Three Essays on Econometric Analysis

Zhiyuan J. ZHENG

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Economics

Richard Ashley, Chair
Christopher F. Parmeter , Co-Chair
Pang Du
Suqin Ge
Hans Haller

May 2, 2011
Blacksburg, Virginia

Keywords: *LSCV Bandwidths, Bootstrap Significant Testing, Data Sharpening, Income Inequality, Health Dynamics*

Three Essays on Econometric Analysis

Zhiyuan J. ZHENG

(Abstract)

This dissertation consists of three essays on econometric analysis including both parametric and nonparametric econometrics. The first chapter outlines three topics involved and briefly discusses the motivations and methods, as well as some conclusions in each of the following chapters.

Both chapter 2 and chapter 3 are in the field of kernel smoothed nonparametric econometrics. Chapter 2 conducts large volumes of simulations to explore the properties of various methods proposed in the literature to detect irrelevant variables in a fully nonparametric regression framework. We focus our attention to two broadly sets of methods, the least square cross-validated bandwidth selection procedure and the conventional nonparametric significance testing frameworks.

In chapter 3, a bootstrap test statistic is proposed to test the validity of imposing some arbitrary restrictions on higher order derivatives of a regression function. We use data sharpening method to enforce the desired constraints on the shape of the conditional means and then measure the distance between the unrestricted and restricted models. The empirical distribution of the test statistic is generated by bootstrapping and the asymptotic distribution for the bootstrap test statistic is also provided.

The last chapter examines the relationship between population health and income inequality in China. We use a multilevel dynamic panel model to test the absolute income hypothesis, various versions of relative income hypothesis, and health selection hypothesis empirically.

Acknowledgments

First, I would like to acknowledge my dissertation committee: Christopher Parmeter, Richard Ashley, Hans Haller, Suqin Ge, and Pang Du. I thank my advisor, Christopher Parmeter, for his support and guidance. I acknowledge the contribution of my coauthor, Christopher Parmeter and Patrick Mccann, toward part of this dissertation. Chapter 2 is a joint work with them. I thank Byron Tsang for his comments on several occasions. I wish to thank Sherry Williams, Amy Stanford, Teresa Ritter, and Chris Hale for their help and assistance. I sincerely thank my fellow students, Hengheng Chen, Tiefeng Qian, Golnaz Taghvatalab, Jongwon Shin, Paul Hallmann, and Atanu Rakshit for their warm concern and encouragement. I must also express my gratitude to all ACSS (Association of Chinese Students and Scholars) members here at Virginia Tech.

Above all, I would like to thank my lovely family: my mother Huiying Zheng, and my aunts Lili Bai and Yunyun Bai. They always supported me with their love. Finally, I wish to thank my girlfriend, Shaojuan Liao who has been always by my side and has made me cheerful and comfortable so that I could devote myself to studying.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The breaking down of the curse of dimensionality is paramount in nonparametric econometric studies due to the fact that as the number of covariates increases, the theoretical properties and finite sample performances of a nonparametric estimator deteriorate rapidly, see Stone (1980). Throughout literature, different approaches have been proposed to detect redundant variables in order to reduce the dimensionality of the model under consideration.

The development of the asymptotic theory and simulation evidences documented in Hall, Li & Racine (2007) show that the data-driven least square cross-validation method can smooth away irrelevant regressors. The inclusion of irrelevant variables drives the associated bandwidths toward their theoretical upper bounds as the sample size becomes larger, rather than converging to zero in the classical analysis of nonparametric methods. In the local constant setting, a large bandwidth removes a variable from kernel regression, and this phenomenon holds for both continuous variables and discrete variables. However, in the case of local linear regression, this only applies to discrete variables. For continuous variables, a large bandwidth forces the associated continuous variable enter the model linearly, locally speaking, which contrastingly signify the relevance of this variable in the model. As a matter of fact, in any setting, such as local constant, local linear, and local polynomial, a discrete variable whose bandwidth hits its theoretical upper bound is deemed to be irrelevant.

However, bandwidth selection procedures can erroneously assign a large bandwidth to a relevant variable and associate a small bandwidth with an irrelevant variable. On the other hand, many conventional nonparametric significance tests proposed in the past, see Racine (1997), Lavergne & Vuong (2000), Racine, Hart & Li (2006), Gu, Li & Liu (2007), etc., provide formal testing framework upon which we can examine the degree of variables significance. They are widely accepted, empirically tested, and their performances are well appreciated. The size of cross-validated bandwidths can tell us the degrees of smoothness of each variable, nonetheless, they don't reveal a p value for the corresponding regressor. The pros come from the fact that cross-validated bandwidths not only minimize the least squares but also provide information about the relevance of variables at no cost. The cons are also obvious in a sense that they don't yield a degree of irrelevance or relevance. This can only be considered as *ad hoc*.

In Chapter 2 of this dissertation, we assess the accuracy of using cross-validated bandwidths as rule-of-thumb to determine the significance of variables and compare it with traditional nonparametric significance tests to find a sound strategy to reduce the dimensionality in various settings.

In chapter 3, we study the problem of imposing some arbitrary restrictions on a fully nonparametric model. Kernel smoothed nonparametric model has the appealing features of less restrictive assumptions on functional form, however, due to the nature of its building foundations, unlike its parametric counterpart, it has a great difficulty when a researcher needs to impose a conventional economic constraint, such as monotonicity, convexity, or any other restrictions applied to higher order derivatives, to the resulting conditional expectations. In this chapter, we propose a procedure which utilizes data sharpening technique to resolve this issue.

Data sharpening method, which was originally introduced by Choi & Hall (1999) and Choi, Hall & Rousson (2000), is employed to perturb the data so as to enhance properties of the relatively conventional statistical procedures. Braun & Hall (2001) suggested to extend the scope of data sharpening method to estimate a curve subject to some qualitative constraints, such as monotonicity. The method involves altering the positions of the data values as little as possible subject to a constraint. Neither the bootstrapping procedure in Braun & Hall (2001) discussed how to impose general restrictions on higher order derivatives nor can it formally test the validity of these restrictions. In this chapter, we resort to the local polynomial regression framework of Masry (1996*a*) and Masry (1996*b*) and then propose a distance-based test statistic which measure the change of conditional means before and after imposing the null hypothesis. We use bootstrap method to generate the empirical distribution of this test statistic and obtain the p value.

We also introduce the constrained weighted bootstrap method proposed by Racine, Parmeter & Du (2008), which alters the weight matrix to impose the null. Both data sharpening and constrained weighted bootstrap method can be viewed as alternative ways of ”data tuning”. By varying weights, one can shift the response variables up and down just enough to satisfy constraints under consideration. We compare both methods finite sample performances by looking at their impact on the shape of conditional means estimated from various $DGPs$. We also present Monte Carlo Simulation results and an empirical example.

The last chapter attempts to study the relationship between rising income inequality and population's health status in China, using China Health and Nutrition Survey(CHNS). To date, there have been many studies conducted to examine reasons and decompositions of the expanding inequality. However, the strong association between the public health status and income has not been fully investigated in terms of their directions of causations, let alone the role of income inequality.

The *absolute income hypothesis*, positing that a higher income level yields a better health outcome, has become a consensus view. However, the many debates over income inequality and health still leave us with an unresolved controversy regarding the so called *relative income hypothesis*, which states that income inequality is a destructive factor for individual health. The arguments focus on the role of income inequality. Mixed empirical results for different countries have been obtained. In this chapter, we use a multilevel model to break income inequality into individual and community levels and simultaneously examine their effects on health outcomes.

This chapter intends to answer the question that whether income inequality affects health outcomes independently when dual causations between income and health are already accounted. We use a dynamic panel framework to account for health selection process, then proceed to test various versions of the hypothesis between income inequality and health outcomes.

The relative income hypothesis has two versions. The strong version states that income inequality deteriorates health outcomes for both the rich and the poor; the weak version emphasizes the negative effect of income inequality on the poor's health outcomes, and it has little to do with the health status of the rich.

We find that the individual level inequality shows a persistent significant effect on health in our sample. The community level inequality measured by Gini coefficients proves to be irrelevant when health dynamics are correctly accounted. The results from CHNS data support the weak version of the relative income hypothesis; the strong version is rejected. Furthermore, the long run income level is more important than the short run income variation. The policy implication here is that reducing the individual level inequality and lifting the level of investments on public goods can improve the population's health.

Chapter 2

Cross-Validated Bandwidths and Significant Testing

(ABSTRACT)

The asymptotic theory relating to the size of a bandwidth that belongs to a variable in a kernel smoothed regression function is becoming well known. However, the comparison between the performances of conventional nonparametric significance testing and the size of bandwidths signaling whether the corresponding variables are *smoothed away* remains to be studied. In the first chapter of my dissertation, we try to answer the question that how reliable it is to use the procedure of cross-validated bandwidths to determine the degree of relevance for a variable or a group of variables that enter(s) the regression function. Our approach to solve this question involves setting up a variety of Monte Carlo simulation exercises with different scenarios. In each scenario, the priorities are different than the others, such as hypothesis focusing only on continuous variables, only on discrete variables, and finally a mix of discrete and continuous variables. In another word, we select tests that can handle both individual and joint data and compare their performances with respect to the performances the size of bandwidths. Our simulation results indicate the bandwidth

selection procedure works as good as conventional tests of significance, but to test a group of variable, it is better to resort to formal significance testing rather than joint bandwidths recognition. In terms of a sound strategy, we recommend that conventional significance testing is necessary to determine the final nonparametric model.

2.1 Introduction

The development of the asymptotic theory that data-driven least squares cross-validation method can remove irrelevant regressors through oversmoothing has been well recognized. The simulation studies documented in Hall et al. (2007) also reveal that this "automatic reduction" of the redundant (irrelevant) variables works well in finite samples. The inclusion of irrelevant variables drives the bandwidths associated with them go to their theoretical upper bounds as the sample size becomes larger, rather than converges to zero in the classical analysis of data-driven methods. In a local constant setting, a large bandwidth removes a variable from kernel regression, and it holds for both continuous variables and discrete variables. However, in the case of local linear regression, it is only true for discrete variables. For continuous variables, a large bandwidth forces the associated continuous variable enter the model linearly, which contrastingly signify the relevance of this variable. As a matter of fact, in any setting, local constant, local linear, local polynomial, a discrete variable whose bandwidth hits its upper bound is deemed irrelevant.

The benefits of bandwidth selection procedure is at no cost if least squares cross-validated method was used to obtain the optimal bandwidths given the data. However, the cons are that it can erroneously assign a large bandwidth to a relevant variable or give a small bandwidth to an irrelevant variable. On the other hand, many conventional nonparametric significance testings proposed in the past literature, see Racine (1997), Lavergne & Vuong

(2000), Racine et al. (2006), Gu et al. (2007), etc., provide formal testing framework upon which we can examine the degree of variables' significant levels. They are widely accepted and their performances are well known. Bandwidth selection can tell us the degrees of smoothness associated with each variable, nonetheless, the size of the bandwidth doesn't reveal a p value for the corresponding regressor. In this chapter, we assess the accuracy of cross-validated bandwidth selection procedure in terms of determining the significance of variables by merely looking at the size of the bandwidth. We also provide the results from traditional testing frameworks to see which one is more reliable, both in individual and joint settings.

At this point of discussion, it is worth to point out that the results of nonparametric significance testing, sometimes, are influenced by the choice of bandwidths used to perform the test, especially for the power, see Gu et al. (2007). In most applied work, rules-of-thumb and *ad hoc* bandwidth procedures are overwhelmingly practiced in reality, even though there is no reason to do so. The ideal nonparametric significance tests should not be affected by the choice of different bandwidth selection methods. The robustness of the testing results with respect to the choice of bandwidths is a necessitate for a research to make sound judgement and convincing inferences. On the other hand, with the appealing features of cross-validated bandwidth, we can pre-determine the suspected variables those are subject to scrutinization. In a sense, bandwidth selection procedure can be used as complementary information for conventional testing frameworks, rather than being viewed as competitors.

This chapter of my dissertation attempts to provide empirical guidance about how researchers should practice in nonparametric modelling when certain set of the regressors are seemed to be irrelevant. This is particularly important in applied nonparametric settings because nonparametric models suffer from the curse of dimensionality. By correctly detecting the irrelevant variables, lower dimensional model can be produced and therefore alleviating the curse of dimensionality. We examine a number of different but inter-related approaches to deciding which estimation method to use. We suggest that interpreting the data-driven bandwidths first and testing the significance of over-smoothed variables subsequently. The simulations in this chapter focus on least squares cross-validated (LSCV), see Hall et al. (2007), and the wild-bootstrap test of Gu et al. (2007), which we extend to include discrete variables though their paper only considered continuous case. The underlying $DGPs$ considered here admit both continuous and discrete variables, and some of them including a high number of irrelevant variables. Also the null hypothesis presented in this chapter consider both individual and joint tests of significance for continuous, discrete and mixed type of regressors. The bandwidths are selected by local constant kernel methods via LSCV method, which has the ability of detect irrelevant variables with high accuracy in the case of individual variables, see Hall et al. (2007).

The rest of the first chapter proceeds as follows. We first introduce local constant nonparametric estimation method, then discuss the LSCV bandwidth selection procedure. We also outline the two nonparametric tests that can determine the both individual and joint significance. 2.3 presents our simulation setup and provides evidence for our comparison. Section 2.4 summarizes some concluding remarks.

2.2 Nonparametric Local Constant Estimation

The model under consideration is in the most general form:

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where y_i is a scalar, and x_i is a $d \times 1$ vector of covariates and ε_i represents the random component of the model. The initial incentive here is to identify all the irrelevant variables so that we can reduce the dimension of our model. We would like to impose null hypothesis that individual (joint) variable(s) are redundant. The variables under consideration can be both continuous and discrete ones. We adopt kernel smoothing methods that was proposed by Racine & Li (2004), the generalized product kernels.

The name of local constant estimation comes from the fact that the conditional mean at each point of interest is actually a weighted sum of the dependent variable, and the weight is determined by kernel function and the bandwidth associated with the corresponding variables. Contrasting to the cases of local linear, local higher polynomials where the conditional mean are estimated through linear and higher order polynomials of the explanatory variables, we fit in a constant into the window-width, measured by the bandwidth. First, let us introduce the local constant estimation framework, see Nadaraya (1965) and Watson (1964) for more details. The conditional mean can be expressed as follows

$$\hat{m}(x) = \frac{\sum\limits_{i=1}^{n} y_i K_h(x, x_i)}{\sum\limits_{i=1}^{n} K_h(x, x_i)} = \sum_{i=1}^{n} A_i(x) y_i, \tag{2.2}$$

where

$$K_h(x, x_i) = \prod_{s=1}^{q} h_s^{-1} l^c \left(\frac{x_s^c - x_{si}^c}{h_s} \right) \prod_{s=1}^{r} l^u \left(x_s^u, x_{si}^u, \lambda_s^u \right) \prod_{s=1}^{p} l^o \left(x_s^o, x_{si}^o, \lambda_s^o \right), \tag{2.3}$$

where $K_h(x, x_i)$ is the product kernel admitting both continuous and discrete variables, see Pagan & Ullah (1999). x_s^c represents continuous variables, x_s^o represents ordered discrete variables and x_s^o represents unordered discrete variables. Notice that $q + r + p = d$. The conventional method to deal with discrete variables are to split samples according to categorical variables. However, Racine & Li (2004) pointed out that there could be large losses in efficiency. For continuous variables, we use the standard normal distribution as the kernel function to assign weights to data points according to their relative distance which is measured by the bandwidth h_s. For the smoothness of discrete variables, Aitchison & Aitken (1976) designed a kernel function for unordered categorical data and Wang & Ryzin (1981) considered kernel function for ordered categorical variables. The smoothing mechanisms adopted in this generalized product kernels are very similar to those of Aitchison & Aitken (1976) and Wang & Ryzin (1981).

For the unordered discrete variables, $0 \leq s \leq r$, define the kernel function as

$$l^u\left(x_s^u, x_{si}^u, \lambda_s^u\right) = \begin{cases} 1 & \text{if } x_s^u = x_{si}^u, \\ \lambda_s^u & \text{if } x_s^u \neq x_{si}^u, \end{cases} \tag{2.4}$$

where $0 \leq \lambda_s^u \leq 1$ is the smoothing parameter for x_s^u. Next we display the kernel function for the ordered discrete variables,

$$l^o\left(x_s^o, x_{si}^o, \lambda_s^o\right) = \begin{cases} 1 & \text{if } x_s^o = x_{si}^o, \\ \lambda_s^o & \text{if } x_s^o \neq x_{si}^o, \end{cases} \tag{2.5}$$

where $\lambda_s^o = \lambda_s^{|x_s^o - x_{si}^o|}$. Finally, we let $\lambda = (\lambda_1^u, ..., \lambda_r^u, \lambda_1^o, ..., \lambda_p^o)$.

From the definitions of the kernel function for discrete variables above, it is straight forward to see the connection between the size of the bandwidths and their relevance in the regression function. As λ approaches to 0, both ordered and unordered discrete variables become indicators and split the data sample into grids where within each grid, we have $x_i^o \equiv x_j^o$ and $x_i^u \equiv x_j^u$. This is the traditional frequency estimator since no weights are assigned to points with different values of discrete variables. On the other hand, if λ come near at 1, the upper bound, different values of discrete variables have no impact on the resulting nonparametric regression. The variable is completely smoothed out. Therefore, a large bandwidth for a discrete variable always indicates less relevance for the nonparametric regression.

2.2.1 Cross-Validated Bandwidth Selection

The choice of bandwidth is at the heart of nonparametric estimation. It determines the relative distance between data points. A small bandwidth requires data points clustered in order to get a smoothed estimate, otherwise the fitted values will experience high variance and bounce up and down like wiggling. A large bandwidth is good for scattered data set since it can include enough points to make estimation. But an exceedingly large bandwidth might take too much information under consideration, resulting in a over-smoothed estimate and leading to high bias. On the extreme case, it becomes a flattened line. Typically for the same data sample, there is a trade-off between variance and bias as bandwidth grows: variance shrinks and bias rises. The least square cross-validation (LSCV) chooses a set of bandwidths that minimize the sum of total variances and total squared biases, see Hall et al. (2007) (HLR hereafter), and it has the ability to automatically smooth out irrelevant variables by assigning bandwidths approaching the theoretical upper bounds.

