(0.0005)	(0.0005)						
Bump	800	0.50	0.0115	0.0114	0.0353	0.0343	0.1126	0.1120	0.1115	0.1114
			(0.0016)	(0.0016)						
Bump	800	1.00	0.0300	0.0300	0.0485	0.0451	0.2415	0.2447	0.2342	0.2395
			(0.0055)	(0.0058)						
Bump	800	2.00	0.0769	0.0710	0.0676	0.0614	0.4314	0.4516	0.4494	0.4724
			(0.0185)	(0.0154)						
Hardle	200	0.25	0.0239	0.0238	0.0168	0.0166	0.0522	0.0531	0.0547	0.0552
			(0.0027)	(0.0027)						
Hardle	200	0.50	0.0635	0.0635	0.0217	0.0215	0.1436	0.1436	0.1371	0.1368
			(0.0080)	(0.0081)						
Hardle	200	1.00	0.1601	0.1561	0.0290	0.0286	0.3086	0.3090	0.2959	0.3035
			(0.0238)	(0.0223)						
Hardle	200	2.00	0.3952	0.3550	0.0491	0.0458	0.5082	0.5260	0.5187	0.5261
			(0.0802)	(0.0631)						
Hardle	400	0.25	0.0152	0.0151	0.0147	0.0146	0.0374	0.0377	0.0363	0.0364
			(0.0015)	(0.0015)						
Hardle	400	0.50	0.0400	0.0399	0.0193	0.0192	0.0979	0.0973	0.0949	0.0950
			(0.0043)	(0.0043)						
Hardle	400	1.00	0.1021	0.1010	0.0253	0.0250	0.2147	0.2136	0.2296	0.2299
			(0.0133)	(0.0133)						
Hardle	400	2.00	0.2540	0.2454	0.0353	0.0341	0.4068	0.4059	0.3872	0.3940
			(0.0412)	(0.0354)						
Hardle	800	0.25	0.0090	0.0090	0.0128	0.0127	0.0255	0.0255	0.0251	0.0249
			(0.0007)	(0.0007)						
Hardle	800	0.50	0.0243	0.0241	0.0170	0.0169	0.0646	0.0638	0.0654	0.0644

*Values in brackets are the median absolute deviation. *Continued on next page*

Table 5 – *Continued from previous page*

DGP	n	σ	MSE_W	MSE_U	h^W	h^U	λ_1^W	λ_1^U	λ_2^W	λ_2^U
			(0.0023)	(0.0023)						
Hardle	800	1.00	0.0642	0.0635	0.0220	0.0218	0.1530	0.1522	0.1523	0.1514
			(0.0083)	(0.0081)						
Hardle	800	2.00	0.1595	0.1581	0.0287	0.0282	0.3205	0.3165	0.3040	0.3042
			(0.0216)	(0.0215)						

*Values in brackets are the median absolute deviation.

Table 6: Bandwidths from Nonparametric Regressions

Estimator	Age	Gender
WLC	6.1220	0.0005
LC	4.0602	0.0119

C Figures

Figure 1: Population DGPs considered for simulations

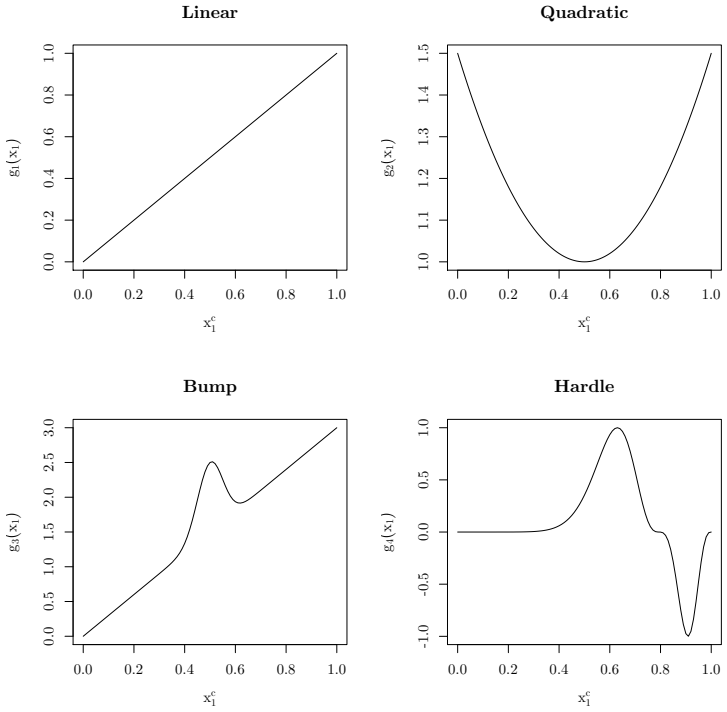
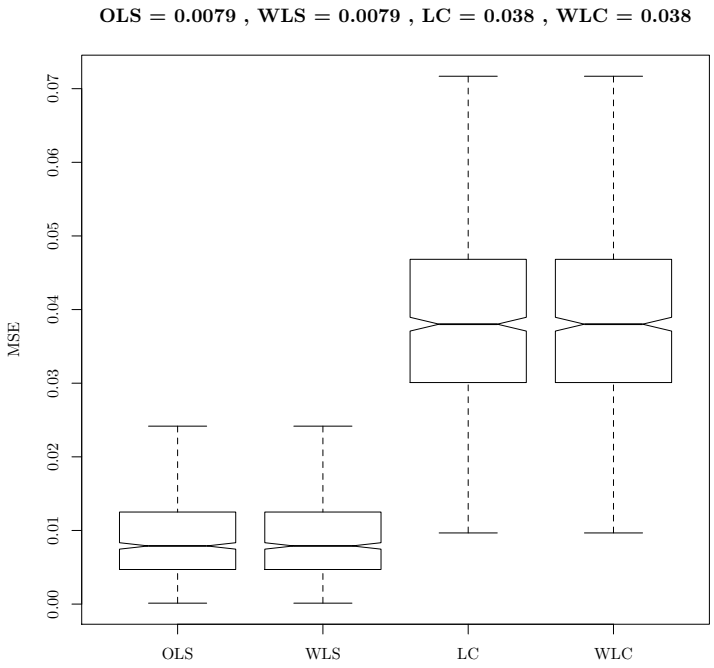
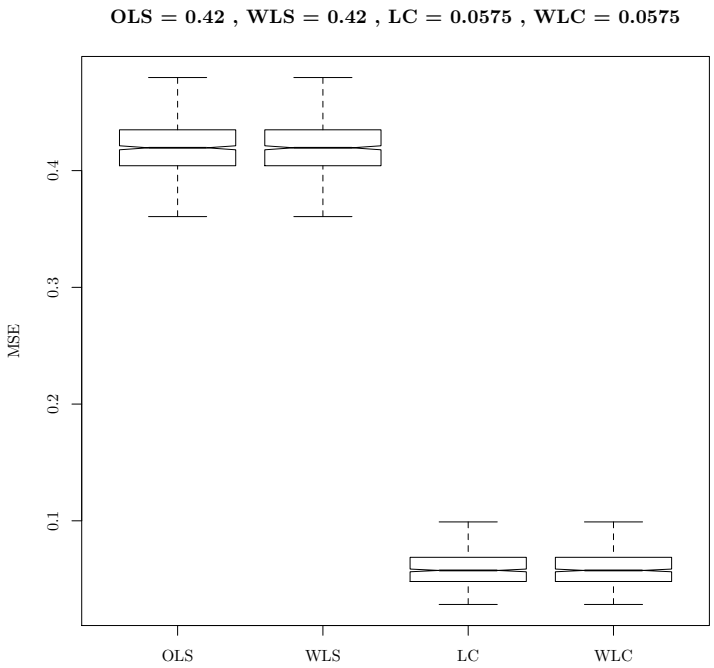


Figure 2: Boxplot of MSE for OLS, WLS, $\hat{g}(x)$, and $\tilde{g}(x)$ under SRS for $n = 400$ and $\sigma = 1.00$

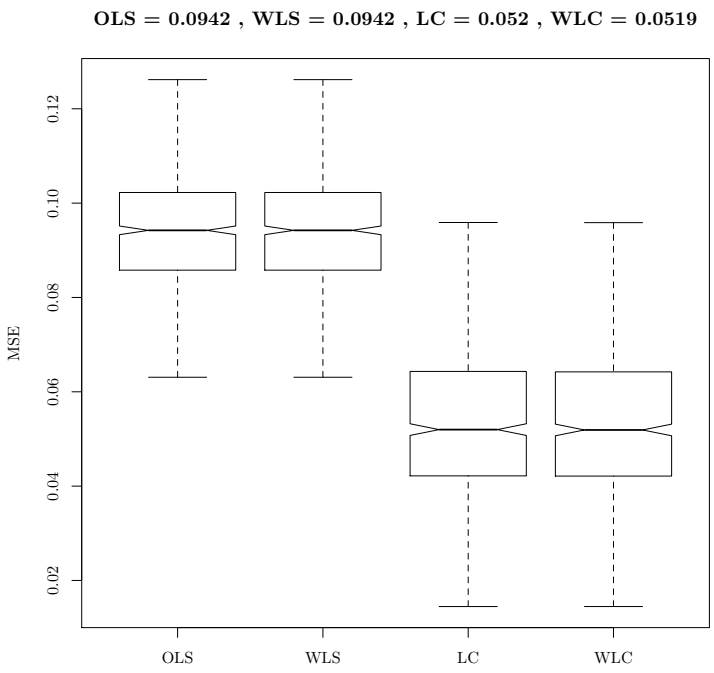
(a) Linear



(b) Quadratic



(c) Bump



(d) Härdle

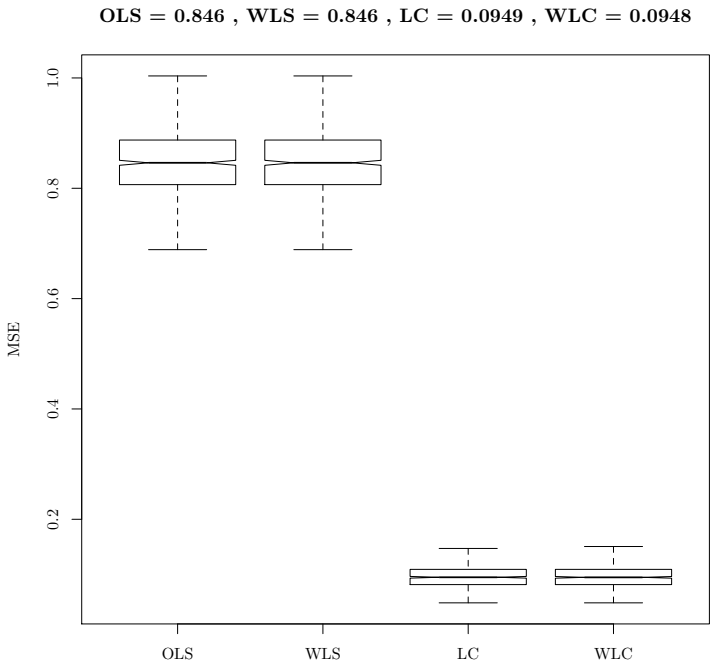
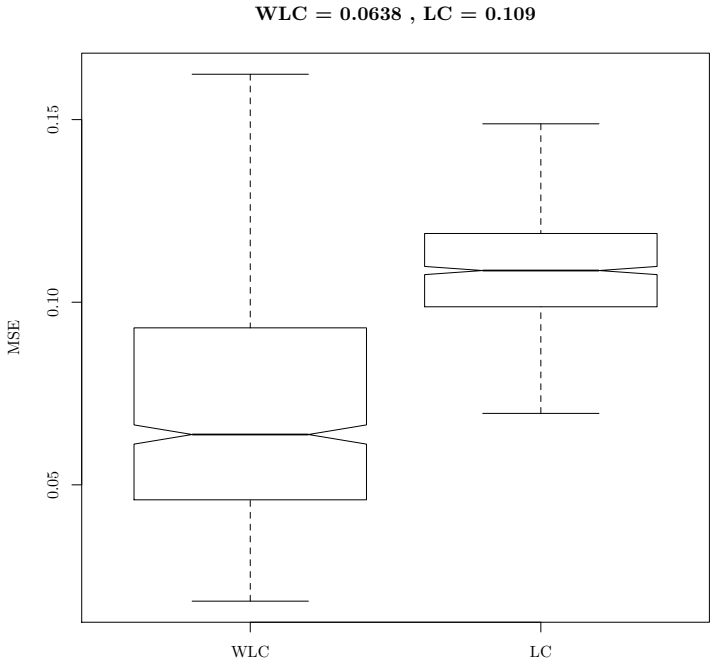
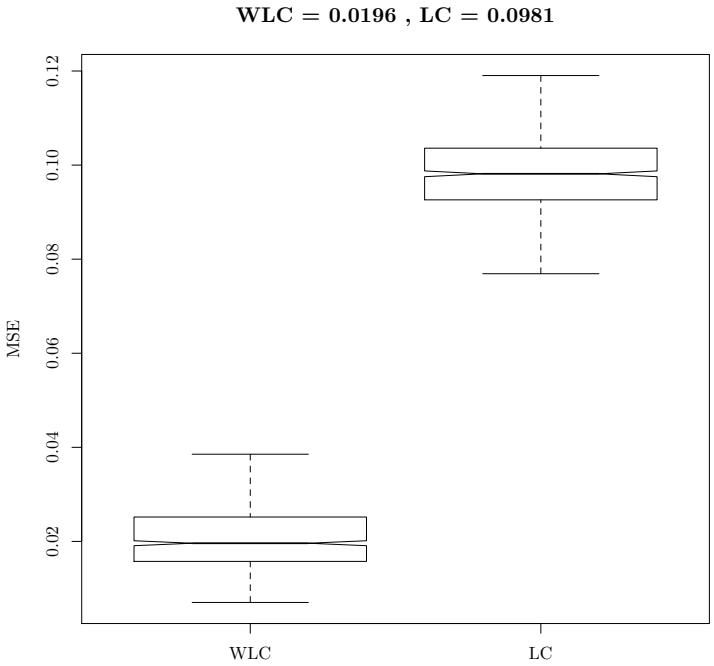


Figure 3: Boxplots for MSE($\hat{g}(x)$) and MSE($\tilde{g}(x)$) under endogenous Stratification, $\sigma = 0.50$