Following Hall et al. (2007), the bandwidths are chosen to minimize

$$CV(h, \lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2\omega(x_i), \tag{2.6}$$

where $\hat{m}_{-i}(x_i) = \sum_{j\neq i}^{n} K_{h,\lambda}(x_i, x_j)y_i / \sum_{j\neq i}^{n} K_{h,\lambda}(x_i, x_j)$ is the leave-one-out estimator and $0 \leq \omega(\cdot) \leq 1$, which serves as a weight function to avoid difficulties caused by dividing by 0. In the simplest case, we can set $\omega(x_i)$ to be a constant. Recall equation (2.1), we can rewrite equation (2.6) in the following manner

$$
\begin{aligned}
CV(h,\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(m(x_i) + \varepsilon_i - \hat{m}_{-i}(x_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_{-i}(x_i)\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(m(x_i) - \hat{m}_{-i}(x_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.
\end{aligned}
$$

From the above equation, we see that the first term is the sum of biases and the third term represents the total variance of our estimation. The middle term has a expectation of zero as $n \to \infty$. The LSCV chooses the smoothing parameter (h, λ) to minimize the sum of biases and variances.

For a variable which is indeed relevant in the modelling of the relationship between response variable and explanatory variables, the bandwidth adhered with this variable will play a role in the determination of the final value of $CV(h, \lambda)$. The fact that as n goes to infinity, the bandwidth usually converges to 0. That is because as the density of data points rises, nonparametric estimator no longer need to include points away from the point of interest to make accurate estimate. The surrounding observations are already providing enough information to make a point estimation. However, in the case of irrelevant variables, this phenomenon doesn't hold true anymore. The presence of a redundant variable should not affect the shape of the conditional mean as n goes to infinity. This is intuitively straightforward because the response variable are completely independent of this variable. Asymptotically, irrelevant variables should be neglected and the value of the objective function $CV(h, \lambda)$ are determined by the bandwidths of those relevant variables. Instead of shrinking to 0 as n becomes larger, the bandwidths of irrelevant variables actually go to infinity for continuous variables and 1 for discrete variables in the local constant estimation setting. They act like they were not there at the first place. The LSCV has this automatic mechanism of removing irrelevant variables. Once this is achieved, the curse of dimensionality is reduced by the action of oversmoothing redundant variables. This is another advantage of using nonparametric methods, besides its robustness of functional form specification.

Despite the appealing features of LSCV method discussed above, there are potential minefield which researchers can misuse the information provided by the size of bandwidths. First, the asymptotic analysis of the size of irrelevant variables' bandwidths does not produce a clear principle upon which we can state the degree of irrelevance for a specific variables.[1] Compared with formal testing frameworks, LSCV bandwidths don't render a p value to assess the confidence of irrelevance. Second, in most applied settings, researchers face the problem of large number of independent variables. It is unclear whether the LSCV method works well under large sets of regreesors. Third, the theoretical results of HLR requires regressors to be independent of each other, even though their simulations results indicate that correlations among regressors are not a severe damaging factor. But still their simulations only consider two variables which is rarely encountered in applied work.

Though the HLR provides a handy tool to examine the individual significance, what we find out in this chapter is that as the number of irrelevant variables goes up, LSCV bandwidths are a poor marker for assessing the joint significance of a group of variables, which is frequently asked in applied settings. Therefore, in the next section we introduce some conventional nonparametric significant tests.

[1] HLR method uses the rule-of-thumb of 2 times standard deviations as a deciding criterion.

2.2.2 Existing Nonparametric Significant Tests

As we discussed so far, LSCV bandwidths selection procedure doesn't provide a p value for the significance of a variable, nor does it yield an efficient framework to determine the joint significance of a group of variables. The need for formal tests are paramount. Also the test of significance is the most frequently encountered tests in applied economic analysis. They serve as evidence either to confirm or refute economic theories. Nonparametric significant tests don't rely on the correctly functional specifications and therefore are consistent under less restrictive assumptions.

First of all, let us outline an approach based on which we can test hypothesis in a fully nonparametric setting. Let $X \in R^d$ consists of both continuous and discrete variables, with q continuous variables, $r + p$ discrete variables. We partition X into two groups, (W, Z), where Z denotes the variables subject to hypothesis testing. The null hypothesis that the conditional mean of Y does not depend on Z can be expressed as following,

$$H_0 : E(Y|w, z) = E(Y|w) \quad a.e. \tag{2.7}$$

Z in our setup admits both continuous and discrete variables, or a mixture of them. However, w must contain at least one continuous variables. Define $u = Y - E(Y|W)$, then if the null hypothesis is true, we have $E(u|X) = 0$ a.e.. The test statistic constructed is based on $E\{uf_w(W)E[uf_w(W)|X]f(X)\}$, where $f_w(\cdot)$ and $f(\cdot)$ are probability density functions of W and X, respectively.

Consider the following nonparametric regression model of the form

$$y_i = m\left(w_i, z_i\right) + u_i. \tag{2.8}$$

Let $\hat{f}_{w_i}$ and $f(\cdot)$ denote the leave-one-out kernel estimators of $f_w(w_i)$ and $E(y_i|w_i)$, respectively, i.e.,

$$\hat{f}_{w_i} = \frac{1}{n-1}\sum_{j\neq i}^{n} K_{h_w}(w_i, w_j),$$
$$\hat{Y}_i = \frac{1}{n-1}\sum_{j\neq i}^{n} K_{h_w}(w_i, w_j)y_j/\hat{f}_{w_i}.$$

Next we continue by discussing cases of Z being all continuous variables, discrete variables and a mix of both of them.

All Continuous Case

For the moment, let z denote the continuous variables that might be redundant. Given the setup above, a feasible test statistic can be written as

$$\widehat{I_n^c} = \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(y_i - \widehat{y_i}\right)\widehat{f_w}\left(w_i\right)\left(y_j - \widehat{y_j}\right)\widehat{f_w}\left(w_j\right)K_{h_x}\left(x_i, x_j\right), \tag{2.9}$$

where $K_{h_x}(\cdot)$ is the product kernel introduced by Racine & Li (2004).

Under the conditions listed in Li & Racine (2007) (Page 372) and H_0, we have

$$T_n^c = \left(nh_1h_2\cdots h_q\right)^{1/2}\widehat{I_n^c}/\widehat{\sigma}_n^c \rightarrow N\left(0,1\right) \tag{2.10}$$

where $\widehat{\sigma}_n^{c2} = \frac{2h_1h_2\cdots h_q}{n^2}\sum\limits_{i=1}^{n}\sum\limits_{j=1,j\neq i}^{n}\tilde{u}_i^2\widehat{f}_w(w_i)\tilde{u}_j^2\widehat{f}_w(w_j)K_{h_x}\left(x_i,x_j\right)$ and $\tilde{u}_i = (y_i - \widehat{y_i})$. At this point, it is worthy to point out that h_w and h_x are estimated separately with different rates and the asymptotic distribution of the proposed test statistic is sensitive to the choice of smoothing parameters. Alternatively, one can use bootstrap procedures to improve finite sample performances, see Gu et al. (2007) (GLL hereafter).

In the simulations presented in the next section, we extend the bootstrap test of Lavergne & Vuong (2000) and Gu et al. (2007) to admit both continuous and discrete variables since the null hypothesis doesn't change and generalized kernel function listed above can embody both types of variables.

Wild Bootstrap Procedure

Let u^* denote the wild bootstrap error that is obtained from the fitted residuals $\tilde{u}_i = (y_i - \widehat{y}_i)$,

$$u^* = \begin{cases} \frac{1-\sqrt{5}}{2}\tilde{u}_i & \text{with probability } r \\ \frac{1+\sqrt{5}}{2}\tilde{u}_i & \text{with probability } 1-r, \end{cases} \tag{2.11}$$

where $r = \frac{1+\sqrt{5}}{2\sqrt{5}}$. Following the steps stated in Gu et al. (2007), the bootstrap test statistic is obtained via

1. Use the wild bootstrap error u_i^* to construct $y_i^* = \widehat{y}_i + u_i^*$, then obtain the kernel estimator of $E^*\left(y_i^*|w_i\right) f_w(w_i)$ via

$$\widehat{y_i^*}\widehat{f}_w(w_i) = \frac{1}{n-1}\sum_{j\neq i}^{n} y_j^* K_{h_w}(x_i, x_j)$$

The estimated density-weighted bootstrap residual is

$$\begin{aligned} \widehat{u_i^*}\widehat{f}_w\left(w_i\right) &= \left(y_i^* - \widehat{y_i^*}\right)\widehat{f}_w\left(w_i\right) \\ &= y_i^*\widehat{f}_w\left(w_i\right) - \widehat{y_i^*}\widehat{f}_w\left(w_i\right). \end{aligned}$$

2. Compute the bootstrap test statistic

$$\widehat{I}_n^c * = \frac{1}{n\left(n-1\right)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \tilde{u}_i \widehat{f}_w\left(w_i\right) \tilde{u}_j \widehat{f}_w\left(w_j\right) K_{h_x}\left(x_i, x_j\right),$$

   and then obtain the estimated asymptotic variance through

$$\widehat{\sigma}_n^{c*2} = \frac{2h_1h_2\cdots h_q}{n^2} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \tilde{u}^*2_i \widehat{f}_w(w_i) \tilde{u}^*2_j \widehat{f}_w(w_j) K_{h_x}\left(x_i, x_j\right).$$

   Next, calculate the standardized bootstrap statistic by $T_n^{c*} = \left(nh_1h_2\cdots h_q\right)^{1/2} \widehat{I}_n^{c*}/\widehat{\sigma}_n^{c*}$.

3. Repeat previous two steps a large number of times, say $(B = 399)$ times, and obtain the empirical distribution of the B bootstrap test statistics. Let $T_{n(\alpha)}^{c*}$ denote the the α-percentile of the bootstrap distribution. We will reject the null hypothesis at significance level α if $T_n^c > T_{n(\alpha)}^{c*}$.

All Discrete Case

For now, instead of letting z be continuous variables, we consider the case that z is (are) discrete variable(s). Our focus is to test whether z is (are) irrelevant. The form of the null hypothesis remains to be the same. Now suppose z is of dimension $d - r$ and z_s, the sth component of z takes value of $0, 1, \ldots, c_s - 1$ $(s = 1, ..., d - r)$. Then the null hypothesis can be rewritten as $m(w, z = l) = m(x, z = 0)$, then a feasible test statistic can be constructed as

$$\widehat{I}_n^d = \frac{1}{n} \sum_{i=1}^{n} \sum_{z} \left[\hat{m}(w_i, z) - \hat{m}(w_i, z_1 = 0, \ldots, z_{d-r} = 0)\right]^2 \tag{2.12}$$

where $\sum\limits_{z}$ denote the summation of all possible values of $z \in \prod_{s=1}^{d-r}\{0, 1, ..., c_s - 1\}$. Therefore, there are totally $c = \prod_{s=1}^{q-r} c_s$ distinct values of z.

Define the nonparametric residuals

$$\hat{u}_i = Y_i - \hat{m}(x_i) - \bar{\delta}, \quad , i = 1, ..., n, \tag{2.13}$$

where $\hat{m}(x_i) = \sum_z = \hat{m}(x_i, z)/c$ and $\bar{\delta} = \sum_{i=1}^{n}(y_i - \hat{m}(x_i))/n$. As argued in Racine et al. (2006), the asymptotic null distribution of $\widehat{I_n^d}$ is quite complex, therefore they proposed the following bootstrap procedure:

Let u^* denote the wild bootstrap error that is obtained from the fitted residuals $\tilde{u}_i = (y_i - \widehat{y}_i)$,

$$u^* = \begin{cases} \frac{1-\sqrt{5}}{2}\tilde{u}_i & \text{with probability } r \\ \frac{1+\sqrt{5}}{2}\tilde{u}_i & \text{with probability } 1-r, \end{cases} \tag{2.14}$$

where $r = \frac{1+\sqrt{5}}{2\sqrt{5}}$.

1. Construct $y_i^* = \hat{m}(x_i) + u_i^*$ to obtain the bootstrap sample (y_i^*, x_i, z_i) where $i \in (1, ..., n)$.
2. Use the bootstrap sample to calculate the bootstrap test statistic $\widehat{I_n^{d*}}$, using the same LSCV bandwidth obtained initially.
3. Repeat steps 2-3 a large number (B) of times and obtain the empirical distribution of the bootstrap test statistics. Let $I_{n(\alpha)}^{d*}$ denote the α-percentile of the bootstrap distribution. We will reject the null hypothesis at significance level α if $I_n^d > I_{n(\alpha)}^{d*}$.

The advantage of the above bootstrap procedure is that the smoothing parameters only need to be computed by LSCV once. Second, wild bootstrap method is robust to heteroskedasticity.

Mixed Discrete-Continuous Case

The testing framework of Gu et al. (2007) admits discrete variables in the null hypothesis, however, their asymptotic results only retain to continuous variables. With the development of the generalized product kernel, we extend the bootstrap test statistic of Gu et al. (2007) to cases including both continuous and discrete variables in z. The obvious reason to do so is that no such tests exist in the literature to test both continuous and discrete variables simultaneously. Also, our simulation results confirm our conjecture that there is no reason in the finite sample that this procedure won't be able to include discrete variables, as long as the generalized product kernel mentioned above are properly used.

2.3 Monte Carlo Illustration

In this section, we present Monte Carlo Simulation results to compare the performances of lSCV bandwidths selection procedure and various tests introduced in the previous section. We select four data generating processes:

$$
\begin{aligned}
DGP_{2.1}: &\quad y = x_1 + \delta_1 x_2 + \delta_2 x_3 + \varepsilon. \\
DGP_{2.2}: &\quad y = x_1 + \delta_1 x_1 x_2 + \delta_2 x_1 x_3^2 + \varepsilon. \\
DGP_{2.3}: &\quad y = x_1 + x_2 + x_3 + \delta_1 x_1 (1 + x_2^2) \sin(.5\pi x_3) + \delta_2 x_3 \sin(x_2^3) + \varepsilon. \\
DGP_{2.4}: &\quad y = x_1 + x_2 + x_1 x_2 + \delta_1 x_1 x_3^2 + x_1^2 x_4 + \delta_2 x_2 x_3 x_5 + \delta_3 x_6^3 + \varepsilon.
\end{aligned}
$$

$DGP_{2.1}$ is a simple linear model and $DGP_{2.2}$ becomes more complicated in a sense of interactions among x_is. $DGP_{2.3}$ is designed to be a high frequency one. And $DPG_{2.4}$ increases the number of covariates from 3 to 6. Six covariates are rarely seen in most of the simulation studies and we would like to see to what extent the performances of LSCV bandwidth procedure and formal significance tests will deteriorate as dimensionality increases.

For all four $DGPs$, we use $\delta = (\delta_1, \delta_2, \delta_3)$ to control the degree of significance of variables. When $\delta = 0$, we assess the size properties otherwise we examine the power properties. We set $\delta_i \in \{0, .1, .5, 1\}$. We apply $DGP_{2.1}$, $DGP_{2.2}$ and $DGP_{2.3}$ to cases of z being all continuous variables and all discrete variables separately. We use $DGP_{2.4}$ to the case that z consists of both continuous and discrete variables. $x_i \sim N(0, 1)$, as well as ε. For discrete cases, we let

$$x_{2i} = \begin{cases} 0, & \text{with probability } 0.65, \\ 1, & \text{with probability } 0.35, \end{cases}$$

and

$$x_{3i} = \begin{cases} 0, & \text{with probability } 0.25, \\ 1, & \text{with probability } 0.4, \\ 2, & \text{with probability } 0.35, \end{cases}$$

where x_2 is unordered variable and x_3 is an ordered one.

We set sample size to be 100 and 200 with 399 bootstrap replications to generate the empirical distribution of the test statistic in each scenario accordingly.

For the part of LSCV bandwidth selection procedure, we first obtain the cross-validated bandwidths first and then use the rule-of-thumb to determine whether a variable is significant. For continuous variables, the criterion is two standard deviations (2sd) and the inter-quartile range (IQR); For discrete variables, the criterion becomes 80 of the theoretical upper bound of the bandwidth. For example, when x_3 is a ordered discrete variable, the upper bound is 1, therefore we have .8 to serve as a rule-of-thumb.

2.3.1 Continuous only case

Tables 2.1, 2.2 and 2.3 report results for all $x_i, i \in \{1, 2, 3\}$ are continuous variables. Part (a) of all three tables present size and power properties of the bootstrap test statistic using the same bandwidth selection strategies where $h = c \cdot sd_i n^{-1/(4+d)}, i \in \{1, 2, 3\}$. As we can see, the rate of convergence equals to $n^{-1/(4+d)}$ and d is the dimension of the model. sd_i is the standard deviation of variable x_i. They choose different c to control the degrees of smoothness and assess the sensibility of their test statistic to different bandwidths. [2] The first column of part (b) of all three tables gives us the size and power properties when we use LSCV bandwidths scaling factors chosen to automatically smooth out the irrelevant variables, that is c is chosen by equation 2.6. Except for the rest of the two columns in part (b) of all three tables, all size and power are examined at 1%, 5% and 10% significant levels. The last two columns of part (b) report the probabilities of failing to reject the null for $x_i, i \in \{1, 2, 3\}$ being irrelevant according to the rule-of-thumb $2sd$ and IQR respectively.

First, we observe that the the size properties of Gu et al. (2007) bootstrap test statistic are very impressive, the actual size approaches to the nominal size. However, using scaling factors coming from LSCV bandwidth selection method, their bootstrap test statistic is severely over sized, even though the power seems to be superior. It is common to see that both strategies improve the size and power properties when sample size increases. Moreover, using IQR as the rule-of-thumb to reject the null hypothesis seems to be more effective when the null is true compared with $2sd$, however, it has the cost of higher probability to erroneously accept the null when variables are deemed relevant. We also notice that when sample size increases, both IQR and $2sd$ improve their performances.

For the raw interpretation of the LSCV bandwidths, we see that individually speaking, it has

[2] In Gu et al. (2007), they chose $c = 0.25, 0.5, 1, 2$, therefore we adopt the same strategy. Also, their theories requires to undersmooth the unrestricted model while keeping the restricted model optimal smoothed.