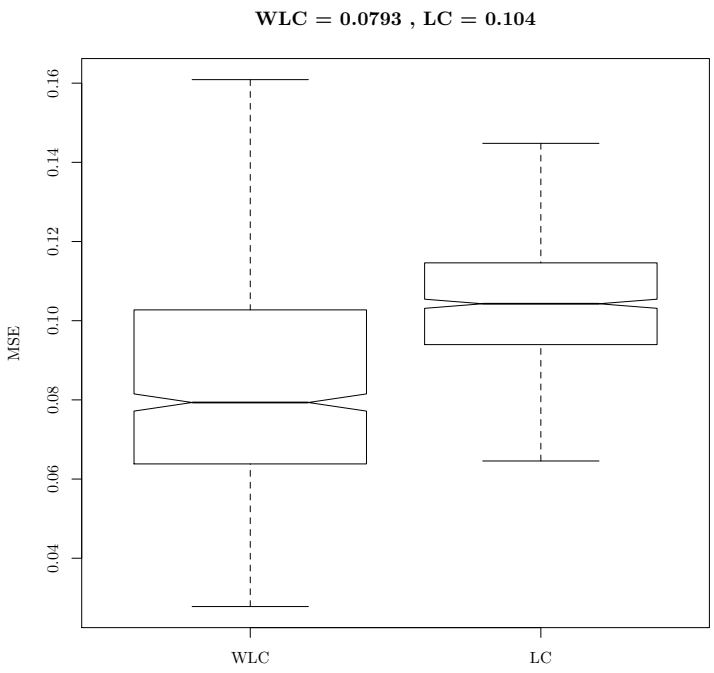
(a) Linear, n=200



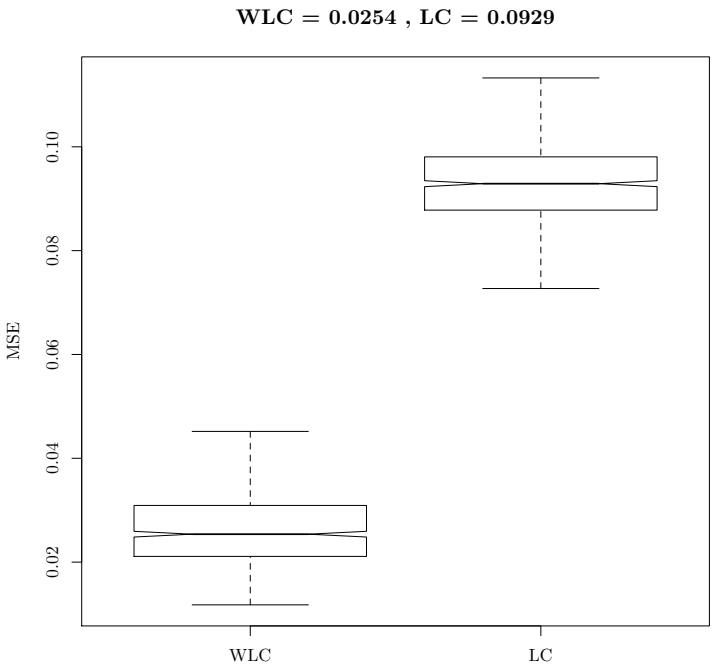
(b) Linear, $n = 800$



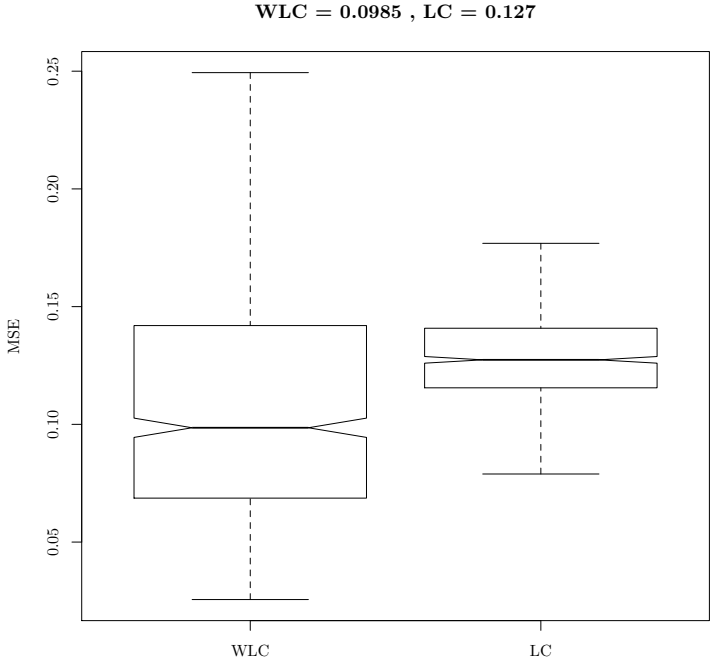
(c) Quadratic, $n = 200$



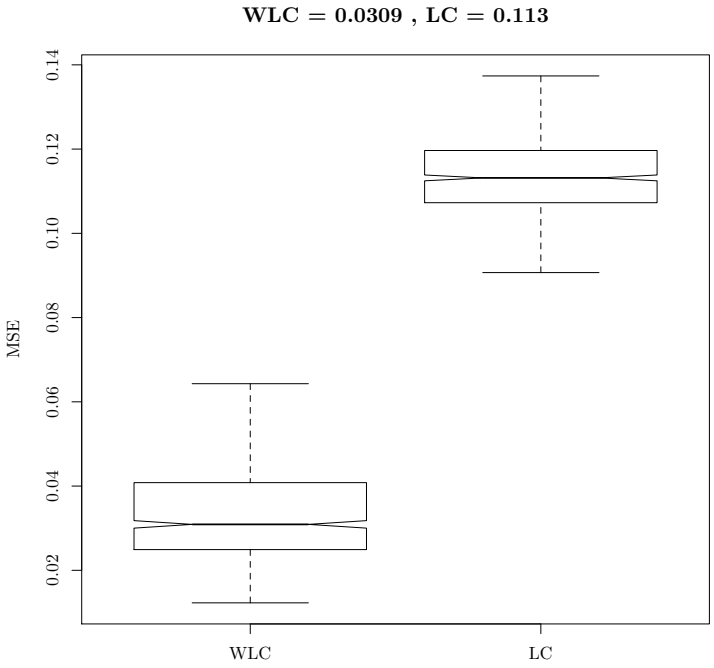
(d) Quadratic, $n = 800$



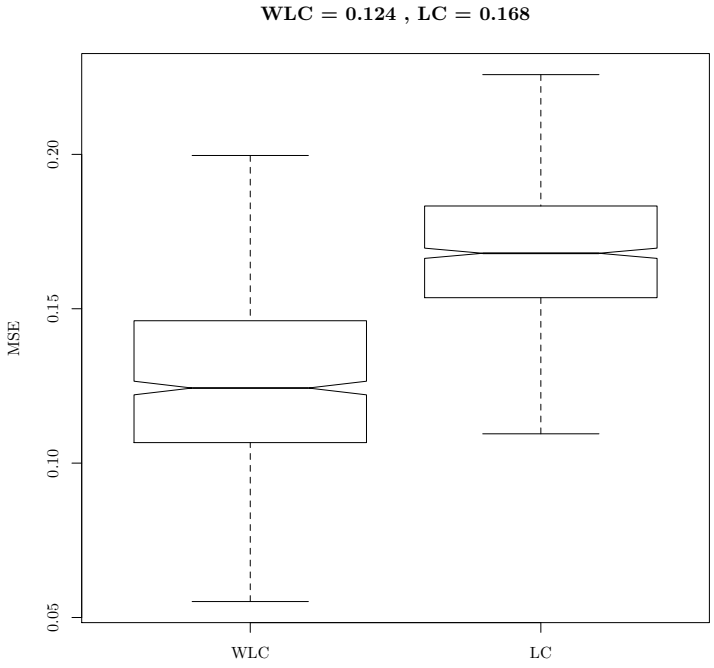
(e) Bump, $n = 200$



(f) Bump, $n = 800$



(g) Härdle, $n = 200$



(h) Härdle, $n = 800$

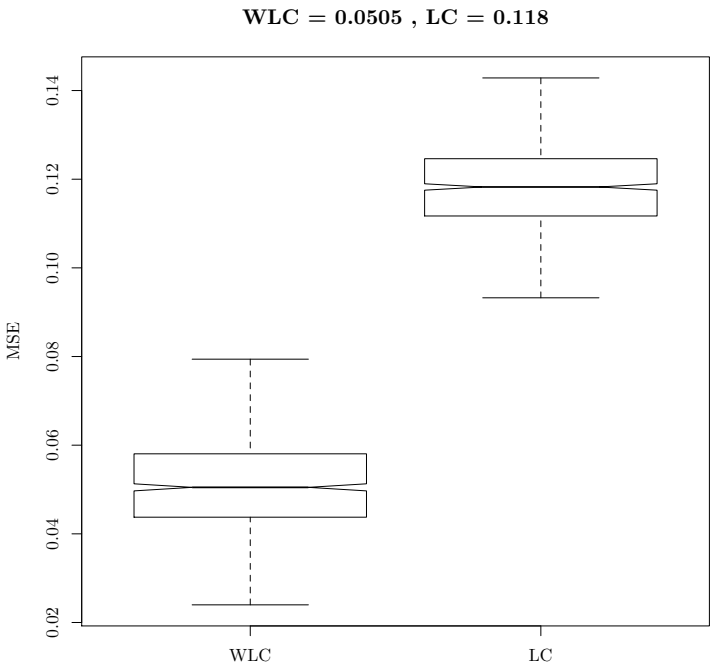
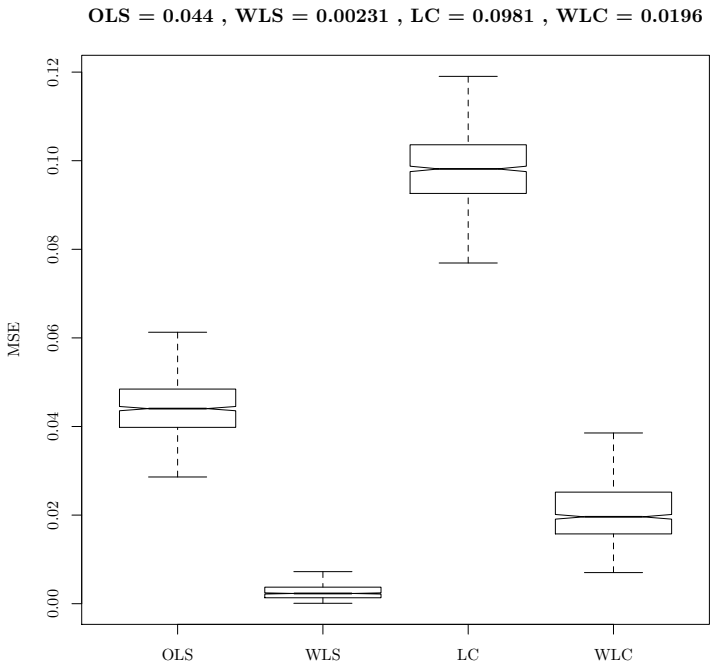
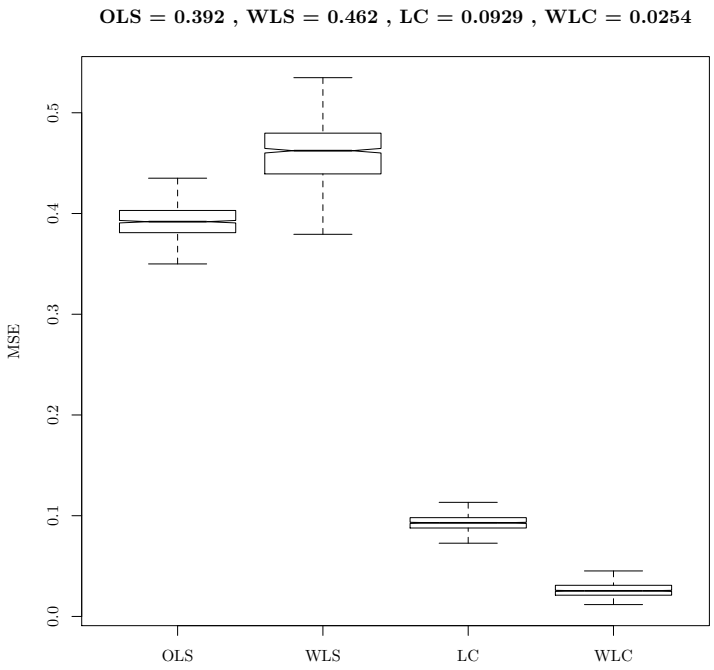


Figure 4: Boxplots for MSE of OLS, WLS, $\hat{g}(x)$, and $\tilde{g}(x)$ under endogenous Stratification, $n = 800$ and $\sigma = 0.50$

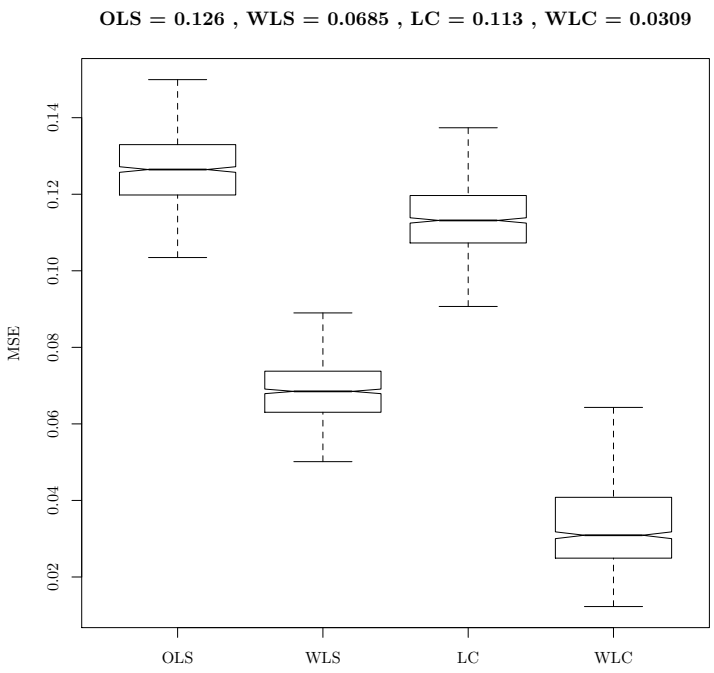
(a) Linear



(b) Quadratic



(c) Bump



(d) Härdle

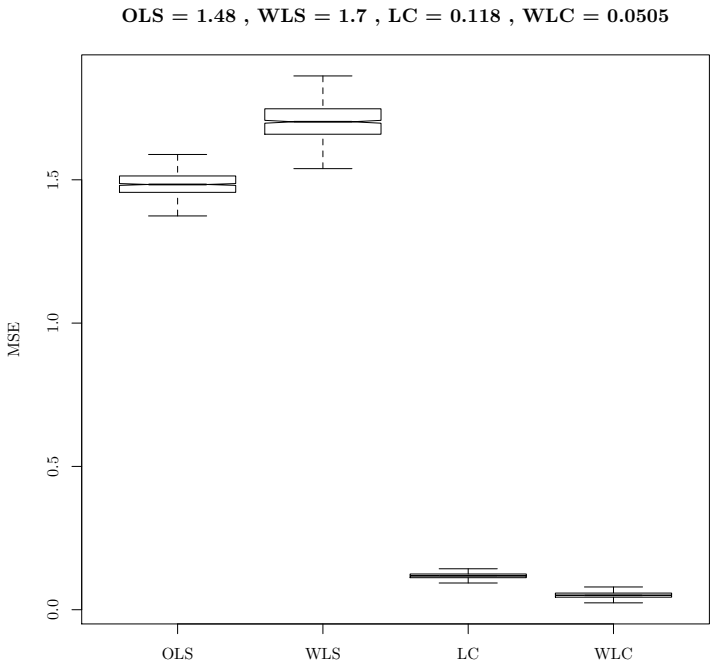
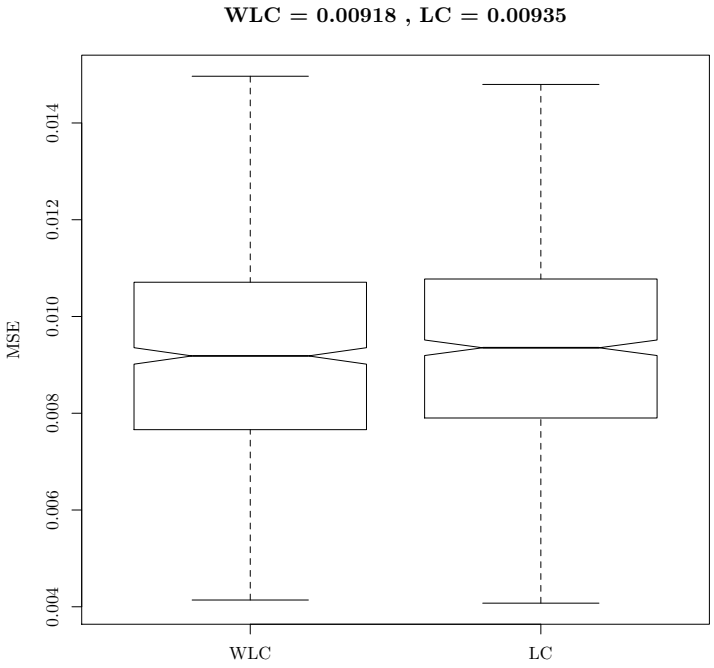
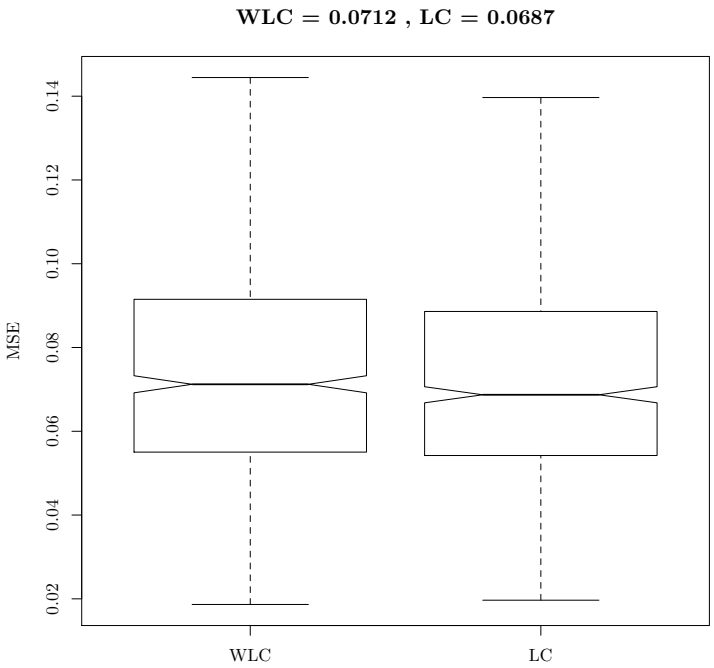


Figure 5: Boxplots for MSE($\hat{g}(x)$) and MSE($\tilde{g}(x)$) under exogenous Stratification for Linear DGP