Table 2.1: $DGP_{2.1}$

(a) Gu, Li & Liu 2007 Bandwidths

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.003	0.045	0.110	0.013	0.065	0.120	0.013	0.050	0.123	0.038	0.103	0.163
$\delta = 0.5$	0.015	0.080	0.155	0.080	0.228	0.346	0.378	0.612	0.742	0.832	0.965	0.987
$\delta = 1$	0.020	0.168	0.318	0.366	0.659	0.784	0.967	1.000	1.000	1.000	1.000	1.000
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.053	0.080	0.013	0.053	0.105	0.015	0.063	0.125	0.015	0.058	0.103
$\delta = 0.1$	0.010	0.028	0.090	0.015	0.063	0.113	0.035	0.103	0.155	0.040	0.138	0.223
$\delta = 0.5$	0.023	0.103	0.188	0.158	0.373	0.489	0.722	0.892	0.945	0.987	1.000	1.000
$\delta = 1$	0.090	0.358	0.524	0.799	0.957	0.980	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth Results

$n = 100$	*LSCV*			2*sd				IQR			
α	1%	5%	10%	$x1$	$x2$	$x3$	Joint	$x1$	$x2$	$x3$	Joint
$\delta = 0$	0.045	0.133	0.213	0.000	0.687	0.604	0.426	0.000	0.757	0.712	0.561
$\delta = 0.1$	0.038	0.130	0.236	0.000	0.551	0.564	0.318	0.000	0.669	0.659	0.446
$\delta = 0.5$	0.466	0.737	0.852	0.000	0.018	0.013	0.000	0.000	0.048	0.043	0.000
$\delta = 1$	0.927	0.982	0.992	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$n = 200$	*LSCV*			2*sd				IQR			
α	1%	5%	10%	$x1$	$x2$	$x3$	Joint	$x1$	$x2$	$x3$	Joint
$\delta = 0$	0.040	0.098	0.178	0.000	0.669	0.639	0.441	0.000	0.900	0.865	0.769
$\delta = 0.1$	0.048	0.138	0.243	0.000	0.486	0.489	0.263	0.000	0.822	0.837	0.687
$\delta = 0.5$	0.835	0.942	0.965	0.000	0.000	0.000	0.000	0.000	0.168	0.153	0.008
$\delta = 1$	0.997	1.000	1.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000

a higher chance to smooth out a irrelevant variable than jointly remove multiple irrelevant variables. On average speaking, the chance of accepting the null drops approximately 20%. For example, let us look at table 2.3 part (b) column 2. When $\delta = 0$, the probability of smooth away variable x_2 is 0.687 and the probability of smooth away variable x_3 is 0.604. But it is only 0.426 for the LSCV to jointly remove both x_2 and x_3. This is also true when using

Table 2.2: $DGP_{2.2}$

(a) Gu, Li & Liu 2007 Bandwidths

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.008	0.065	0.093	0.010	0.065	0.120	0.005	0.048	0.103	0.010	0.065	0.108
$\delta = 0.5$	0.003	0.088	0.078	0.020	0.088	0.155	0.058	0.148	0.218	0.073	0.228	0.323
$\delta = 1$	0.008	0.175	0.143	0.043	0.175	0.301	0.263	0.536	0.657	0.499	0.769	0.857
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.038	0.080	0.013	0.053	0.105	0.015	0.063	0.125	0.015	0.058	0.103
$\delta = 0.1$	0.003	0.038	0.080	0.015	0.045	0.108	0.023	0.088	0.128	0.015	0.058	0.100
$\delta = 0.5$	0.005	0.055	0.095	0.020	0.088	0.158	0.068	0.213	0.346	0.213	0.469	0.612
$\delta = 1$	0.033	0.143	0.256	0.198	0.466	0.602	0.732	0.902	0.952	0.952	0.992	0.995

(b) LSCV Bandwidth Results

$n = 100$	*LSCV*			2*sd				IQR			
α	1%	5%	10%	$x1$	$x2$	$x3$	Joint	$x1$	$x2$	$x3$	Joint
$\delta = 0$	0.045	0.133	0.213	0.000	0.687	0.604	0.426	0.000	0.757	0.712	0.561
$\delta = 0.1$	0.033	0.120	0.190	0.000	0.554	0.586	0.341	0.000	0.662	0.672	0.451
$\delta = 0.5$	0.045	0.203	0.308	0.000	0.100	0.185	0.018	0.000	0.221	0.203	0.030
$\delta = 1$	0.188	0.451	0.604	0.000	0.020	0.038	0.005	0.000	0.053	0.043	0.005
$n = 200$	*LSCV*			2*sd				IQR			
α	1%	5%	10%	$x1$	$x2$	$x3$	Joint	$x1$	$x2$	$x3$	Joint
$\delta = 0$	0.040	0.098	0.178	0.000	0.669	0.639	0.441	0.000	0.900	0.865	0.769
$\delta = 0.1$	0.028	0.105	0.165	0.000	0.491	0.617	0.343	0.000	0.872	0.825	0.724
$\delta = 0.5$	0.115	0.286	0.398	0.000	0.018	0.035	0.003	0.000	0.514	0.070	0.018
$\delta = 1$	0.559	0.777	0.887	0.000	0.000	0.000	0.000	0.000	0.133	0.003	0.000

IQR as the rule-of-thumb. We see that IQR has the problem of higher chance to accept the null when variables are indeed revelment. If one were worried to mistakenly remove important variables in the model, $2sd$ is a better choice. Also, from $DGP_{2.1}$ to $DGP_{2.2}$, neither $2sd$ nor IQR is unaffected by the interactions among variables. They perform better when variables enter the model separately, as we expected.

Table 2.3: $DGP_{2.3}$

(a) Gu, Li & Liu 2007 Bandwidths

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.005	0.060	0.095	0.008	0.060	0.123	0.005	0.060	0.125	0.025	0.090	0.138
$\delta = 0.5$	0.008	0.080	0.165	0.083	0.206	0.308	0.271	0.481	0.607	0.579	0.837	0.917
$\delta = 1$	0.028	0.160	0.333	0.396	0.664	0.769	0.957	0.985	0.992	1.000	1.000	1.000
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.038	0.080	0.013	0.053	0.105	0.015	0.063	0.125	0.015	0.058	0.103
$\delta = 0.1$	0.008	0.028	0.085	0.018	0.055	0.103	0.025	0.100	0.143	0.038	0.118	0.198
$\delta = 0.5$	0.023	0.103	0.170	0.138	0.318	0.429	0.586	0.772	0.880	0.942	0.980	0.985
$\delta = 1$	0.085	0.346	0.509	0.797	0.927	0.960	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth Results

$n = 100$	*LSCV*			2*sd				IQR			
α	1%	5%	10%	$x1$	$x2$	$x3$	Joint	$x1$	$x2$	$x3$	Joint
$\delta = 0$	0.045	0.133	0.213	0.000	0.687	0.604	0.426	0.000	0.757	0.712	0.561
$\delta = 0.1$	0.033	0.120	0.190	0.000	0.554	0.586	0.341	0.000	0.662	0.672	0.451
$\delta = 0.5$	0.045	0.203	0.308	0.000	0.100	0.185	0.018	0.000	0.221	0.203	0.030
$\delta = 1$	0.188	0.451	0.604	0.000	0.020	0.038	0.005	0.000	0.053	0.043	0.005
$n = 200$	*LSCV*			2*sd				IQR			
α	1%	5%	10%	$x1$	$x2$	$x3$	Joint	$x1$	$x2$	$x3$	Joint
$\delta = 0$	0.040	0.098	0.178	0.000	0.669	0.639	0.441	0.000	0.900	0.865	0.769
$\delta = 0.1$	0.028	0.105	0.165	0.000	0.491	0.617	0.343	0.000	0.872	0.825	0.724
$\delta = 0.5$	0.115	0.286	0.398	0.000	0.018	0.035	0.003	0.000	0.514	0.070	0.018
$\delta = 1$	0.559	0.777	0.887	0.000	0.000	0.000	0.000	0.000	0.133	0.003	0.000

2.3.2 Discrete only case

In this section, we set x_2 and x_3 to be discrete variables define above. We let x_1 be a continuous variables. Unlike the upper bounds of continuous variables which go to infinity, the upper bound of an unordered discrete variables is 0.5 and ordered discrete variable is 1. Therefore, we use 80% of their upper bounds to serve as the rule-of-thumb to determine the corresponding variables significance. Table 2.4, 2.5 and 2.6 provides size and power results for Racine et al. (2006) test statistic when z in our null hypothesis is consists of two discrete variables, x_2 and x_3. For the test statistic of Racine et al. (2006), we adopt the same bandwidth selection strategy as in Gu et al. (2007).

It is surprisingly to see that the test statistic has very good size and power properties in both individual testing and joint settings. [3] The LSCV bandwidth criterion according to the rule of 80% of upper bounds shows similar patterns as it was for all continuous case. In the joint settings, LSCV is inferior compared with individual settings when we use the size of bandwidth to determine the relevance of different discrete variables. Also, as n increases, LSCV bandwidth criterion also performs better.

Compared with LSCV bandwidth selection criterion, we find that the formal testing framework of Racine et al. (2006) delivers remarkably size and power properties . While LSCV bandwidths criterion leads to a high probability of smoothing away a relevant variable, especially when δ is small. For example, in table 2.5 part (b) column one, we see that for x_2, there is no significant difference in terms of the percentage of the time that LSCV removes x_2 when δ changes its value from 0 to 0.1. Even at $\delta = 0.5$, there is still 59.9% of the chance to conclude that x_2 is irrelevant. More importantly, for joint tests, conventional testing framework is far more accurate to determine the significance of discrete variables.

[3] Again, we undersmooth the unrestricted model as we did in all continuous case and use the standard smoothing parameters in the unrestricted models.

Table 2.4: $DGP_{2.1}$ where x_2 and x_3 are discrete variables

(a) Gu, Li & Liu 2007 Bandwidths

x_2 and x_3 Joint Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.015	0.053	0.105	0.023	0.063	0.130	0.015	0.080	0.135	0.020	0.075	0.125
$\delta = 0.1$	0.020	0.070	0.110	0.023	0.078	0.135	0.023	0.088	0.160	0.035	0.088	0.175
$\delta = 0.5$	0.135	0.328	0.461	0.258	0.506	0.619	0.424	0.662	0.767	0.544	0.777	0.872
$\delta = 1$	0.727	0.925	0.972	0.965	0.992	0.995	0.997	0.997	1.000	0.997	1.000	1.000
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.003	0.050	0.083	0.015	0.045	0.073	0.008	0.055	0.088	0.003	0.050	0.103
$\delta = 0.1$	0.005	0.050	0.110	0.025	0.053	0.118	0.015	0.070	0.118	0.023	0.110	0.165
$\delta = 0.5$	0.386	0.664	0.764	0.642	0.832	0.907	0.792	0.947	0.972	0.925	0.980	0.990
$\delta = 1$	0.997	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x_2 Individual Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.055	0.090	0.005	0.038	0.108	0.005	0.048	0.108	0.015	0.050	0.108
$\delta = 0.1$	0.010	0.050	0.095	0.005	0.053	0.125	0.013	0.063	0.140	0.018	0.083	0.135
$\delta = 0.5$	0.038	0.160	0.261	0.098	0.241	0.356	0.135	0.338	0.471	0.198	0.444	0.571
$\delta = 1$	0.013	0.058	0.128	0.080	0.293	0.451	0.238	0.576	0.729	0.404	0.752	0.870
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.095	0.015	0.048	0.100	0.020	0.050	0.083	0.010	0.055	0.095
$\delta = 0.1$	0.013	0.053	0.095	0.018	0.055	0.118	0.025	0.055	0.130	0.015	0.078	0.150
$\delta = 0.5$	0.093	0.286	0.401	0.198	0.439	0.574	0.358	0.617	0.742	0.471	0.712	0.842
$\delta = 1$	0.504	0.764	0.842	0.779	0.907	0.950	0.922	0.967	0.987	0.957	0.987	0.992

x_3 Individual Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.018	0.063	0.115	0.015	0.068	0.135	0.023	0.073	0.148	0.023	0.078	0.150
$\delta = 0.5$	0.123	0.331	0.444	0.236	0.474	0.579	0.353	0.609	0.727	0.474	0.724	0.830
$\delta = 1$	0.759	0.915	0.947	0.927	0.990	0.995	0.985	1.000	1.000	1.000	1.000	1.000
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.040	0.115	0.005	0.050	0.110	0.003	0.038	0.095	0.010	0.045	0.100
$\delta = 0.1$	0.008	0.063	0.130	0.018	0.075	0.130	0.015	0.085	0.153	0.020	0.110	0.175
$\delta = 0.5$	0.363	0.627	0.742	0.576	0.810	0.895	0.762	0.942	0.970	0.885	0.970	0.985
$\delta = 1$	0.995	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth results using 80% of the upper bound.

	$n = 100$			$n = 200$		
	x_2	x_3	Joint	x_2	x_3	Joint
$\delta = 0$	0.697	0.551	0.411	0.772	0.602	0.501
$\delta = 0.1$	0.684	0.506	0.378	0.707	0.521	0.398
$\delta = 0.5$	0.363	0.030	0.010	0.211	0.000	0.000
$\delta = 1$	0.028	0.000	0.000	0.000	0.000	0.000

Table 2.5: $DGP_{2.2}$, where x_2 and x_3 are discrete variables.

(a) Gu, Li & Liu 2007 Bandwidths

x_2 and x_3 Joint Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.015	0.053	0.105	0.023	0.063	0.130	0.015	0.080	0.135	0.020	0.075	0.125
$\delta = 0.1$	0.015	0.063	0.110	0.020	0.073	0.150	0.015	0.100	0.163	0.023	0.090	0.145
$\delta = 0.5$	0.188	0.409	0.514	0.308	0.566	0.694	0.436	0.699	0.789	0.414	0.722	0.812
$\delta = 1$	0.759	0.910	0.932	0.887	0.967	0.990	0.962	0.997	1.000	0.962	0.997	0.997
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.003	0.050	0.083	0.015	0.045	0.073	0.008	0.055	0.088	0.003	0.050	0.103
$\delta = 0.1$	0.010	0.045	0.100	0.018	0.063	0.118	0.018	0.075	0.138	0.010	0.090	0.153
$\delta = 0.5$	0.509	0.707	0.799	0.692	0.845	0.915	0.822	0.952	0.980	0.865	0.980	0.992
$\delta = 1$	0.997	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x_2 Individual Significance Test

$n = 100$	c=.25			c=.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.040	0.078	0.020	0.050	0.090	0.005	0.045	0.123	0.005	0.033	0.110
$\delta = 0.1$	0.013	0.040	0.083	0.018	0.053	0.088	0.005	0.043	0.120	0.008	0.040	0.113
$\delta = 0.5$	0.018	0.075	0.118	0.023	0.100	0.158	0.030	0.098	0.185	0.023	0.108	0.195
$\delta = 1$	0.073	0.168	0.298	0.123	0.281	0.431	0.145	0.371	0.531	0.133	0.401	0.559
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.005	0.053	0.095	0.005	0.050	0.115	0.000	0.043	0.113	0.005	0.030	0.073
$\delta = 0.1$	0.010	0.060	0.085	0.013	0.050	0.090	0.015	0.048	0.088	0.010	0.045	0.105
$\delta = 0.5$	0.020	0.073	0.135	0.028	0.105	0.178	0.025	0.128	0.218	0.018	0.095	0.203
$\delta = 1$	0.075	0.223	0.393	0.160	0.381	0.519	0.233	0.494	0.637	0.216	0.534	0.687

x_3 Individual Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.108	0.010	0.055	0.113	0.005	0.055	0.108	0.013	0.068	0.110
$\delta = 0.1$	0.018	0.063	0.133	0.015	0.075	0.150	0.015	0.078	0.160	0.018	0.088	0.165
$\delta = 0.5$	0.281	0.506	0.622	0.454	0.689	0.772	0.596	0.812	0.890	0.544	0.835	0.917
$\delta = 1$	0.865	0.960	0.980	0.962	0.992	0.995	0.990	0.997	1.000	0.977	0.997	1.000
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.040	0.115	0.005	0.050	0.110	0.003	0.038	0.095	0.010	0.045	0.100
$\delta = 0.1$	0.020	0.070	0.148	0.013	0.080	0.165	0.018	0.083	0.185	0.020	0.088	0.193
$\delta = 0.5$	0.629	0.810	0.890	0.817	0.955	0.980	0.937	0.987	0.992	0.957	0.992	0.997
$\delta = 1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth results using 80% of the upper bound.

	$n = 100$			$n = 200$		
	x_2	x_3	Joint	x_2	x_3	Joint
$\delta = 0$	0.697	0.551	0.411	0.772	0.602	0.501
$\delta = 0.1$	0.699	0.409	0.311	0.762	0.298	0.233
$\delta = 0.5$	0.599	0.000	0.000	0.546	0.000	0.000
$\delta = 1$	0.378	0.000	0.000	0.103	0.000	0.000

Table 2.6: $DGP_{2.3}$, where x_2 and x_3 are discrete variables

(a) Gu, Li & Liu 2007 Bandwidths

x_2 and x_3 Joint Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.015	0.053	0.105	0.023	0.063	0.130	0.015	0.080	0.135	0.020	0.075	0.125
$\delta = 0.1$	0.018	0.060	0.120	0.028	0.080	0.145	0.018	0.095	0.158	0.038	0.108	0.185
$\delta = 0.5$	0.150	0.333	0.449	0.258	0.499	0.609	0.409	0.659	0.797	0.544	0.805	0.872
$\delta = 1$	0.832	0.935	0.965	0.955	0.992	0.997	0.987	1.000	1.000	0.995	1.000	1.000
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.003	0.050	0.083	0.015	0.045	0.073	0.008	0.055	0.088	0.003	0.050	0.103
$\delta = 0.1$	0.008	0.060	0.093	0.018	0.053	0.110	0.018	0.070	0.128	0.020	0.088	0.158
$\delta = 0.5$	0.366	0.617	0.742	0.586	0.837	0.917	0.815	0.947	0.977	0.920	0.982	0.992
$\delta = 1$	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x_2 Individual Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.055	0.090	0.005	0.038	0.108	0.005	0.048	0.108	0.015	0.050	0.108
$\delta = 0.1$	0.010	0.048	0.088	0.005	0.040	0.120	0.010	0.055	0.123	0.020	0.080	0.130
$\delta = 0.5$	0.030	0.133	0.236	0.078	0.211	0.311	0.113	0.328	0.434	0.175	0.409	0.536
$\delta = 1$	0.190	0.381	0.484	0.318	0.544	0.697	0.496	0.724	0.815	0.589	0.807	0.887
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.095	0.015	0.048	0.100	0.020	0.050	0.083	0.010	0.055	0.095
$\delta = 0.1$	0.010	0.055	0.098	0.015	0.060	0.103	0.020	0.063	0.123	0.015	0.075	0.145
$\delta = 0.5$	0.100	0.246	0.378	0.165	0.391	0.546	0.318	0.574	0.707	0.409	0.687	0.789
$\delta = 1$	0.516	0.742	0.837	0.752	0.902	0.940	0.880	0.952	0.972	0.917	0.970	0.982

x_3 Individual Significance Test

$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.108	0.010	0.055	0.113	0.005	0.055	0.108	0.013	0.068	0.110
$\delta = 0.1$	0.015	0.068	0.120	0.013	0.073	0.123	0.015	0.083	0.155	0.023	0.088	0.158
$\delta = 0.5$	0.113	0.293	0.421	0.213	0.439	0.594	0.311	0.602	0.702	0.464	0.692	0.787
$\delta = 1$	0.609	0.817	0.887	0.787	0.927	0.965	0.920	0.982	0.997	0.972	0.997	0.997
$n = 200$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.040	0.115	0.005	0.050	0.110	0.003	0.038	0.095	0.010	0.045	0.100
$\delta = 0.1$	0.015	0.060	0.120	0.008	0.075	0.125	0.020	0.083	0.143	0.015	0.100	0.165
$\delta = 0.5$	0.291	0.506	0.639	0.456	0.702	0.789	0.639	0.832	0.920	0.764	0.932	0.975
$\delta = 1$	0.955	0.992	1.000	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth results using 80% of the upper bound.