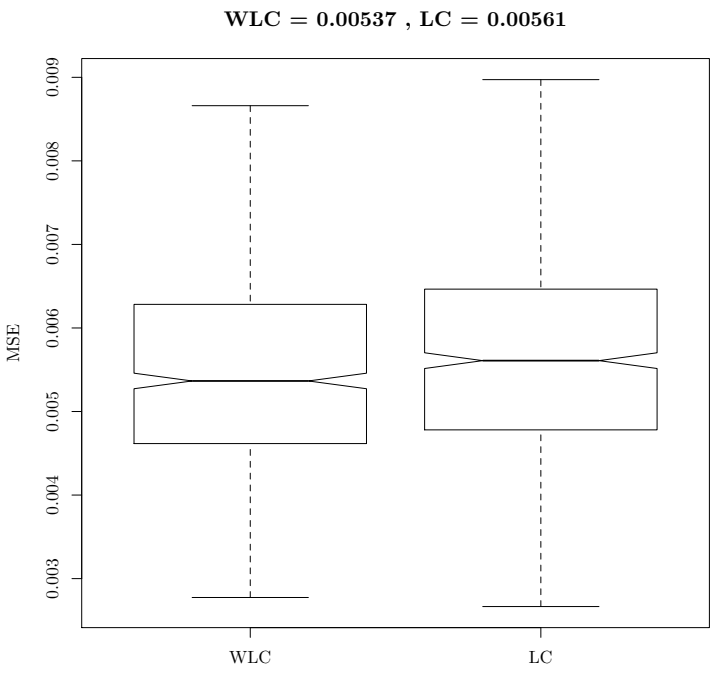
(a) $n = 200$ and $\sigma = 0.25$



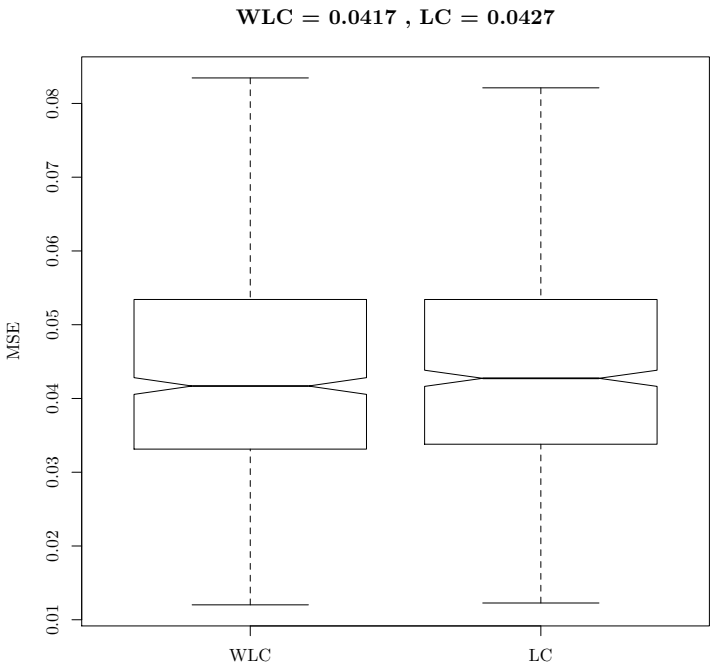
(b) $n = 200$ and $\sigma = 1.00$



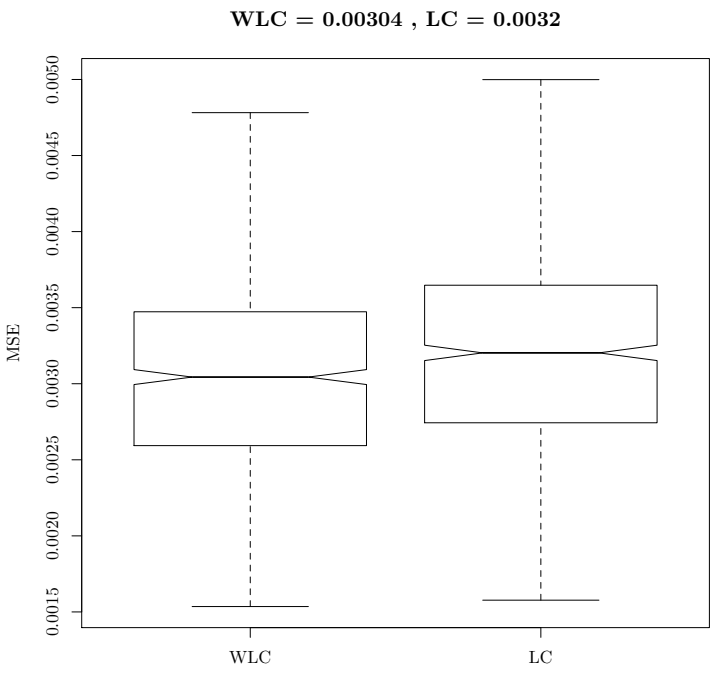
(c) $n = 400$ and $\sigma = 0.25$



(d) $n = 400$ and $\sigma = 1.00$



(e) $n = 800$ and $\sigma = 0.25$



(f) $n = 800$ and $\sigma = 1.00$

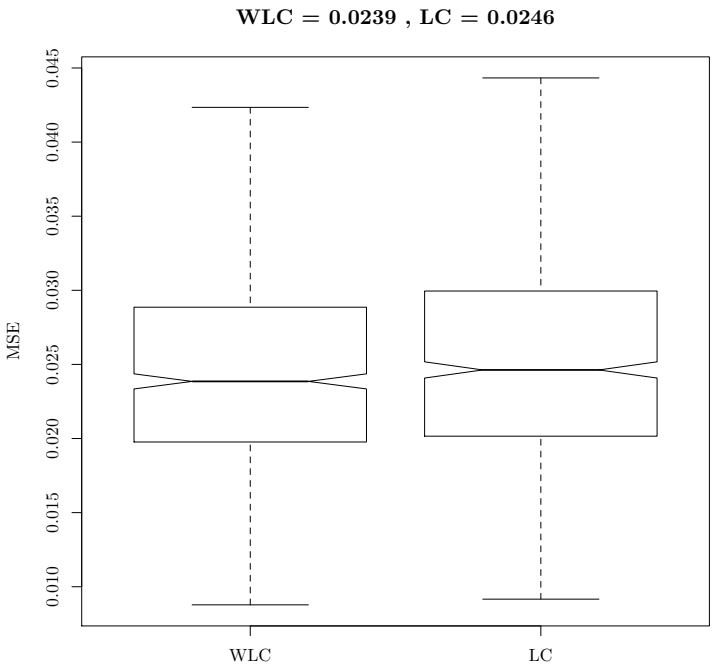
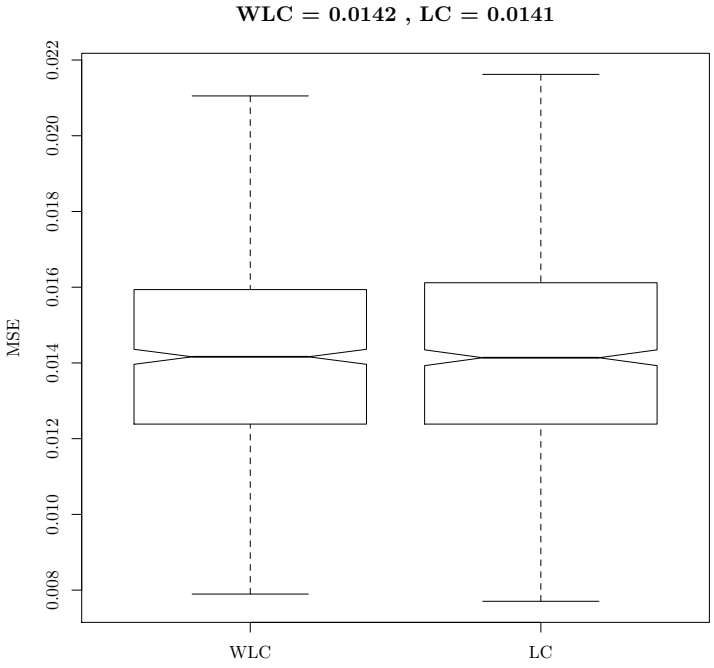
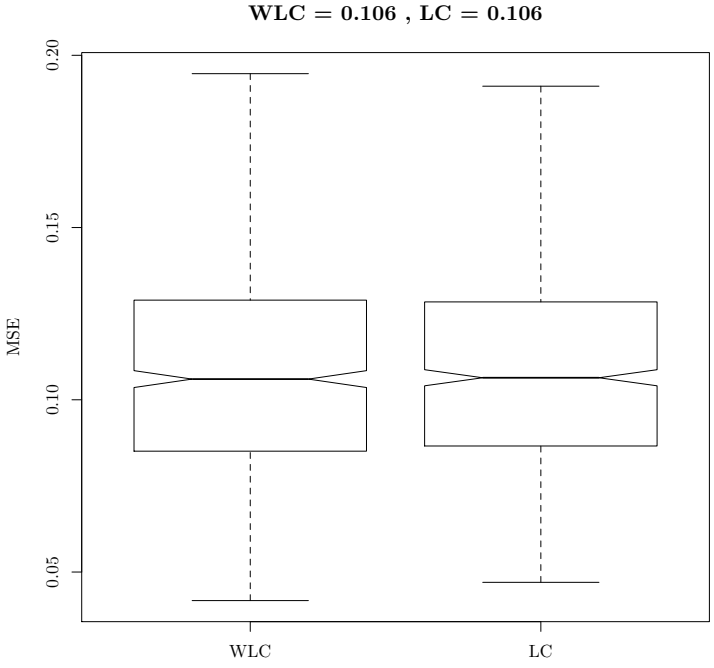


Figure 6: Boxplots for MSE($\hat{g}(x)$) and MSE($\tilde{g}(x)$) under exogenous Stratification for Quadratic DGP