	$n = 100$			$n = 200$		
	x_2	x_3	Joint	x_2	x_3	Joint
$\delta = 0$	0.697	0.551	0.411	0.772	0.602	0.501
$\delta = 0.1$	0.682	0.476	0.356	0.719	0.454	0.353
$\delta = 0.5$	0.393	0.023	0.003	0.251	0.000	0.000
$\delta = 1$	0.050	0.000	0.000	0.000	0.000	0.000

2.3.3 Mixed discrete-continuous case

For the tests of mixed discrete and continuous variables jointly, we use $DGP_{2.4}$ which has six variables totally so that we can test a variety of different null hypothesis. Though the lack of theoretical support limits the application of formal testing framework, there is no reason the test of Lavergne & Vuong (2000) and Gu et al. (2007) can not be extended to include discrete variables.

The first null hypothesis is $H_0 : x_3, x_5, x_6$ are insignificant. Here x_3 is the continuous variable and x_5 and x_6 are discrete ones. We see that the performance of the proposed test statistic is sensitive to the choice of scaling factor c, which is not a desirable property we would like to have. Size is a little above the nominal size and power seems to be affected by c to a larger extent.

For the null hypothesis $H_0 : x_1, x_2, x_4$ are insignificant, we can only report power properties since they represent the extreme case of a false null hypothesis, where all three variables are actually relevant in the model. Looking at the first column, we see that when scaling factor equals to 0.25, as δ increases, the power does not increase much. But when the scaling factor increases, we see power goes up quickly. The scaling factor seems to paly an important role here. The third null hypothesis $H_0 : x_2, x_5$ are insignificant, which is only partially right, only power can be presented. And the pattern looks like the case of the second null hypothesis.

Next let us take a look at LSCV bandwidth results. Totally we have three continuous variables and three discrete variables. The first two columns test six variables in the individual settings. First we notice that when a variable is relevant, LSCV bandwidth criterion never smooth it out, e.g. x_1, x_2 and x_4. For $d = 6$ which is a large dimension considering sample sizes, we see that LSCV still has a decent chance to detect an irrelevant variable, either using $2sd$ or IQR criterion. Also, like the previous cases, IQR results in a higher percent-

Table 2.7: $DGP_{2.4}$,where x_4, x_5 and x_6 are discrete variables.

(a) Gu, Li & Liu 2007 Bandwidths

Joint Significance Test(Size and Power) $H_0 : x_3, x_5$ and x_6 are insignificant												
$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.017	0.082	0.150	0.020	0.067	0.120	0.017	0.050	0.105	0.007	0.050	0.102
$\delta = 0.1$	0.018	0.090	0.165	0.048	0.173	0.263	0.173	0.348	0.469	0.356	0.619	0.729
$\delta = 0.5$	0.173	0.494	0.704	0.885	0.980	0.987	1.000	1.000	1.000	1.000	1.000	1.000
$\delta = 1$	0.223	0.609	0.832	0.970	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Joint Significance Test(Only Power) $H_0 : x_1, x_2$ and x_4 are insignificant												
$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.028	0.180	0.338	0.291	0.604	0.729	0.890	0.965	0.992	1.000	1.000	1.000
$\delta = 0.1$	0.028	0.183	0.336	0.291	0.604	0.737	0.882	0.970	0.992	1.000	1.000	1.000
$\delta = 0.5$	0.065	0.193	0.358	0.336	0.637	0.792	0.887	0.977	0.992	0.997	1.000	1.000
$\delta = 1$	0.075	0.246	0.396	0.404	0.712	0.837	0.897	0.972	0.992	0.990	1.000	1.000

Joint Significance Test(Only Power) $H_0 : x_2$ and x_5 are insignificant												
$n = 100$	c=0.25			c=0.5			c=1			c=2		
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.058	0.195	0.301	0.211	0.449	0.564	0.619	0.805	0.890	0.947	0.987	0.992
$\delta = 0.1$	0.058	0.168	0.288	0.228	0.446	0.561	0.642	0.817	0.902	0.952	0.992	0.992
$\delta = 0.5$	0.128	0.296	0.388	0.358	0.586	0.682	0.820	0.935	0.967	0.987	0.995	1.000
$\delta = 1$	0.223	0.436	0.561	0.617	0.810	0.880	0.950	0.977	0.990	0.995	0.995	1.000

(b) LSCV Bandwidth Results

$n = 100$	Continuous(2*sd)			Discrete(0.8)			Joint(2*sd,0.8)	Continuous(IQR)			Joint(IQR,0.8)
Variable	x_1	x_2	x_3	x_4	x_5	x_6	Joint	x_1	x_2	x_3	Joint
$\delta = 0$	0.000	0.000	0.647	0.000	0.734	0.694	0.341	0.000	0.000	0.774	0.401
$\delta = 0.1$	0.000	0.000	0.632	0.000	0.674	0.311	0.093	0.000	0.000	0.767	0.108
$\delta = 0.5$	0.000	0.000	0.183	0.000	0.569	0.000	0.000	0.000	0.003	0.308	0.000
$\delta = 1$	0.000	0.000	0.028	0.000	0.471	0.000	0.000	0.000	0.000	0.053	0.000

age to smooth out irrelevant variables, but also has the tendency of mistakenly removing relevant ones. Moreover, as we expect in a higher dimension setting, the joint measure of determination from LSCV bandwidths is worse than other cases.

2.4 Conclusion

This chapter of my dissertation attempts to study various methods that can be used to reduce the dimensionality of a nonparametric model by detecting irrelevant variables. We broadly discuss two sets of methods, one is the cross-validated bandwidth selection procedures and the other one is the conventional testing frameworks. The LSCV bandwidths has the ability to automatically smooth away irrelevant variables through assigning a large bandwidth to a redundant variable. Our simulation studies reveal that this appealing feature of LSCV procedure works well for a individual test, but as the number of variables in the null hypothesis goes up, its performance breaks down quickly. When a variable indeed relevant, LSCV bandwidth criterion never oversmooth it. However, compared with other formal testing frameworks, we discover that for joint significance setting, the formal testing frameworks are still paramount in terms of size and power properties.

The conventional tests introduced by Lavergne & Vuong (2000), Racine et al. (2006) and Gu et al. (2007) are remarkably accurate in terms of size and power performance. We extend the framework of Gu et al. (2007) to admit both categorical and continuous variables in the null hypothesis simultaneously. Even though there is not any theoretical properties for this action, we find that it still has decent size and power performances, though a little bit of oversized. We suggest to use LSCV bandwidth selection procedure to detect suspicious variables whose bandwidths exceed the *ad hoc* criterion used in Hall et al. (2007) and then apply formal testing frameworks to confirm in order to conduct sound reduction of dimensionality in nonparametric models.

Further research should focus on developing testing mechanisms that can test not only the significance of a variable or a group of variables, but also the validity of imposing any arbitrary constraints on the regression model.

Chapter 3

Bootstrap Nonparametric Significance Testing for Restrictions on Higher Order Derivatives by Data Sharpening Technique

(ABSTRACT)

The need for imposing restrictions on nonparametric kernel smoothed regressions has been increasing as nonparametric models become more popular and acknowledged. By so far, to the best of our knowledge, the only general framework that can provide a simple yet approach toward constrained nonparametric regression was proposed by Racine et al. (2008), which is referred as the constrained weighted bootstrap (CWB) method. The competing method, so-called data sharpening methods, see Braun & Hall (2001), was designed to perturb the data in order to improve the performance of a statistical method. In this chapter, we expand the application of data sharpening technique to the field of imposing arbitrary constraints on the curve estimation, including both the conditional mean and its higher order derivatives, either equality constraints or non-equality constraints. A testing framework based on a distance criterion is proposed and the asymptotic distribution of the test statistic is also

provided. Numerical examples are demonstrated in a way that the finite sample performances of both CWB method and data sharpening method can be compared. Various Monte Carlo simulations are illustrated and an empirical application is examined.

3.1 Introduction

The kernel smoothing method is appraised for its robustness against misspecifications of functional forms since the date of its born, and it continues to gain popularity among different fields. However, due to the nature of its nonparametric formulation, unlike the parametric counterpart, kernel smoothed regression has the limitation of being incapable to impose a conventional economic constraint, such as monotonicity, convexity, or any other restrictions applied to higher order derivatives, etc.. It remains to be a source of frustration since the appeal of being functional form free is weakened by the inability of imposing a null hypothesis on the shape of the conditional expectation, which in turn can be easily achieved by its parametric counterparts.

Data sharpening method, which was originally introduced by Choi & Hall (1999) and Choi et al. (2000), is employed to perturb the data so as to enhance properties of the relatively conventional statistical procedures. For example, by tilting the data in a way that making them more clustered than before, the corresponding density estimator becomes less biased. Recently, Braun & Hall (2001) suggested to extend the scope of data sharpening method to estimate a curve subject to qualitative constraints, such as monotonicity. The method involves altering the positions of the data values, controlled by minimizing the total distance that data are moved, subject to a constraint. In the various scenarios described in Braun & Hall (2001), data sharpening can either change the values of the response variable or the explanatory variables.

Yet the bootstrap hypothesis testing methodology presented in Braun & Hall (2001) discussed how to impose general constraints on the conditional expectations and statistically test the validity of these null hypothesis. In this paper, we fill in this blank by imposing general constraints on the local polynomial regression framework of Masry (1996*a*) and Masry (1996*b*) and utilizing the bootstrap hypothesis testing procedure of Braun & Hall (2001) to obtain the empirical distribution of proposed distance test statistic and test the validity of the null hypothesis.

The closest competitor of data sharpening method was the biased-bootstrap method, see Hall & Presnell (1999), which alters the weights attached to data values without adjusting the original observed data points. Data sharpening, on the other hand, keeps the weights to be fixed but shift the data points directly. Both can be viewed as ways of "data tuning". However, the most related method with respect to data sharpening nowadays is the CWB method demonstrated in Racine et al. (2008), which can be viewed as the generalization of biased-bootstrap method. Traditionally, the weights attached to data points are considered as of a density distribution; while the CWB method still retains the sum of weights to be 1, but allowing individual weights to vary enough that they can be both positive and negative. The null hypothesis can be generalized to be multiple constraints and the validity of the change of the weights is accessed by a bootstrap procedure.

As noted in Braun & Hall (2001), a disadvantage of the biased-bootstrap in the setting is that in order to satisfy the condition that the summation of varying weights equals to 1, the distance criterion based on the total change of weights accommodates the low weights points with assigning a large weight on other points. This can result in large mean squared errors. In our framework, we translate data sharpening method into an analogous of the CWB method without the density restriction. Data points are altered as little as possible according to a distance measurement subject to general constraints that are conventionally imposed on higher order derivatives. Our numerical examples illustrate that data sharpening method dose have better finite sample performances in terms of lower mean squared errors.

3.2 The Data Sharpening Estimator

From now on, we let $\{X_i, Y_i\}$ denote sample pairs we wish to adjust as little as possible, according to some distance criterion, subject to one or more economic constraints, where $i \in (1, \ldots, n)$. Y_i is a scalar and X_i is of dimension d and n is the sample size. Data sharpening directly change the data values, but it is not necessary to alter every components of the data vectors. As a matter of fact, it is more straightforward to focus our attention on the response variable Y_i, which means we only shift the response variable without tilting the explanatory variables. This is because of the following reasons: First of all, practitioners are, more often than not, interested in the average mean response $m(x) = E(Y|X = x)$. The enforcement of constraints on the conditional expectations $m(x)$ restricts its shape to $\tilde{m}(x)$, which satisfies the null hypothesis. Let $\tilde{Y}_i$ denote the newly selected values through data sharpening technique to replace the original observations, therefore we have $\tilde{m}(x) = E(\tilde{Y}|X = x)$. Then the distance, measured by some distance criterion $D(\cdot)$, between the new data pairs $\{\tilde{Y}_i, X_i\}$ and $\{X_i, Y_i\}$ can be expressed as $D\left(Y, \tilde{Y}\right)$. Hence researchers can directly compare the sharpened $\tilde{Y}_i$ with respect to Y_i, which is of direct interest. Second, X_i is a vector rather than a scalar in most applied settings. The choice of appropriate components among explanatory variables can be difficult. Suppose a constraint can be imposed by alter ether X_{1i} or X_{2i}, or both of them. It could be the case that contrasting strategies give rise to different statistical significant levels, which lead to contradicting conclusions and false inferences thereafter. Also sometimes it is inappropriate to change the values of some explanatory variables, for example, a male/female indicator. Third, by changing the data pairs exclusively on the value of Y_i, both biased bootstrap and CWB methods can be viewed as a special edition of data sharpening method, since tilting the empirical distribution of the data to the least amount is just another way of transform the response variable as little as possible subject to some constraints. However, data sharpening method is not bounded by the condition that the summation of individual weights equal to 1. As we will show later on, this enables us to compare the numerical performances between data sharpening and CWB methods.

The model under consideration is

$$y = m(\mathbf{x}) + \epsilon, \tag{3.1}$$

where $\mathbf{x}$ is a $d \times 1$ vector. Throughout this paper, we assume $m(\mathbf{x})$ has continuous derivatives up to order $p+1$ at point $\mathbf{x}$. In what follows, in order to breviate notations, we let $\{x_i, y_i\}_{i=1}^n$ to denote the original observations and $\hat{y_i}$ to replace $\hat{m}(x_i) = E(Y|X = x_i)$ the unrestricted conditional expectation at point x_i. Let $\{x_i, \tilde{y_i}\}_{i=1}^n$ represent the new data pairs selected by data sharpening subject to one or more constraints. Following the notation of Masry (1996*a*) and Masry (1996*b*), for a $d \times 1$ vector $\mathbf{k}$, define

$$\mathbf{k} = (k_1, ..., k_d), \quad \mathbf{k}! = k_1! \times \cdots k_d!, \quad |\mathbf{k}| = \sum_{i=1}^{d} k_i, \tag{3.2}$$

$$\mathbf{x}^{\mathbf{k}} = x_1^{k1} \times \cdots \times x_d^{kd}, \quad \sum_{0 \le |\mathbf{k}| \le p} = \sum_{j=0}^{p} \sum_{k_1=0}^{j} \cdots \sum_{k_d=0; k_1+\cdots+k_d=j}^{j}, \tag{3.3}$$

and

$$m^{(k)}(\mathbf{x}) = \frac{\partial^{k_1} m(x)}{\partial x_1^{k_1}} \cdots \frac{\partial^{k_d} m(x)}{\partial x_d^{k_d}}. \tag{3.4}$$

Therefore, if $k = (0, \ldots, 0)$, equation (3.4) represents the conditional mean itself, while $k = (1, 1, 0, \ldots, 0)$ represents the cross partial derivatives of $\partial^2 m(x)/\partial x_1 \partial x_2$. The general constraints we wish to impose is of form

$$l(x) \leq m^{(k)}(x) \leq u(x) \tag{3.5}$$

for some lower bounds $l(\cdot)$ and upper bound $u(\cdot)$ and $0 \leq |k| \leq p$. We consider local polynomial regression framework introduced in Masry (1996*a*) and Masry (1996*b*), which essentially is a linear smoother of the response variable. To be consistent with the CWB method proposed in Racine et al. (2008), we adopt their notation and write the conditional expectation of $m^{(k)}(x)$ in the following way

$$\hat{m}^{(k)}(x) = \sum_{i=1}^{n} A_i^{(k)}(x) y_i = \sum_{i=1}^{n} \psi_i(x), \tag{3.6}$$

where the footnote i represents data pair (x_i, y_i). Let $y = (y_1, y_2, \ldots, y_n)^T$ and $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)^T$. An general application of data sharpening method subject to constraints can be expressed as following,

$$min_{\tilde{y}} \sum_{i=1}^{n} \left(\frac{y_i}{n} - \frac{\tilde{y}_i}{n} \right)^2, \quad s.t. \quad \sum_{i=1}^{n} \psi_{\tilde{i}}(x_j) \geq 0, \quad 1 \leq j \leq N, \tag{3.7}$$

where $\tilde{i}$ denotes data pair $(x_i, \tilde{y}_i)$ and correspondingly we have

$$\tilde{m}^{(k)}(x) = \sum_{i=1}^{n} A_i^{(k)}(x) \tilde{y}_i = \sum_{i=1}^{n} \psi_{\tilde{i}}(x). \tag{3.8}$$

Notice that in equation (3.5), we only have one-sided constraints that $\sum_{i=1}^{n} \psi_{\tilde{i}}(x) \geq 0$. This is due to the fact that $l(x) \leq m^{(k)}(x)$ can be easily transformed into $m^{(k)}(x) - l(x) \geq 0$ and $m^{(k)}(x) \leq u(x)$ can be transformed into $-m^{(k)}(x) + u(x) \geq 0$. Hence condition expressed as in equation (3.5) can be cast into $\sum_{i=1}^{n} \psi_{\tilde{i}}(x) \geq 0$ with minor modifications of adding or subtracting some constants at x_i.