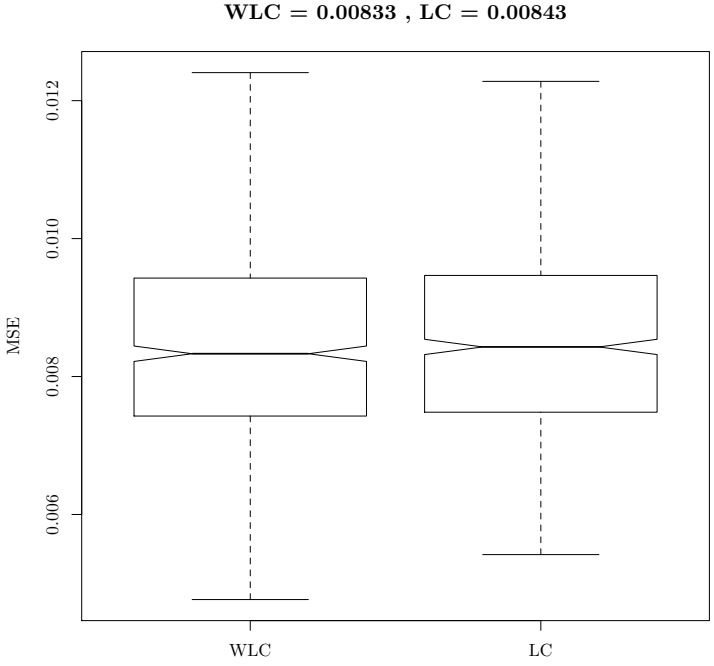
(a) $n = 200$ and $\sigma = 0.25$



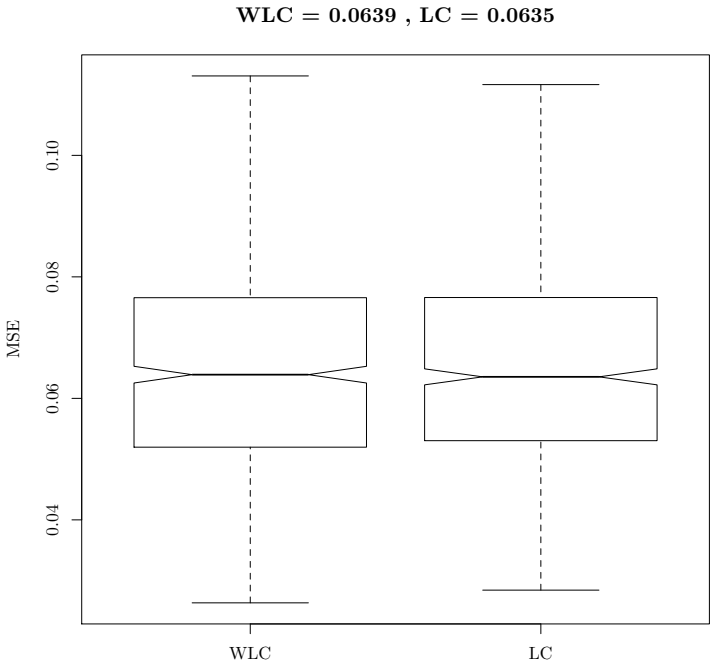
(b) $n = 200$ and $\sigma = 1.00$



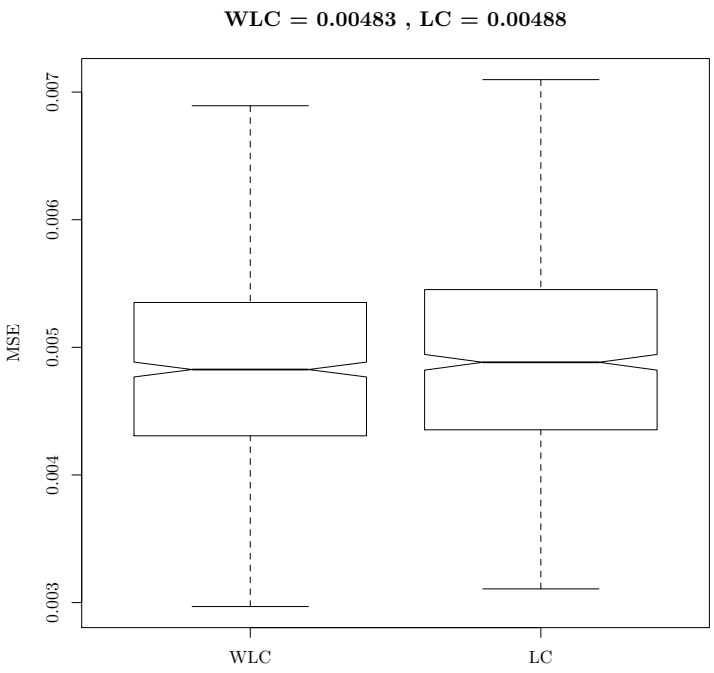
(c) $n = 400$ and $\sigma = 0.25$



(d) $n = 400$ and $\sigma = 1.00$



(e) $n = 800$ and $\sigma = 0.25$



(f) $n = 800$ and $\sigma = 1.00$

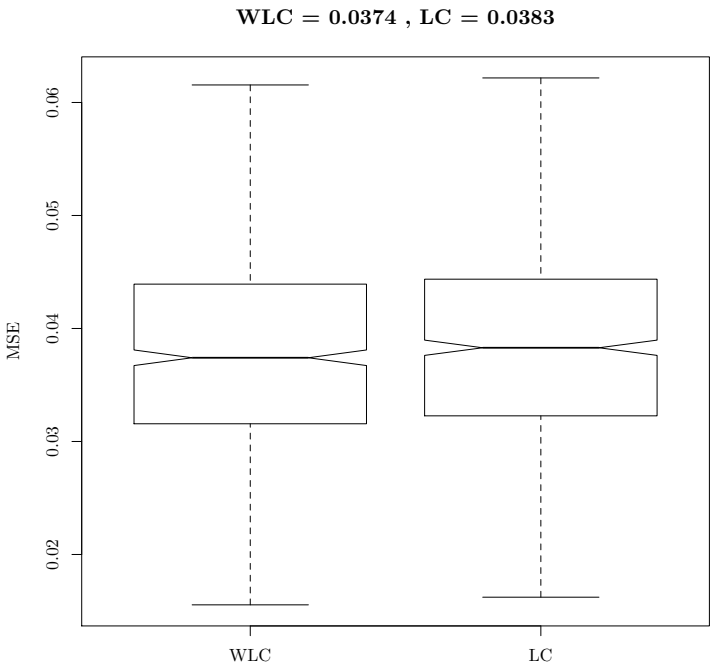
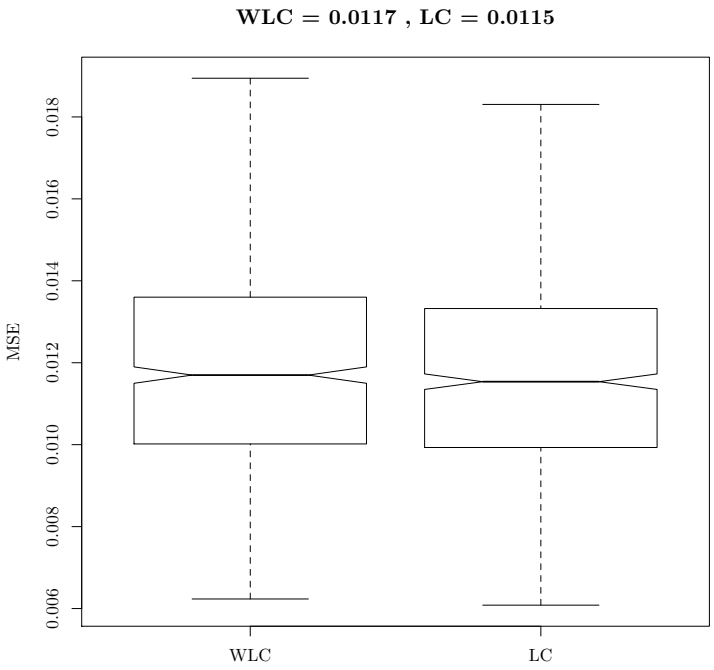
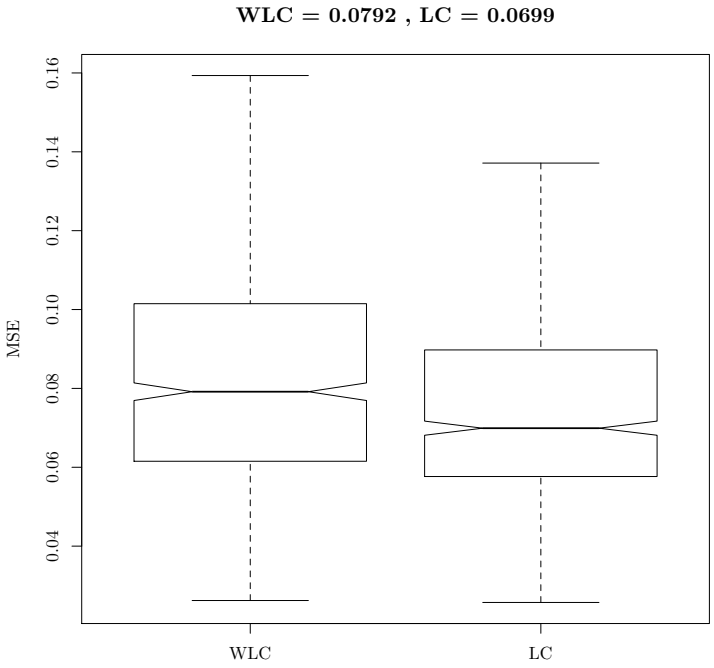


Figure 7: Boxplots for MSE($\hat{g}(x)$) and MSE($\tilde{g}(x)$) under exogenous Stratification for Bump DGP