Since $\tilde{y}_i$ is a scalar, just like y_i, it can be expressed as a multiple of y_i that

$$p_i = \frac{\tilde{y}_i}{ny_i}. \tag{3.9}$$

Define $p = (p_1, p_2, \ldots, p_n)^T$, the objective function in equation (3.7) now becomes

$$min_p \sum_{i=1}^{n} \left(\frac{y_i}{n} - p_i y_i\right)^2, \quad s.t. \quad \sum_{i=1}^{n} \psi_i(x_j) p_i \geq 0, \quad 1 \leq j \leq N, \tag{3.10}$$

where $\sum_{i=1}^{n} \left(\frac{y_i}{n} - p_i y_i\right)^2 = \sum_{i=1}^{n} \left(\frac{1}{n} - p_i\right)^2 y_i^2$.

It is clear that traditionally p is a weight matrix coming from a density function such that $\sum_i^n p_i = 1$ and $p_i \geq 0$. CWB method relaxes this condition by allowing $p_i \leq 0$ and $p_i \geq 1$, maintaining $\sum_i^n p_i = 1$. One consequence is that for some "*outliers*", CWB might assign negative values which in turn needs to be compensated by assigning larger positive values to all the other points, even though other regions of the datum don't need to be adjusted. The condition requiring the summation of p_i equals to 1 interrelate every data pair $\{x_i, y_i\}$ together. It penalizes any kind of dramatic change of a small group of data pairs. However, data sharpening method doesn't restrain itself by this condition. Data sharpening alters the observations just enough to satisfy the desirable qualitative properties we would like to impose on our model. The adjustment of one particular point has no impact on the others. As we will show in later sections, data sharpening method tends to move fewer points than CWB method but with larger magnitude changes.

Before we continue to discuss the test statistics based on data sharpening, we shall first take a small detour to discuss the CWB estimator proposed by Racine et al. (2008). In their paper, instead of searching for a new set of response variable $\tilde{y}$ satisfying some null hypothesis, they select a new set of weights p to replace the uniform weights $1/n$ so that constraints are successfully imposed. To be more specific, a standard nonparametric Nadaraya-Watson estimator could be written as following

$$E(Y|x) = \frac{\sum_{i=1}^{n} y_i K_\gamma\left(\frac{y_i - x}{h}\right)}{\sum_{j=1}^{n} K_\gamma\left(\frac{y_j - x}{h}\right)} = \sum_{i=1}^{n} \frac{K_\gamma\left(\frac{y_i - x}{h}\right)}{\sum_{j=1}^{n} K_\gamma\left(\frac{y_j - x}{h}\right)} = \sum_{i=1}^{n} A_i(x) y_i, \tag{3.11}$$

where $K_\gamma(\cdot)$ is a generalized product kernel that admits both continuous and categorical data, γ represents a vector of bandwidths, and $A_i(x)$ is defined as follows

$$A_i(x) = \frac{K_\gamma\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K_\gamma\left(\frac{X_j - x}{h}\right)}. \tag{3.12}$$

Recall that $m(x) = E(Y|x)$, one can easily see that the conditional mean is a function of the uniform weights $1/n$, where

$$\hat{m}(x) = \sum_{i=1}^{n} A_i(x) y_i = n \sum_{i=1}^{n} A_i(x) \frac{1}{n} y_i. \tag{3.13}$$

As argued in their paper, one can impose a null hypothesis by choosing a different p vector to replace the uniform weight, denoted as $p_0 = (\frac{1}{n}, \ldots, \frac{1}{n})$, where $p = (p_1, p_2, ..., p_n)$ and

$$\hat{m}(x|p) = n \sum_{i=1}^{n} p_i A_i(x) y_i. \tag{3.14}$$

In practice, p is chosen according to minimization of a distance function

$$min_p \sum_{i=1}^{n} \left(\frac{1}{n} - p_i\right)^2 \quad s.t. \quad \sum_{i=1}^{n} p_i = 1 \quad and \quad \sum_{i=1}^{n} \psi_i(x_j) p_i \geq 0, \quad for \quad 1 \leq j \leq N, \tag{3.15}$$

where $\|\cdot\|$ is the L_2 norm and $\psi_i(x_j) = nA_i(x_j)y_i$.

As readers can see, the objective function of CWB method is solely depending upon vector p without involvement of the response variable. It has a natural interpretation of being a density, though they relaxed the condition $0 \leq p_i \leq 1$ to allow both positive and negative weights while retaining $\sum_{i=1}^{n} p_i = 1$. Because of this, CWB method has to forgo the power divergence metric used in biased-bootstrap method, see Hall & Huang (2001), since it only works for probability weights. Therefore, they turn to the well-know L_2 metric. The consequence of focusing on p vector in the objective function is that the change over p_i can has to be accompanied with changes over the rest of weights $p_j, j \neq i$,

$$\Delta p_i + \sum_{j=1, j\neq i}^{j=n} \Delta p_j = 0. \tag{3.16}$$

even though there is no reason to do so.

Data sharpening method, though the objective function can be expressed in a manner that is very similar as the CWB method, individual weights there not only can have both positive and negative weights, but also immune from the extra constraint of $\sum_{i=1}^{n} p_i = 1$. This enables data sharpening method to adjust the individual weights separately. Another benefit of considering y_i when we select a new set of weights in data sharpening is that the objective function tends to minimize the distance between sharpened $\tilde{y}$ and y. The inference thereafter has much smaller mean squared error which we will explain in details in later sections.

Next we define a distance measurement based on L_2 norm to be our test statistic

$$D(y, \tilde{y}) = (\frac{y}{n} - \frac{\tilde{y}}{n})^T(\frac{y}{n} - \frac{\tilde{y}}{n}) = \sum_{i=1}^{n} \left(\frac{y_i}{n} - p_i y_i\right)^2 = D(y, p). \tag{3.17}$$

3.3 Local Polynomial Regression Framework

In order to derive the asymptotic distribution of the proposed distance test statistic, we need to rely heavily on the notations introduced by Masry (1996*a*) and Masry (1996*b*), which considered the estimation of a multivariate regression function and its higher order derivatives.

According to Taylor's Theorem, we can approximate $m(z)$ by a multivariate local polynomial up to order p as following,

$$m(z) \approx \sum_{0\leq|k|\leq p} \frac{1}{k!} m^{(k)}(x) \, (z-x)^k \,, \tag{3.18}$$

where $K_h(\cdot)$ is a standard product kernel and $h = (h_1, \ldots, h_d)$ is the bandwidth, see Li & Racine (2007) for more details. Next, we solve the following multivariate weighted least squares with respect to b_k,

$$\min_{b_k(x)} \sum_{i=1}^{n} \left[y_i - \sum_{0\leq|k|\leq p} b_k(x) \, (x_i - x)^k \right]^2 K\left(\frac{x_i - x}{h} \right), \tag{3.19}$$

to get an estimate of $\hat{b}_k(x)$. Combined with equation (3.18), it is straight forward to see that $m^{(k)}(x) = k!\hat{b}_k(x)$.

The minimization of equation (3.19) gives us the following relationship:

$$t_{n,j}(x) = \sum_{0\leq|\mathbf{k}|\leq p} h^{|k|}\hat{b}_k(x) s_{n,j+k}(x), \quad 0 \leq |j| \leq p, \tag{3.20}$$

where

$$t_{n,j}(x) = \frac{1}{n}\sum_{i=1}^{n} y_i \left(\frac{x_i - x}{h} \right)^j K_h\left(\frac{x_i - x}{h} \right), \tag{3.21}$$

$$s_{n,j}(x) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i - x}{h} \right)^j K_h\left(\frac{x_i - x}{h} \right), \tag{3.22}$$

and

$$K_h\left(\frac{x_i - x}{h} \right) = \frac{1}{h^d} k\left(\frac{x_{i1} - x_1}{h_1} \right) \times \cdots \times k\left(\frac{x_{id} - x_d}{h_d} \right). \tag{3.23}$$

We follow the setup of Masry (1996a) where equation (3.20) can be cast into a matrix form. Using the same lexicographical order as in Masry (1996a), let $D_s = \begin{pmatrix} s+d-1 \\ d-1 \end{pmatrix}$ be the distinct d-tuples with $|k| = s$. The highest priority is given to last position so that $k_1 = (0, \ldots, 0, s)$ is the first one and $k_{D_s} = (s, 0, \ldots, 0)$ is the last element, there is a one-to-one map which can be denoted by g_k^{-1}, so that

$$\left(\boldsymbol{\tau}_{n,s}\right)_k = t_{n,g_s(k)}. \tag{3.24}$$

Define

$$\boldsymbol{\tau}_n = \begin{pmatrix} \boldsymbol{\tau}_{n,0} \\ \boldsymbol{\tau}_{n,1} \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\tau}_{n,p} \end{pmatrix}. \tag{3.25}$$

$\boldsymbol{\tau}_n$ is of dimension $D = \sum_{s=0}^{p} D_s \times 1$. Using the same lexicographical order to rearrange $h^{|k|}\hat{b}_k$ for $0 \leq |k| \leq p$, we have

$$\hat{\boldsymbol{\beta}}_n = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{n,0} \\ \hat{\boldsymbol{\beta}}_{n,1} \\ \cdot \\ \cdot \\ \cdot \\ \hat{\boldsymbol{\beta}}_{n,p} \end{pmatrix}. \tag{3.26}$$

Eventually, let $[S_{n,|j|,|k|}]_{(l,m)} = s_{n,g_{|j|}(l)+g_{|k|}(m)}$ so that $\mathbf{S}_{n,|j|,|k|}$ is a $D_{|j|} \times D_{|k|}$ matrix, then we have a $D \times D$ matrix given by

$$S_n = \begin{pmatrix} S_{n,0,0} & S_{n,0,1} & \cdots & S_{n,0,p} \\ S_{n,1,0} & S_{n,1,1} & \cdots & S_{n,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p,0} & S_{n,p,1} & \dots & S_{n,p,p} \end{pmatrix}. \tag{3.27}$$

Equation (3.20) now can be rewritten in the matrix form

$$\boldsymbol{\tau}_n(x) = \mathbf{S}_n(x)\hat{\boldsymbol{\beta}}_n(x). \tag{3.28}$$

As argued in Masry (1996a), S was proved to be positive definite with appropriate assumptions, therefore

$$\hat{\boldsymbol{\beta}}_n(x) = \mathbf{S}_n^{-1}(x)\boldsymbol{\tau}_n(x). \tag{3.29}$$

Note that the $r - th$ element of $\hat{\boldsymbol{\beta}}_n(x)$ represents an estimate of the derivative of $m(x)$ via the relationship

$$\left(\hat{\boldsymbol{\beta}}_n(x)\right)_r = \frac{h^{|k|}\hat{m}^{(k)}(x)}{k!}, \quad r = g_k^{-1}(k) + \sum_{s=0}^{|k|-1} N_s. \tag{3.30}$$

Recall that $\sum_{i=1}^n \psi_i(x_j)p_i \geq 0$, for $1 \leq j \leq N$. At this point, we are able to write out $\psi_i(\cdot)$ explicitly. There is one more notation we need to introduce. Define

$$\left(\phi_{i,|j|}(x)\right)_k = \frac{1}{n}\left(\frac{x_i - x}{h}\right)^k K_h\left(\frac{x_i - x}{h}\right), \tag{3.31}$$

then let $\left(\boldsymbol{\phi}_{|j|}\right)_k = \left(\left(\phi_{1,|j|}\right)_k, \dots, \left(\phi_{n,|j|}\right)_k\right)^T$, which is an $n \times 1$ vector. Also following the same lexicographical order, we have

$$\boldsymbol{\phi}_{|j|} = \begin{pmatrix} \left(\boldsymbol{\phi}_{|j|}\right)^T_{k_1} \\ \left(\boldsymbol{\phi}_{|j|}\right)^T_{k_2} \\ \cdot \\ \cdot \\ \cdot \\ \left(\boldsymbol{\phi}_{|j|}\right)^T_{k_{D_s}} \end{pmatrix}, \tag{3.32}$$

which is a $D_s \times n$ matrix, and a $D \times n$ matrix can be defined as following

$$\boldsymbol{\phi}_n = \begin{pmatrix} \boldsymbol{\phi}_0 \\ \boldsymbol{\phi}_1 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\phi}_p \end{pmatrix}. \tag{3.33}$$

Note that $\boldsymbol{\tau}_n(x) = \boldsymbol{\phi}_n(x)y$, then equation (3.28) can be rewritten as

$$\hat{\boldsymbol{\beta}}_n(x) = \mathbf{S}_n^{-1}(x)\boldsymbol{\phi}_n(x)y. \tag{3.34}$$

Let $\mathbf{M}_r$ denote the $r - th$ row of matrix $\mathbf{M}$. We already know $m^{(k)}(x) = \frac{k!}{h^{|k|}} \left(\hat{\boldsymbol{\beta}}_n(x)\right)_r$ for some $r = g_k^{-1}(k) + \sum_{s=0}^{|k|-1} N_s$. Now it is clear that

$$m^k(x) = \tilde{\boldsymbol{\psi}}(x)y, \tag{3.35}$$

where $\tilde{\boldsymbol{\psi}}(x) = \frac{k!}{h^{|k|}} \left(S_n^{-1}(x)\boldsymbol{\phi}_n(x)\right)_r = \frac{k!}{h^{|k|}} \left(S_n^{-1}(x)\right)_r \boldsymbol{\phi}_n(x)$. Now we can write out $\psi_i(x)$ in the following way

$$\psi_i(x) = \frac{k!}{h^{|k|}} \left(S_n^{-1}(x)\right)_r \left(\boldsymbol{\phi}_n^T(x)\right)_i^T y_i, \tag{3.36}$$

and

$$\tilde{\psi}_i(x) = \frac{k!}{h^{|k|}} \left(S_n^{-1}(x)\right)_r \left(\boldsymbol{\phi}_n^T(x)\right)_i^T. \tag{3.37}$$

3.4 Theoretical Properties

In this section, we derive the asymptotic distribution of the proposed test statistic of $D(y, p)$ in equation (3.17). First of all, we invoke the following conditions.

Assumption 3.4

(i) *$m(x)$ has continuous derivatives up to order $p+1$ in $\mathcal{D} \in \mathcal{R}^d$; $f(x)$ is continuous and nonvanishing in $\mathcal{D}$; the ϵ_i are independent and distributed with zero mean and variance σ_i, moreover, are independent of x_i.*

(ii) *$K \in L_1$ is bounded and $\|\nu\|^{4p} K(\nu) \in L_1$ and $\|\nu\|^{4p+d} K(\nu) \in L_1$ and $K(\nu) \to 0$ as $n \to \infty$. Additionally $K(\cdot)$ is a symmetric kernel function with compact support.*

(iii) *Let $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$. h also satisfies that $h = O(n^{-1/(d+2p+2)})$.*

(iv) *$N \to \infty$ as $n \to \infty$ and $N = O(n)$.*

(v) *Let $d_N = inf_{1 \leq j_1, j_2} \leq N|x_{j_1} - x_{j_2}|$ be the minimum distance between data points. As $n \to \infty$, we require $d_N \to 0$ and $h^{-1} d_N \to \infty$.*

Assumption 3.4(i) is necessary for the local polynomial regression function and its higher order derivatives to exist. It ensures the smoothness of $m(x)$ and $f(x)$. We admit the variance to be heteroscedasticitic but remains to be independent distributed. The independence between ϵ_i and x_i is actually equivalent to the independence between the noise term and $m(x)$. Assumption 3.4(ii) is required by the local polynomial fitting. The symmetric assumption with compact support is standard in the kernel regression literature. Assumption 3.4(iii) is about the rate of convergence associated with the bandwidths attached to the explanatory variables. It is higher than the standard rate where $h \propto n^{-1/n+4}$ because the higher polynomial fittings. The last two conditions are borrowed from Racine et al. (2008). The minimum distance between data points decreases slower than h so that the correlations between two different point estimates is zero as $n \to \infty$.

Let $\tilde{y}$ be the solution to equation (3.17) and $\tilde{p}$ to be the corresponding percentage change defined in the previous chapters. Next, we give the asymptotic distribution of $D(y, p)$ in the following theorem,

Theorem 3.1. *Under Assumption 3.4, as* $n \to \infty$*, we have*

$$\frac{n^3 h^{-1}}{k! C_{p,k}^2 \left(\Sigma_{m=1}^{M} m^{(k)}(x_m^*)\right)^2} \check{D}(y, p) \sim \chi^2(n, \lambda),$$

where $\check{D}(y, p)$ *and* $C_{p,k}$ *are defined in appendix A, and* λ *is the noncentrality parameter, which is also given in appendix A.*

The proof of Theorem 3.4 is presented in Appendix A.

3.5 Inference

Ideally, the asymptotic distribution of the test statistic can be used to determine the significant level. However, it might not be a good approximation for finite sample sizes, see Racine et al. (2008). Hence, we propose the following bootstrap approach to simulate the empirical distribution of the test statistic.

Recall that $D(p, y) = min_{\tilde{y}} \sum_{i=1}^{n} (y_i/n - \tilde{y}_i/n)^2$ subject to conventional constraints. Typically, we need to estimate the restricted model through data sharpening method to get $\tilde{y}$ and then reject the null if $(y - \tilde{y})^T (y - \tilde{y})$ is two large. The resampling approach that is used to generate the null distribution of the of $D(p, y)$ involves generating new data pairs (x_i, y_i^*) via *iid* residual resampling, which requires us to obtain $\tilde{y}$ first and take $(x_i, \tilde{y}_i)_{i=1}^n$ as if the original sample. Therefore, the resamples are generated under the null hypothesis. However, there is one problem in this approach. The model under consideration is $y = m(x) + \epsilon$. Any constraints that restrict higher order derivatives are in the form of $y^{(k)} = m^{(k)}(x) + \epsilon^{(k)}$, where according to the assumption listed above, we have $\epsilon^{(k)} \equiv 0$. Hence, $\tilde{y}$ are in fact sharpened from $m(x)$ rather than y. Not only $\tilde{y}$ satisfies the null hypothesis, but also it gets rid of the noise part. We can also express this in the following equation

$$\sum_{i=1}^{n}(y_i/n - \tilde{y}_i/n)^2 = \frac{1}{n^2}\sum_{i=1}^{n}(m(x_i) + \epsilon_i - \tilde{y}_i)^2 \tag{3.38}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}(m(x_i) - \tilde{y}_i)^2 + \frac{2}{n^2}\sum_{i=1}^{n}\epsilon(m(x_i) - \tilde{y}_i) + \frac{1}{n^2}\sum_{i=1}^{n}\epsilon_i^2, \tag{3.39}$$

where $\sum_{i=1}^{n}\epsilon_i^2$ is a constant and the expectation of $\sum_{i=1}^{n}\epsilon(m(x_i) - \tilde{y}_i)$ is zero. Therefore, for a finite sample size n, the minimization of $\sum_{i=1}^{n}(y_i/n - \tilde{y}_i/n)^2$ is equivalent of minimizing $\frac{1}{n^2}\sum_{i=1}^{n}(m(x_i) - \tilde{y}_i)^2$. $\tilde{y}$ can also be viewed as tilted from $m(x)$. In practice, we first estimate $\hat{m}(x)$ and then find the sharpened conditional expectation, denoted as $\tilde{m}(x)$.