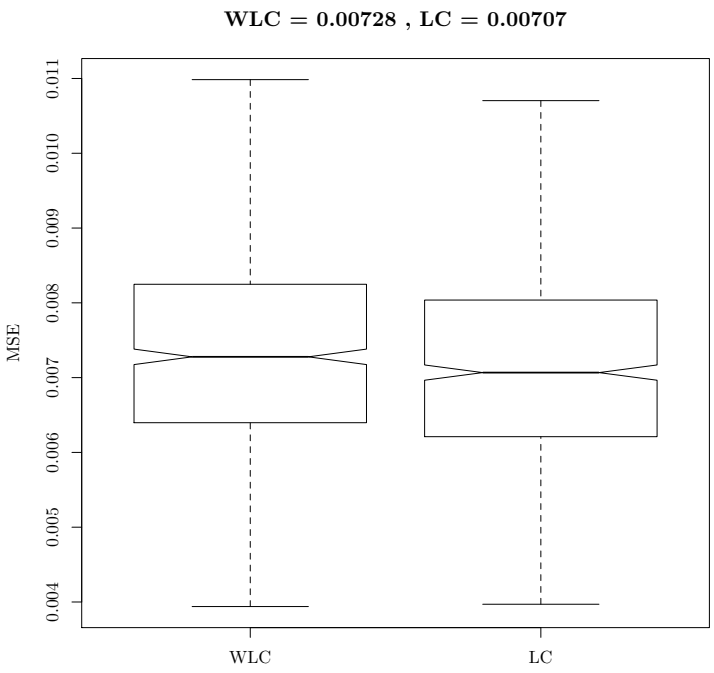
(a) $n = 200$ and $\sigma = 0.25$



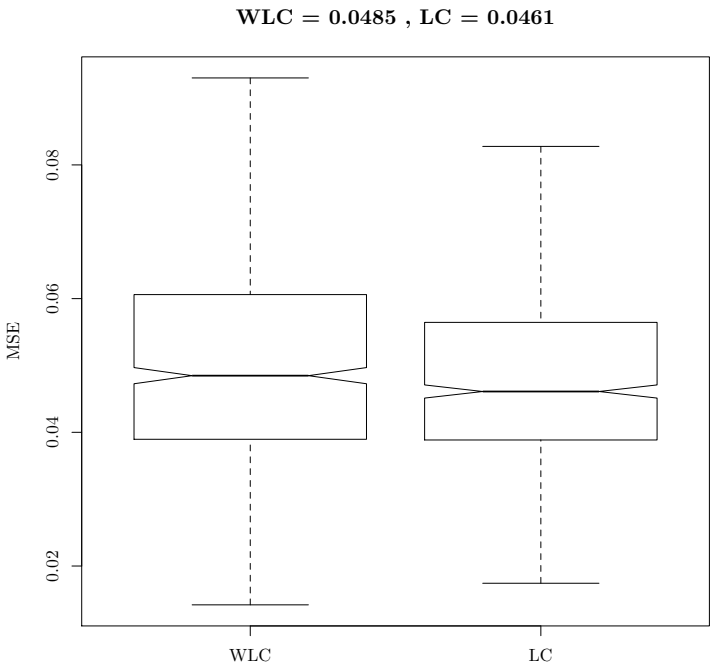
(b) $n = 200$ and $\sigma = 1.00$



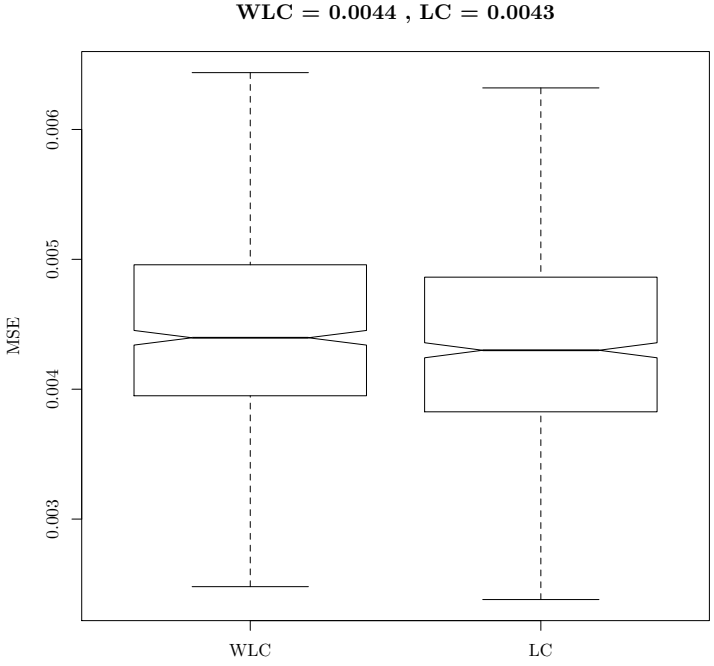
(c) $n = 400$ and $\sigma = 0.25$



(d) $n = 400$ and $\sigma = 1.00$



(e) $n = 800$ and $\sigma = 0.25$



(f) $n = 800$ and $\sigma = 1.00$

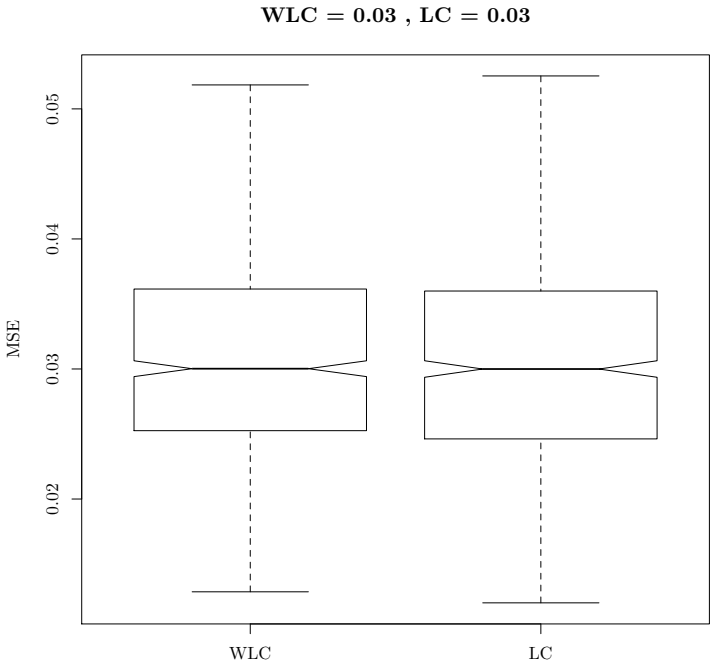
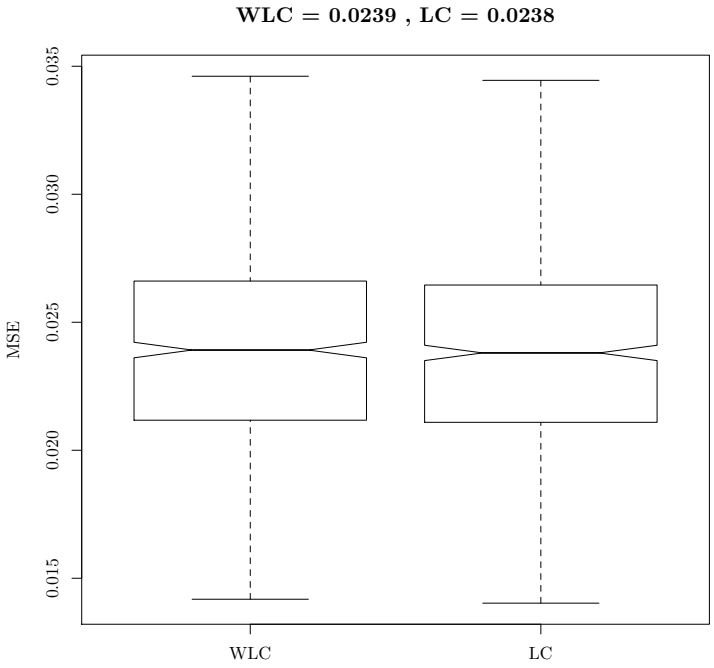
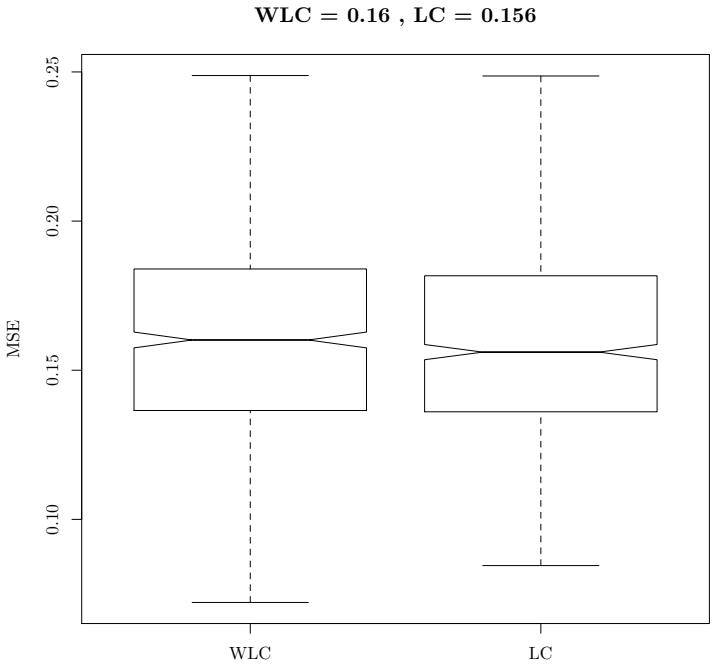