The detailed procedures are described below,

1. Estimate (Y, X) using the nonparametric local quadratic fitting, obtain $Y = \hat{m}(x) + \hat{\epsilon}$;
2. Obtain the restricted conditional mean, $\tilde{m}(x)$ by data sharpening.
3. Record $D_{\hat{m}} = \parallel \hat{m}(x) - \tilde{m}(x) \parallel$, and denote $\tilde{\epsilon} = Y - \tilde{m}(x)$.
4. Resample $\tilde{\epsilon}$ to get ϵ^* and then obtain $Y^* = \tilde{m}(x) + \epsilon*$.
5. Estimate $\hat{m}^*(x)$ using (Y^*, X) nonparametrically, then imposing H_0 to get a new set of $\tilde{m}^*(x)$.
6. Record the distance $D_{\hat{m}^*} = \parallel \hat{m}^*(x) - \tilde{m}^*(x) \parallel$.

7. Repeat $(4)-(6)$ B times to obtain an empirical distribution of $D_{\hat{m}}$, and then calculate p-value according to the following equation,

$$p_B = 1 - Pr\left(D_{\hat{m}} > D_{m^*}\right) = \frac{1}{B}\sum_{i=1}^{B} I\left(D_{\hat{m}} > D_{m^*}\right), \tag{3.40}$$

   where $I(\cdot)$ is the indicator function. One rejects the null at level of α if $p_B < \alpha$.

The bootstrap procedure described above alters $\hat{m}(x)$ directly because we don't know the true mean $m(x)$. This is valid as long as $\hat{m}(x)$ converges to the true mean $m(x)$. Furthermore, as $n \to \infty$, the change of $\hat{m}(x)$ to $\tilde{m}(x)$ diminishes as if the null hypothesis is true, which means if the sample size is large enough and $D(\cdot)$ is practically zero, no further bootstrapping is necessary that one must accept the null, see Hall, Huang, Gifford & Gijbels (2001) for more details.

In the next section, we demonstrate how data sharpening method performs with different underlying DGPs and compare its behavior with CWB method by examining the absolute difference between the unrestricted conditional expectations and restricted ones. We also look at their percentage change by plotting p vector for both data sharpening and CWB methods.

3.6 Significant Testing of Interaction Effects

For what follows we simulate a number of DGPs and then impose a constraint that there are no interactions among explanatory variables. The reason of our particular interests on interaction effect is because economic theory rarely provides justification about how an explanatory variable interacts with other covariates. In practice, adding an additional cross-product term into the model is a common re-specification method that researchers attempt to obtain a better fit, but it has been criticized as ad hoc. Such kinds of modifications usually appear in the analysis of classical linear regression models. Meanwhile, other functional forms can also introduce interactions among different explanatory variables; for example, $sin(X_1 + X_2)$, where X_1 and X_2 are two independent random variables. Without explicit cross-product terms such as X_1X_2, the functional form makes them inseparable from each other. For the discussion of interaction effects in probit and logit models we direct readers to Ai & Norton (2003) where they clear the misunderstanding of interaction effects in nonlinear models. We define interaction effect as cross-partial derivatives of the response variable with respect to the corresponding covariates of interest. More specifically, we are interested in examining the following null hypothesis

$$\boldsymbol{\beta}_k(x_i) \equiv 0, \quad for \quad i \in (1, \ldots, n), \tag{3.41}$$

where $k = (\ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots)$. We can rearrange explanatory variables so that $k = (1, 1, 0, \ldots, 0)$. Therefore, $\boldsymbol{\beta}_k$ can be expressed as $\frac{\partial^2 y}{\partial x_1 \partial x_2}$.

Another reason for us to study interaction effect is that to the best of our knowledge, no formal test is proposed in the literature focusing on interaction effects exclusively, even though information of explanatory variables' significance can be tested either parametrically or nonparametrically, and this remains to be a source of frustration if one wants to make more convincing inferences. The testing of interaction effects potentially has a wide range of applications. First of all, it can lead us one step further to the true data generating process and make the inference more reliable. Second, various interaction effects exist in different economic scenarios, such as the analysis of production theory. A producer would like to know whether two inputs are interacting with each other. Another example would be wage equation. A typical question involved in wage equation is whether educational attainment interacts with years of working experience. Hence, it is of theoretical and practical interests that motivates the significance testing for interaction effect.

Moreover, the testing of interaction effect also has a perspective in the applications of nonparametric models. Unlike its parametric counterparts, where the convergence rate of an estimator is usually $\sqrt{n}$, the convergence rate associated with a fully nonparametric estimator is usually proportional to the inverse of the total number of covariates, for example, $n^{1/(4+d)}$, where n is sample size and d is the number of covariates. This is referred as the *curse of dimensionality.* In the presence of multiple explanatory variables, nonparametric estimators' performances deteriorate quickly as d increases while keeping the sample size fixed, see Stone (1980). One way to break down the curse of dimensionality is to use additive models which requires us to test interaction effect. Ideally, without any sign of interactions between any two covariates, one can use additive models $y = \sum_{d=1}^{q} f(x_d) + u$ instead of $y = m(x_1, \ldots, x_d) + u$ and substantially ameliorate the rate of convergence from $n^{1/(4+d)}$ to $n^{1/5}$, because in additive models each variable is considered independently rather than jointly when we are trying to determine the its relationship with respect to the response variable. In that sense, testing interaction effects is equivalent to testing separability among covariates. Sperlich, Tjostheim & Yang (2002) suggested a test statistic within the scope of additive models. However, additivity is a rather strong assumption that sometimes economists are not comfortable with. The logic here should be that test for interaction effect first then move to the assumption of additivity. See Lavergne & Patilea (2008) and references therein for other methods to reduce the curse of dimensionality.

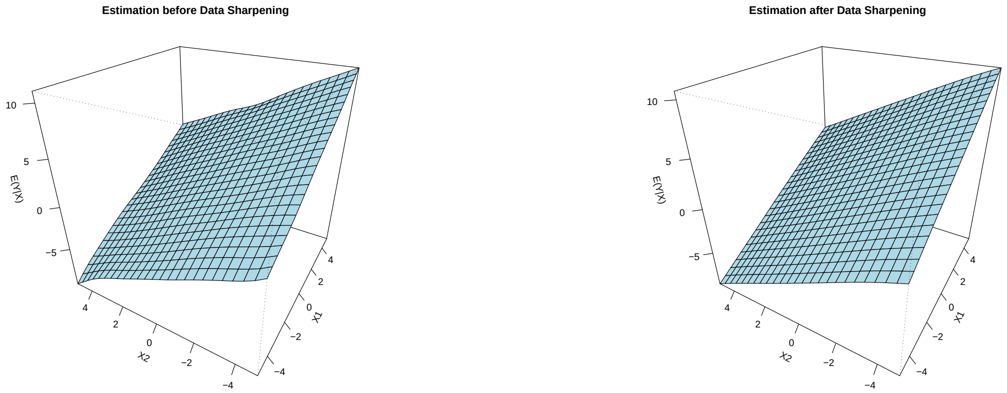
Using our method to test interaction effect has two advantages: First, nonparametric kernel smoothing method are robust for functional form mis-specification. It is well known that inference based upon mis-specified economic model could be misleading and severely undermine the conclusions. However, kernel smoothing methods is widely accepted by its reputation of robustness of functional form. Therefore, our distance test statistic is immune from functional forms. Moreover, we use the local polynomial regression framework developed by Masry (1996*b*) and Masry (1996*a*). In the case of testing interaction effects, where $|k| = 2$, we estimate the model by local quadratic fitting and measure the cross-partial derivatives by the magnitude of the coefficient of local cross product term, $x_1 x_2$. If the true DGP has no interactions among covariates, the coefficient should be zero everywhere. The second advantage of our method is the capability of capturing varying coefficients. This is rather important because for the same data sample, interaction effects can be positively significant in some part of the data space and insignificant or negatively significant in other regions. The data sharpening estimator examines the change of the unrestricted conditional mean estimates, therefore, it admits varying magnitude of interaction effects. While in the world of classical linear regression model, the coefficients are simply assumed to be constants even though there is no reason to do so. By admitting a changing marginal effect our method is more flexible compared with its parametric counterparts.

Next we present a few examples showing how data sharpening shifts $\hat{m}(x)$ under the null hypothesis of being interaction effect free. All examples consist of 100 observations with two explanatory variables $x_j \sim U(-5, 5)$, where $j \in \{1, 2\}$ and $\nu \sim N(0, .1)$. We consider the following DGPs:

$$
\begin{aligned}
&DGP_{3.1} : y = 1 + x_1 - x_2 + \nu,\\
&DGP_{3.2} : y = 1 + x_1 - x_2 + x_1 * x_2 + \nu,\\
&DGP_{3.3} : y = 1 + x_1 - x_2 + x_1^2 - x_2^2 + x_1 * x_2 + \nu,\\
&DGP_{3.4} : y = sin\left(\frac{2(x_1 + x_2)}{pi}\right) + \nu.
\end{aligned}
$$

$DGP_{3.1}$ is a simple linear model without any interaction effects, while $DGP_{3.2}$ adds a cross product term to $DGP_{3.1}$, and $DGP_{3.3}$ includes more second order terms which can be considered as noises when we try to capture interaction effects. $DGP_{3.4}$ is a nonlinear function that interaction effect exists everywhere with changing signs and magnitudes.

For what follows, we put the unrestricted conditional mean in the left sub-figure and put the sharpened restricted conditional mean under the null hypothesis in the right sub-figure. When we compare the performances of data sharpening and CWB methods, the absolute change and percentage change of conditional mean before and after imposing the constraint by the data sharpening method are put in the left sub-figure, and those selected via CWB method are put in the right sub-figure.


Estimation before Data Sharpening

Estimation after Data Sharpening

Figure 3.1: $DGP_{3.1} : y = 1 + x_1 - x_2 + \nu$.

Figure 3.1 shows the case of $DGP_{3.1}$ where no interaction effect exists and the null is true. The estimates $\hat{m}(x)$ before data sharpening is pretty much a plain surface with almost constant slopes in both directions. There are some minor ups and downs on the surface coming from the random component of the DGP, which are completed wiped out after the null is imposed by data sharpening. The restricted surface becomes flatter than it was because all the dents are smoothed out. However, the surface only changed slightly in order to strictly impose the null hypothesis. As we mentioned before, if we let $n \to \infty$, $\hat{m}(x)$ converges to its true mean $m(x)$ and these minor dents will be automatically removed by the estimation. This provides the evidence that data sharpening methods will leave the conditional estimates untouched if the null is true.

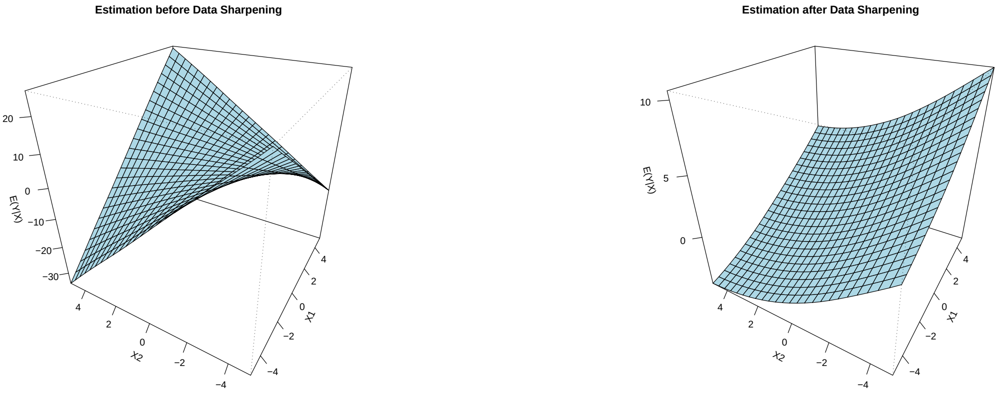


Figure 3.2: $DGP_{3.2} : y = 1 + x_1 - x_2 + x_1 * x_2 + \nu$.

$DGP_{3.2}$ illustrates an example with interaction effects by adding a cross-product term $x_1 x_2$ to $DGP_{3.1}$. The additional cross-product term stretches and curves the surface of $DGP_{3.1}$ into two directions, $x_2 = x_1$ and $x_2 = -x_1$. By looking at the surface at the left sub-figure, it is clear that interaction effect exists and the null hypothesis is false. As we discussed, the change over the surface of the conditional means before and after data sharpening could be dramatic and so it is. We notice that, after imposing the null, data sharpening gives us a surface with some sort of curvature rather than a straight surface. The newly selected conditional means capture the existence of higher order terms in the underlying DGP. This is very informative in a sense that practitioners can make well educated guess the underlying DGP is not linear. Recall that we only control the presence of interaction effect, leaving all the other terms to shift freely. The contrasting shapes here tell us that the null should be rejected.

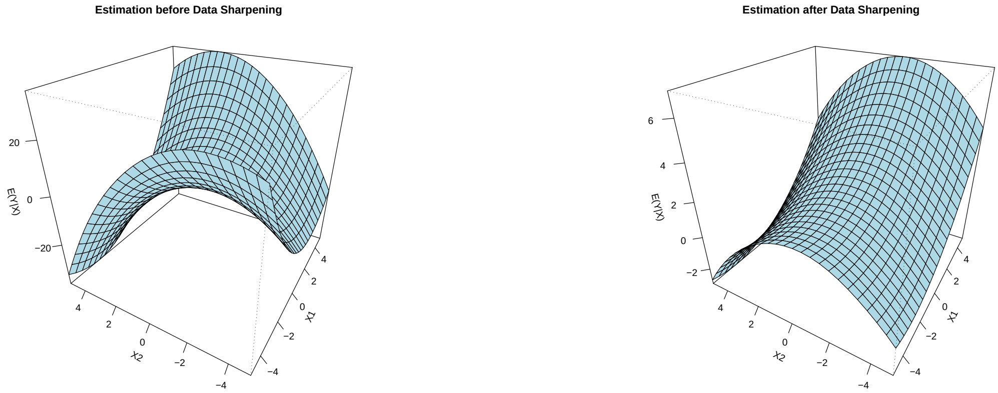


Figure 3.3: $DGP_{3.3} : y = 1 + x_1 - x_2 + x_1^2 - x_2^2 + x_1 * x_2 + \nu$.

Next let us consider $DGP_{3.3}$, with $(x_1^2 - x_2^2)$ added into $DGP_{3.2}$. The unrestricted surface looks like a twisted saddle and right after the null is imposed by data sharpening what we see is that the second order terms are reserved and the saddle shape is tuned to remove interaction effect.

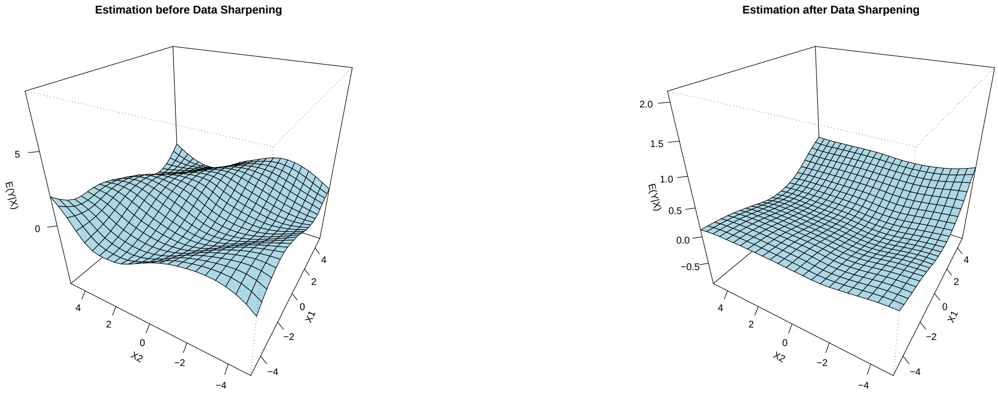


Figure 3.4: $DGP_{3.4} : y = sin\left(\frac{2(x_1+x_2)}{pi}\right) + \nu$.

Lastly, $DGP_{3.4}$, shown in Figure 3.4, is different from all the above three DGPs because of its high frequency. The unsharpened shape looks like a sea wave with interaction effect everywhere. Unlike the previous $DGPs$, it is hard to imagine what the surface would be after the constraint is imposed. As it turns out, data sharpening simply pushes the surface down almost to be a constant. This reminds us that the simplest case for no interaction effects is that the response variable is a constant that it does not depend on any of the covariates. From now on, we would suspect the true DGP has high frequency if the sharpened estimates is a constant.

What is also worth to point out besides the change of the shape is the considerable shrinking of the restricted conditional mean's variance when the null is false. There are two reasons: Firstly, when the null hypothesis only requires no interaction effect, data sharpening will always have a solution which is a constant that can be set to the mean of $\hat{m}(x)$, especially for high frequency $DGPs$, where interaction effect exists everywhere and the magnitude and signs are changing from point to point. In fact, for most of the constraints that can be expressed as restrictions on the conditional means and higher order derivatives, a constant selected by data sharpening can always be considered as a backup answer. Secondly, the objective function in data sharpening requires $\tilde{m}(x)$ to deviate from $\hat{m}(x)$ as little as possible. Therefore, as we will show later on, data sharpening method tends to have smaller mean squared errors compared with CWB method.

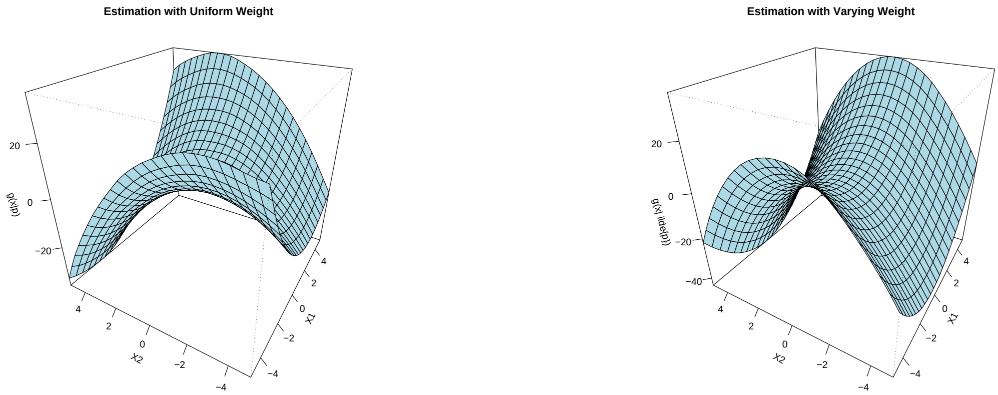


Figure 3.5: $DGP_{3.3} : y = 1 + x_1 - x_2 + x_1^2 - x_2^2 + x_1 * x_2 + \nu.$

Now we take $DGP_{3.3}$ as an example to compare the performances of data sharpening method we proposed and the constrained weighted bootstrap estimator introduced by Racine et al. (2008). The right hand side of Figure 3.5 shows that the null hypothesis is successfully imposed and the restricted conditional mean gives a much steerer saddle shape than the one shown in Figure 3.3. The variance of the restricted conditional mean is much larger when the null is imposed by CWB method. The residuals, $\tilde{u} = Y - \tilde{m}(x)$, systematically oversized compared with the ones from data sharpening method, especially near the edge of the data space. For example, take the point $x_1 = x_2 = 4$ and we see that $\hat{u} \simeq 40$, which is almost two times as from data sharpening. Hence, CWB method results in poorer mean squared errors.

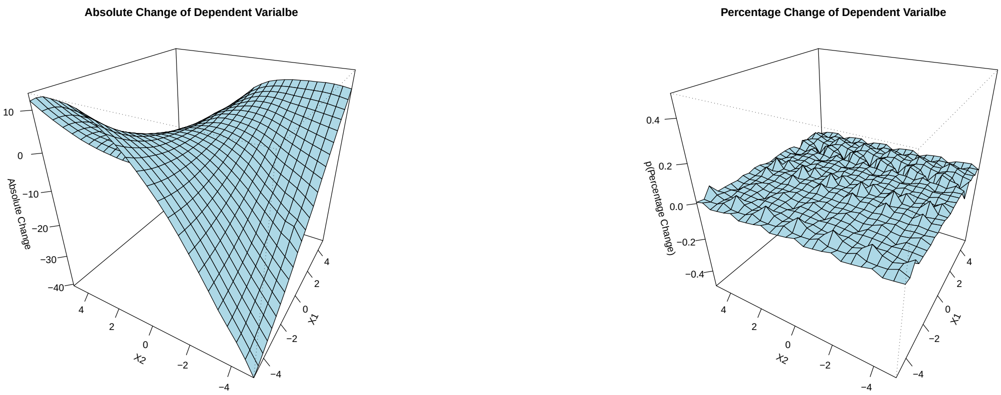


Figure 3.6: $DGP_{3.3}$ Absolute Change v.s. Percentage Value Change

For the purpose of better illustration of their differences, we plot out the absolute change of $\hat{m}(x)$, defined as $\big(\hat{m}(x) - \tilde{m}(x)\big)$, on the left hand side of Figure 3.6, and the percentage change of $\hat{m}(x)$, captured by p on the right hand side of Figure 3.6 for the constrained nonparametric estimator.

We observe that even though the change of p is small for every point estimate, only in decimal levels, the restricted conditional mean is substantially and systematically different from the unrestricted conditional estimates. This is expected when the null is false. CWB method tends to shift points up and down uniformly, due the building mechanism that $\sum_{i=1}^{n} p_i = 1$. As we discussed previously, if one point is changed because of imposing a constraint, all other points need to adjust their position in order to satisfy this condition. The objective function in CWB method limits the change of p to the minimum degree, and consequently the mean squared error is larger. Next we turn to take a look at the results from data sharpening methods.

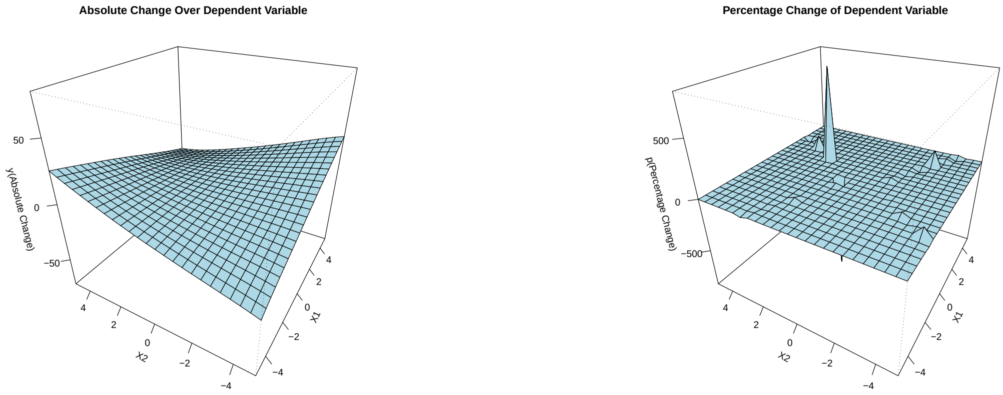


Figure 3.7: $DGP_{3.3}$ Absolute Change v.s. Percentage Change (Data Sharpening Estimator)

On the left hand side of Figure 3.7, we plot the absolute change of the conditional mean before and after imposing the constraint that there is no interaction effect. The percentage change, depicted by p, is shown on the right hand side of Figure 3.7.

By using data sharpening method, the absolute change of the conditional mean has a much smaller variance compared with the CWB method. This is consistent with the design mechanism that data sharpening is minimizing the distance between $\hat{m}(x)$ and $\tilde{m}(x)$. [1] However, in terms of the percentage change, data sharpening method clearly shift a few points immensely leaving most of the other points untouched. This is plausible in a sense that the patterns of the points shifted can tell researchers information about the underlying DGP.

[1] For all the other DGPs discussed in the paper, we attach all the relevant graphics in appendix B, where we find consistent results.

It becomes more obvious if we take a look at Figure 3.8.

In Figure 3.8, we show the case for $DGP_{3.4}$, where a high frequency one is considered. Interestingly, the absolute change of the conditional mean behaves just like a *sin* wave. More importantly, the percentage change pic3ture shows that only points along a few straight lines are adjusted. Compared with CWB method shown in Figure 3.9, which uniformly all observations, the pattern of those lines consisted with the sharpened points can better inform us about the underlying data generating process.

3.7 Monte Carlo Simulations

In this section, we present Monte Carlo Simulations to examine the finite sample performances of the proposed test statistics. We consider the following $DGPs$,

$$DGP^*_{3.2} : y = 1 + x_1 - x_2 + \beta * x_1 * x_2 + \nu, \tag{3.42}$$

$$DGP^*_{3.4} : y = sin\left(\frac{2(x_1 + \beta * x_2)}{pi}\right) + \nu, \tag{3.43}$$

where $x_i \in N(0, 1)$ and $\epsilon \in N(0, 0.45)$. The first DGP is a simple linear function and the second one is a high frequency data generating process. The magnitude of β controls the degrees of interaction effects. We choose sample sizes $n = 100, 200, 400, 800$. The total number of iterations in each Monte Carlo Simulation is 1000 and the number of the bootstrap within each iteration is 399. Bandwidths are selected by $sd(X)n^{-1/8}$, where $sd(X)$ represents the standard deviations of X. We examine the significant level at 1%, 5% and 10%.

Table 3.1 shows the size and power of the proposed two test statistics according to $DGP^*_{3.2}$ and $DGP^*_{3.4}$. Let's first take a look at the size. The size for both $DGPs$ becomes reasonable

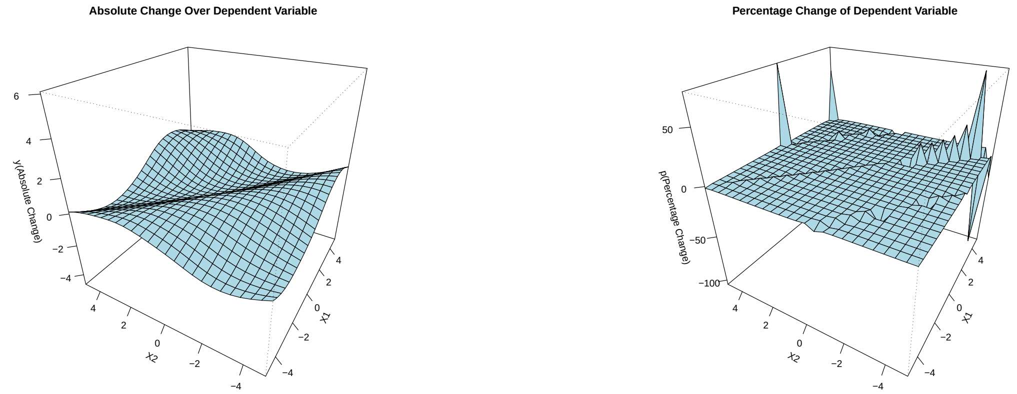


Figure 3.8: $DGP_{3.4}$ Absolute Change v.s. Percentage Change (Data Sharpening Estimator)

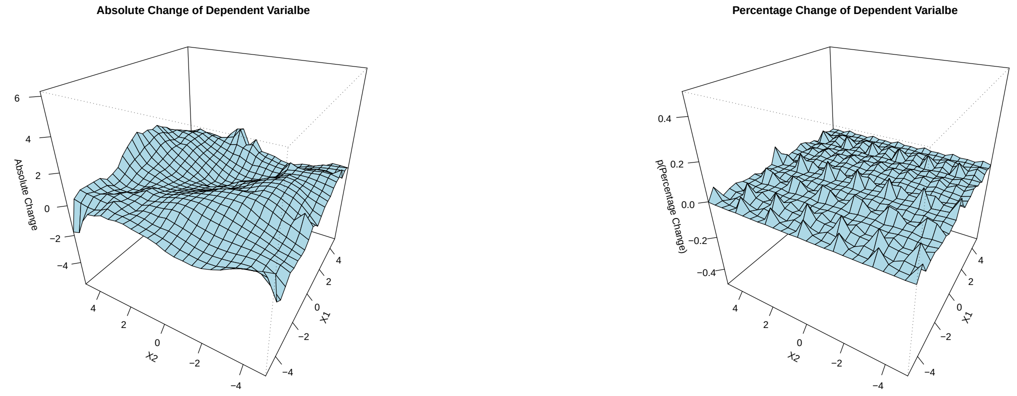


Figure 3.9: $DGP_{3.4}$ Absolute Change v.s. Percentage Change

Table 3.1: Size and Power of Data Sharpening Estimator under $DGP^*_{3.2}$ and $DGP^*_{3.4}$

$DGP^*_{3.2}$	$n = 100$			$n = 200$			$n = 400$			$n = 800$		
Sig Level	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\beta = 0$	0.042	0.093	0.173	0.019	0.083	0.149	0.018	0.075	0.115	0.015	0.051	0.101
$\beta = .1$	0.163	0.275	0.303	0.229	0.445	0.583	0.501	0.779	0.887	0.945	0.986	0.990
$\beta = .5$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\beta = 1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$DGP^*_{3.4}$	$n = 100$			$n = 200$			$n = 400$			$n = 800$		
Sig Level	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\beta = 0$	0.014	0.077	0.140	0.013	0.060	0.117	0.011	0.053	0.098	0.017	0.074	0.134
$\beta = .1$	0.013	0.082	0.148	0.015	0.062	0.130	0.012	0.060	0.110	0.020	0.070	0.147
$\beta = .5$	0.051	0.161	0.270	0.094	0.250	0.382	0.246	0.488	0.629	0.649	0.839	0.903
$\beta = 1$	0.322	0.568	0.698	0.737	0.885	0.930	0.987	0.998	0.998	1.000	1.000	1.000

well when sample size increases to 400. Fixing the sample size and looking at each column, we see that as the degree of interaction effects goes up, the power goes up quickly. Also the power increases when the sample size becomes larger. Even $\beta = .1$, which is a small deviation from the null hypothesis, the proposed test statistic is capable to detect it. Moreover, the power is higher in the simple linear case compared with high frequency DGP.

3.8 Empirical Examples

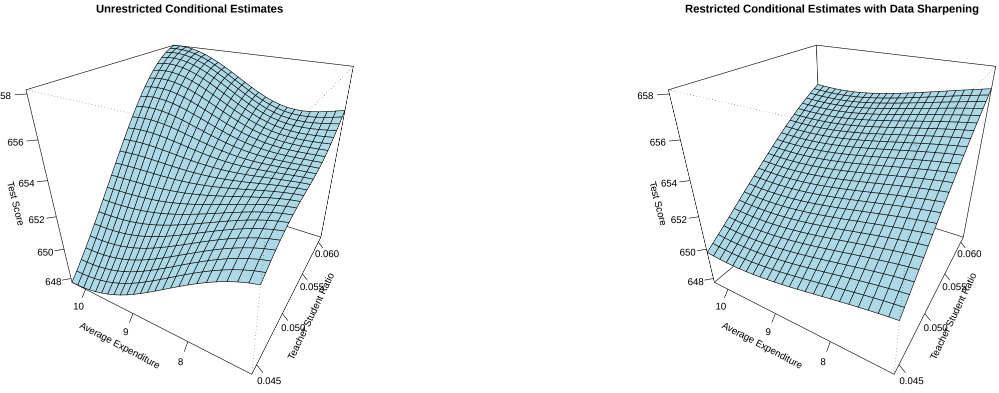
In this section, we apply the proposed test statistics to a dataset which contains test scores on the Standford 9 Standardized test administered to 5th grade students in California. School characteristics(average across the district) includes enrollment, number of teachers, number of computers per classroom, and expenditure per students. The dataset consists of 420 observations[2]. However, to keep things simple, we only consider the number of teachers and the average expenditure to be the relevant variables. We would like to test if there is some interaction effect on the students' test score. However, instead of using the absolute number of teacher, we use the teacher student ratio in our analysis. The test score is defined as the average of math score and reading score. The descriptive statistics for Teacher Student Ratio(TSR) and Average Expenditure(AE) are shown in the upper part of Table 3.2. The lower part of Table 3.2 reports the results for the distanced based test statistic and its corresponding p value. The p-val indicates a strong evidence of interaction effects.

Figure 3.10 shows some curvature over certain regions of the data space. The magnitude of cross-partial derivatives varies from point to point. The restricted conditional mean has constant first order derivatives almost everywhere, suggesting a significant change of the shape.

[2] The dataset is available in Applied Econometrics with R(AER) package. For other detailed information, we direct readers to AER package.

Table 3.2: Descriptive Statistics of Teacher-Student Ratio and Expenditure Example

Descriptive Statistics.	
Name of the Variable	(Mean, Std)
Student Test Score	(654.15, 19.05)
Teacher Student Ratio	(0.0514, 0.0051)
Average Expenditure	(5312.408, 633.937)
Testing Results.	
$(D(\cdot)$, p-val)	(8255.521, 0.005)


Unrestricted Conditional Estimates

Restricted Conditional Estimates with Data Sharpening

Figure 3.10: An Empirical Example of Students' Test Score.

Without formal testing, according to the experience we have at this point, it is naturally to suspect that interaction effect exists here. The test statistic confirms our intuition.

3.9 Appendix A

This appendix provides proofs of Theorem 3.1. It consists of two parts, the first part introduces some results from Masry (1996*a*) and then we proceed to prove the asymptotic distribution of the proposed test statistic.

Proof. For $\mathbf{x}^* = (\mathbf{x}^*_{\mathbf{1}}, \ldots, \mathbf{x}^*_{\mathbf{M}})^{\mathbf{T}}$, equation (3.34) becomes

$$\big(D^{\mathbf{k}} m\big)(\mathbf{x}^*) = \begin{pmatrix} \tilde{\boldsymbol{\psi}}(\mathbf{x}^*_{\mathbf{1}}) \\ \tilde{\boldsymbol{\psi}}(\mathbf{x}^*_{\mathbf{2}}) \\ \cdot \\ \cdot \\ \cdot \\ \tilde{\boldsymbol{\psi}}(\mathbf{x}^*_{\mathbf{M}}) \end{pmatrix} \mathbf{y} = \tilde{\boldsymbol{\psi}}(\mathbf{x}^*)\mathbf{y}. \tag{3.44}$$

Under the Assumption 3.1(i)-(iii), see Masry (1996*a*), define

$$\mu_{\mathbf{j}} = \int_{R^d} \mathbf{u}^{\mathbf{j}} \mathbf{K}(\mathbf{u}) d\mathbf{u}. \tag{3.45}$$

S_n converges as following

$$S_n(\mathbf{x}) \to \boldsymbol{\Lambda} f(\mathbf{x}) \quad \text{as} \quad n \to \infty, \tag{3.46}$$

where the $D \times D$ matrix $\mathbf{\Lambda}$ is defined as

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{(0,0)}, & \mathbf{\Lambda}_{(0,1)}, & \dots & \mathbf{\Lambda}_{(0,p)} \\ \mathbf{\Lambda}_{(1,0)}, & \mathbf{\Lambda}_{(1,1)}, & \dots & \mathbf{\Lambda}_{(1,p)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Lambda}_{(p,0)}, & \mathbf{\Lambda}_{(p,1)}, & \dots & \mathbf{\Lambda}_{(p,p)} \end{pmatrix}. \tag{3.47}$$

$\mathbf{\Lambda}_{(i,j)}$ is a $D_i \times D_j$ dimensional matrix, whose (l, m) element is $\mu_{g_i(l)+g_j(m)}$.