Figure 8: Boxplots for MSE($\hat{g}(x)$) and MSE($\tilde{g}(x)$) under exogenous Stratification for Härdle DGP

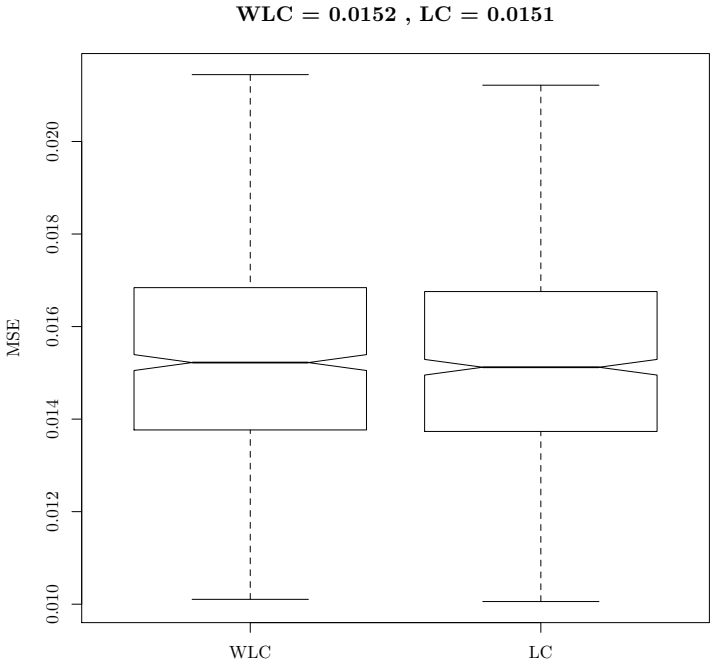
(a) $n = 200$ and $\sigma = 0.25$



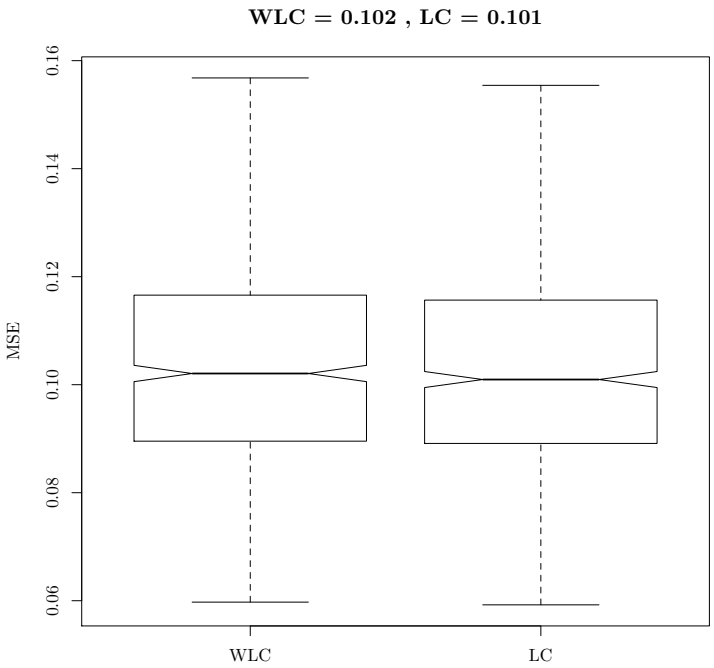
(b) $n = 400$ and $\sigma = 1.00$



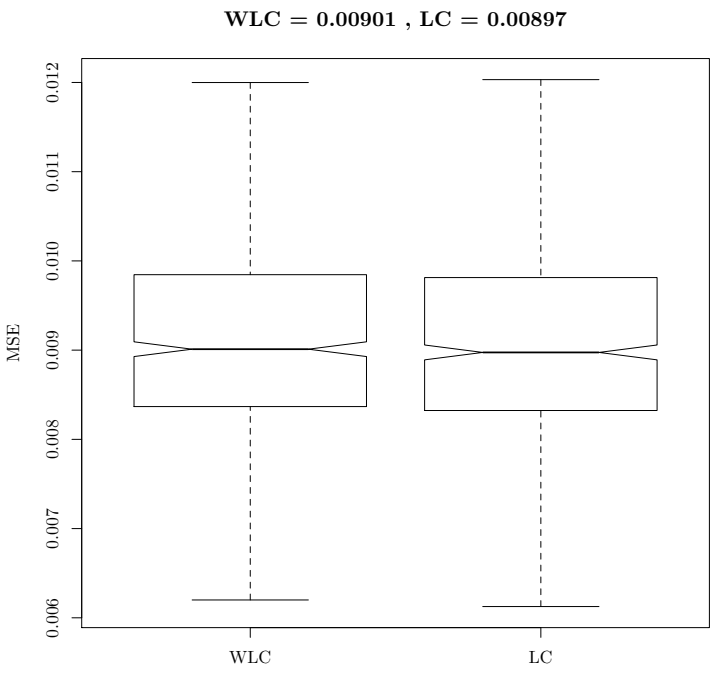
(c) $n = 400$ and $\sigma = 0.25$



(d) $n = 400$ and $\sigma = 1.00$



(e) $n = 800$ and $\sigma = 0.25$



(f) $n = 800$ and $\sigma = 1.00$

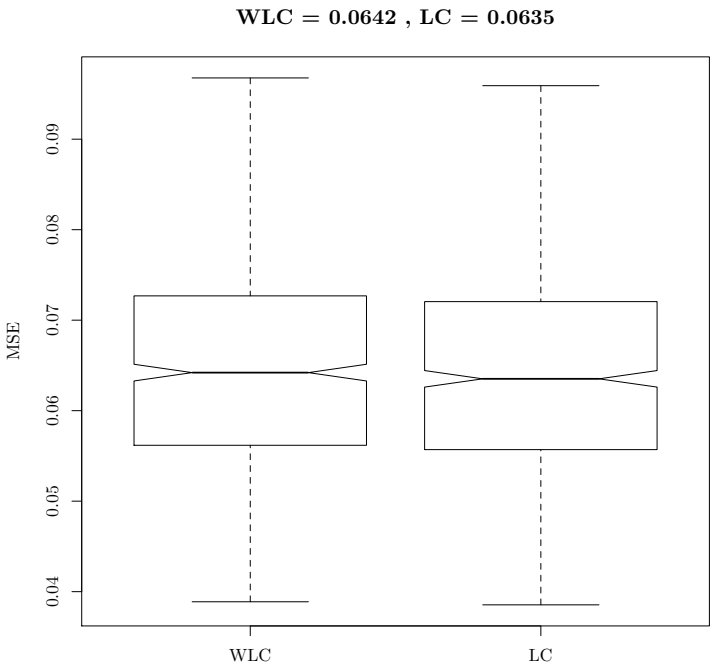
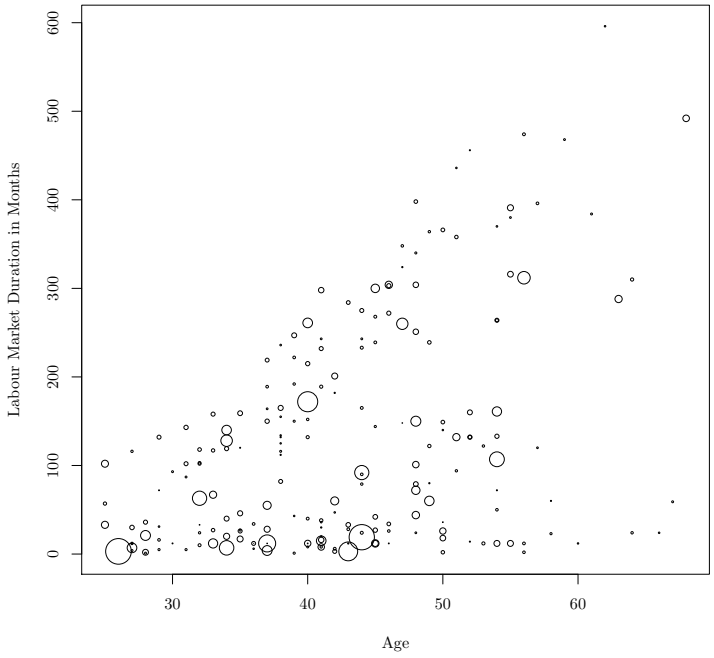


Figure 9: Application using SLID Data

(a) Labour Market Duration Versus Age



(b) Nonparametric Regression: Labour Market Duration versus Age