Next, we proceed from the minimization problem defined in 3.17,

$$min_{\mathbf{p}} \sum_{i=1}^{n} \left(\frac{y_i}{n} - p_i y_i \right)^2, \quad s.t. \quad \sum_{i=1}^{n} \psi_i(\mathbf{x}_j) p_i \geq 0, \quad 1 \leq j \leq N. \tag{3.48}$$

Invoking the Karush-Kuhn-Tucker conditions so that

$$2\left(p_i - \frac{1}{n}\right) y_i^2 + \sum_{j=1}^{N} \lambda_j \psi_i(\mathbf{x}_j) = 0, \quad i = 1, \dots, n. \tag{3.49}$$

$$\lambda_j \sum_{i=1}^{n} \psi_i(\mathbf{x}_j) p_i = 0, \quad j = 1, \dots, N. \tag{3.50}$$

$$\sum_{i=1}^{n} \psi_i(\mathbf{x}_j) p_i \geq 0, \quad \lambda_j \geq 0, \quad j = 1, \dots, N. \tag{3.51}$$

Denote $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, according to equation (3.49),

$$p_i = \frac{1}{n} - \frac{1}{2y_i^2} \sum_{j=1}^{N} \lambda_j \psi_i(x_j). \tag{3.52}$$

Using some permutation strategy, we can rearrange λs to obtain δs so that $\delta_j > 0$ for $j = 1, ..., M$ and zero for $j = M+1, ..., N$. Also permuate $(x_1, ..., x_N)$ accordingly, the above equation becomes

$$p_i = \frac{1}{n} - \frac{1}{2y_i^2} \sum_{j=1}^{M} \delta_j \psi_i(x_j^*). \tag{3.53}$$

From equation (3.50), it is natural to see that $\sum_{i=1}^{n} \psi_i(x_k^*) p_i = 0$, combined with equation (3.53),

$$\sum_{i=1}^{n}\psi_i(x_k^*)\left(\frac{1}{n}-\frac{1}{2y_i^2}\sum_{j=1}^{M}\delta_j\psi_i(x_j^*)\right)=0.$$

After some deductions, we have

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\psi_i(x_k^*) &= \sum_{j=1}^{M}\delta_j\sum_{i=1}^{n}\frac{\psi_i(x_k^*)\psi_i(x_j^*)}{y_i^2}\\ &= n\sum_{j=1}^{M}\delta_j\sigma_{jk}^{(n)} \qquad (3.54)\end{aligned}$$

where

$$\sigma_{jk}^{(n)}=\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_i(x_k^*)}{y_i}\frac{\psi_i(x_j^*)}{y_i}=\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_i(x_k^*)\tilde{\psi}_i(x_j^*). \qquad (3.55)$$

Let $\bar{\psi}(x_j)=\frac{1}{n}\sum_{i=1}^{n}\psi_i(x_j)$, equation (3.54) essentially tells us the following relationship, $\boldsymbol{\delta}=\frac{2}{n}\boldsymbol{\Sigma}_n^{-1}\bar{\boldsymbol{\psi}}(\mathbf{x}^*)$.

$$\begin{aligned}D\left(\mathbf{y},\mathbf{p}\right) &= \sum_{i=1}^{n}\left(\frac{1}{2}\boldsymbol{\delta}^{\mathbf{T}}\boldsymbol{\psi}_i(\mathbf{x}^*)\right)^2\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\bar{\boldsymbol{\psi}}^{\mathbf{T}}(\mathbf{x}^*)\boldsymbol{\Sigma}_{\mathbf{n}}^{-\mathbf{1}}\boldsymbol{\psi}_i(\mathbf{x}^*)\right)^2\\ &= \frac{1}{n^2}\sum_{i=1}^{n}T_i^2, \qquad (3.56)\end{aligned}$$

where $T_i=\boldsymbol{\psi}_i^T(\mathbf{x}^*)\boldsymbol{\Sigma}_{\mathbf{n}}^{-\mathbf{1}}\bar{\boldsymbol{\psi}}(\mathbf{x}^*)$.

Now recall that $T_i=\boldsymbol{\psi}_i^T(\mathbf{x}^*)\boldsymbol{\Sigma}_{\mathbf{n}}^{-\mathbf{1}}\bar{\boldsymbol{\psi}}(\mathbf{x}^*)$.

$$\begin{aligned}\sigma_{lm}^* &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_i(x_l^*)\tilde{\psi}_i(x_m^*) \qquad (3.57)\\ &= E\left(\tilde{\psi}_X(x_l^*)\tilde{\psi}_X(x_m^*)\right) \qquad (3.58)\\ &= E\left(\frac{(k!)^2}{n^2h^{2|k|}f(x_l^*)f(x_m^*)}\sum_{0\le|\mathbf{j}_1|\le p}\boldsymbol{\Lambda}_{r,\mathbf{j}_1}^{-1}\phi_{i,|\mathbf{j}_1|}(\mathbf{x}_l^*)\sum_{0\le|\mathbf{j}_2|\le p}\boldsymbol{\Lambda}_{r,\mathbf{j}_2}^{-1}\phi_{i,|\mathbf{j}_2|}(\mathbf{x}_m^*)\right), \qquad (3.59)\end{aligned}$$

for $m\neq l$, we have $\sigma_{lm}^*=0$.

$$\sigma_{mm}^*=\frac{(k!)^2C_{p,k}^2}{n^2h^{2|k|}}, \qquad (3.60)$$

where $C_{p,k} = \sum_{0\leq|\mathbf{j}|\leq p} f^{-1}(x_m^*)\boldsymbol{\Lambda}_{r,\mathbf{j}}^{-1}E\left(n\phi_{i,|\mathbf{j}|}(\mathbf{x}_m^*)\right) = \sum_{0\leq|\mathbf{j}|\leq p}\frac{\mu^{|\mathbf{j}|}}{\mu^{|\mathbf{j+k}|}}$. Here we see that Σ is a diagonal matrix. Note that $\bar{\boldsymbol{\psi}}(\mathbf{x}^*)$ are consistent estimates of $m^{(k)}(\mathbf{x}^*)$, and under Assumptions 3.4(i)-(v), we have $nh^{d/2+|k|}\psi_i(\mathbf{x}^*)\bar{\boldsymbol{\psi}}(\mathbf{x}^*)$ are independent random variables with means $\check{\mu}_i = \Sigma_{j=1}^M \frac{\psi_i(x_j^*)m^{(k)}(x_j^*)}{\sigma_{mm}^*}$ and variances $\check{\sigma}_i^2 = \left(\Sigma_{j=1}^M \frac{m^{(k)}(x_j^*)}{\sigma_{mm}^*}\right)^2 \sigma_i^2$. Therefore we have the following equations

$$\lambda = \Sigma_{i=1}^n \left(\frac{\check{\mu}_i}{\check{\sigma}_i}\right)^2 = \Sigma_{i=1}^n \left(\frac{\Sigma_{j=1}^M \psi(x_j^*)m^{(k)}(x_j^*)}{\Sigma_{j=1}^M m^{(k)}(x_j^*)\sigma_i}\right)^2. \tag{3.61}$$

Next we define

$$\check{D}(y,p) = \frac{1}{n^2}\sum_{i=1}^n \left(\frac{T_i}{\sigma_i}\right)^2. \tag{3.62}$$

Thus, we have

$$\frac{n^3h^{-1}}{k!C_{p,k}^2\left(\Sigma_{m=1}^M m^{(k)}(x_m^*)\right)^2}\check{D}(y,p) = \Sigma_{i=1}^n Z_i^2 \sim \chi^2(n,\lambda), \tag{3.63}$$

where Z_i is independent variables with means $\check{\mu}_i$ and variances $\check{\sigma}_i^2$. □

3.10 Appendix B

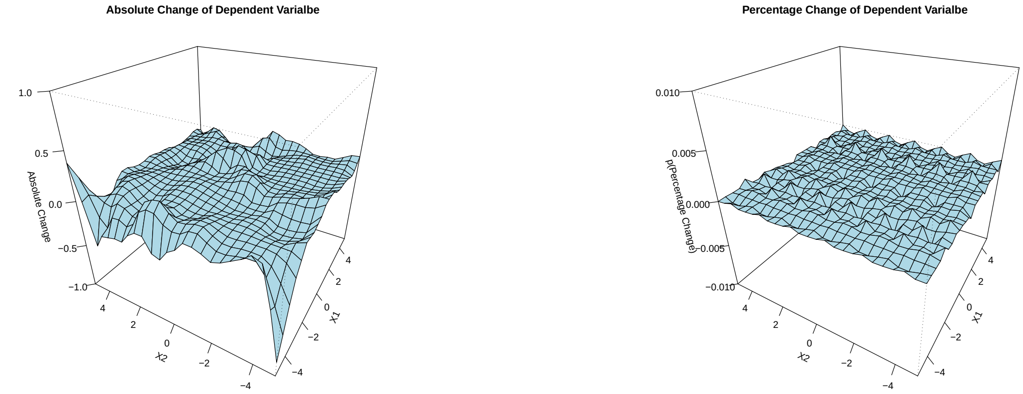


Figure 3.11: $DGP_{3.1}$ CWB Percentage Change v.s. Absolute Change

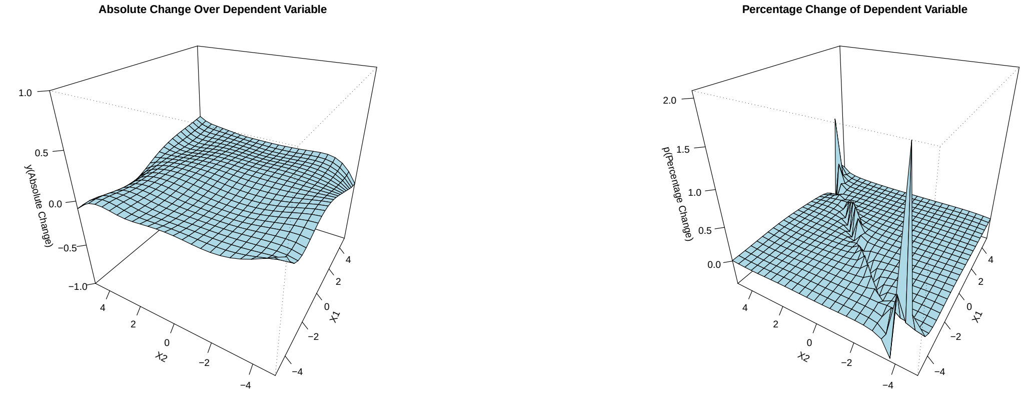


Figure 3.12: $DGP_{3.1}$ Sharpening Percentage Change v.s. Absolute Change

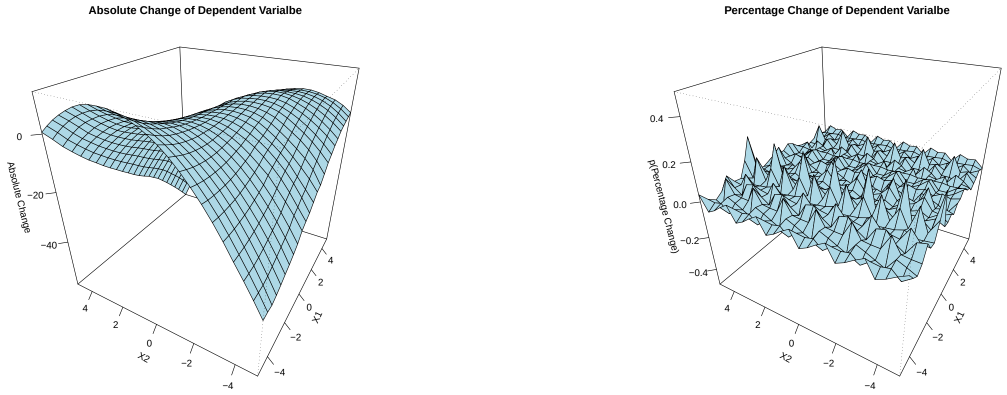


Figure 3.13: $DGP_{3.2}$ CWB Percentage Change v.s. Absolute Change

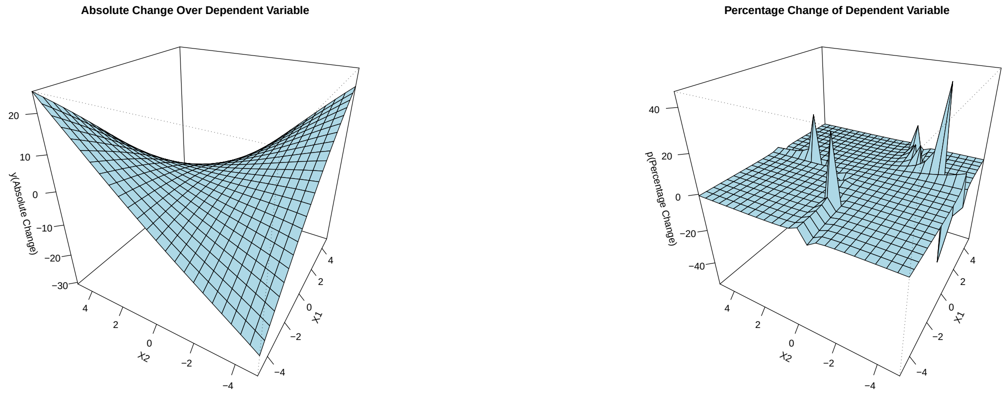


Figure 3.14: $DGP_{3.2}$ Data Sharpening Percentage Change v.s. Absolute Change

Chapter 4

Income Inequality and Health

Income Inequality and Health in China: A Dynamic Approach Accounting for Health Selection Process

(ABSTRACT)

We re-examine the relationship between income, income inequality and health outcome using China Health and Nutrition Survey(CHNS) data. The health selection problem is addressed by the dynamic probit panel model proposed by Heckman (1981*a*) and Heckman (1981*b*). The causations between income, income inequality and health status are explored. The individual level inequality shows a persistent significant effect on health in our sample. The community level inequality measured by Gini coefficients proves irrelevant in the context of health dynamics. The results from CHNS data support the weak version of the relative income hypothesis; the strong version is rejected. Furthermore, the long run income level is more important than the short run income variation. The policy implication here is that reducing the individual level inequality and lifting the level of investments on public goods can improve the population's health.

4.1 Introduction

The rising income inequality in China since the economic reform has caused tremendous concern for economists. To date, there have been many studies examining reasons and decompositions of the expanding inequality. However, the strong association between the public health status and income has not been fully investigated in terms of their directions of causation, let alone the role of income inequality. In short, low income can cause health problems, yet poor health leads to low income. This paper is intended to answer the question whether income inequality affects health outcomes independently, even when dual causations between income and health exist. We use the framework proposed by Heckman(1981a,1981b) to overcome the initial condition problem due to the health selection process; otherwise, the estimates of income and income inequality coefficients on health might be overstated.

There has been an ongoing discussion of the role of income with respect to health outcomes. The *absolute income hypothesis*, positing that a higher income level yields a better health outcome, has become a consensus view. At a very basic level, income provides necessities for health, such as housing, healthy diet, medical care, and a good environment with minimum exposure to harmful hazards. Low income restricts households' access to material resources. For instance, inadequate housing is correlated with respiratory illness and overcrowding increases the risk of infections and leads to low hygiene standards. Also, poverty limits people's food choice and increases the probability of developing an appetite in favor of high calories with low nutrition. Unfortunately, this appetite can become a life-long habit and increase the risk of cardiovascular diseases, see Judge & Paterson (2001) for more details. Since the absolute income hypothesis is well documented, we focus our attention on the relative income hypothesis.

However, the many debates over income inequality and health still leave us with an unresolved controversy regarding the so called *relative income hypothesis*, which states that income inequality is a destructive factor for individual health. The arguments focus on the role of income inequality. Mixed empirical results for different countries have been obtained. Many studies have shown a strong correlation between life expectancy and income inequality, using aggregate level data. This could be simply due to the fact that there is a nonlinear relationship between health and absolute income. If health is a concave function with respect to absolute income, a society with more equality will enjoy a higher health standard compared to a society with the same level of average income but more inequality, since the gradient is steeper among the poor and it diminishes as income increases. Redistribution of income from the rich to the poor can improve the overall health status, because the improvement of the poor's health will exceed the loss of the rich(Lynch, Smith, Kaplan & House 2009). Nonetheless, the effect of inequality itself on individual health cannot be fully addressed by merely exploring the relationship between aggregate income inequality and health outcomes. The concavity of health over income can be estimated by including an additional second order term of income into our analysis. The results for our sample show that individual level inequality does affect people's health as an independent source. Improving a household's socio-economic status by reducing its relative deprivation index turns out to be almost as important as raising the long run income level. We also confirm that inequality acts just like the absolute income level: the rate of return of improving one's health through moving toward a higher position in the income distribution is diminishing, and eventually flattens out. It is important that policy makers in China treat income and income inequality equally important in order to promote the population's health. A higher average income level will benefit the society's health as a whole only if inequality is under control. Otherwise, the effectiveness of a policy that attempts to improve public health by increasing aggregate income in China is doubtful and the facts need to be scrutinized..

4.1.1 The Relative Income Hypothesis

Kawachi & Kennedy (1999) used U.S. state level data and provided an in-depth discussion of several pathways how inequality can affect people's health. One is the psychological stress caused by income inequality. An inferior position in the income distribution may exert a powerful and consistent pressure on a household, see Wilkinson (1999), especially for people who compare their living standards with others from the same area. The anxiety and insecurity created by social comparison with the surroundings lead to an increased probability of self-destructive behaviors, such as smoking and drinking, see Li & Zhu (2004). Secondly, inequality can reduce investments on public goods. That is because rich families' interests diverge from those with lower incomes due to the fact that public investments have less effect on the rich. Compared to the poor, rich families usually have the ability to cope with deficiencies caused by low levels of public investments, however, they bear the burden of higher share of tax revenue. On the other hand, poor families with lower shares of tax revenue rely heavily on public services. The disparities in their interests are getting bigger when income inequality is rising, and this constantly puts pressure on lowering taxes and reducing public services, because of the clout of the elite, see Krugman (1996). Thirdly, inequality can undermine social cohesion and lead to erosion of social capital, which can be measured as the degree of mistrust, levels of reciprocity, and crime rate.

The relative income hypothesis has two versions. The strong version states that income inequality deteriorates health outcomes for both the rich and the poor; the weak version emphasizes the negative effect of income inequality on the poor's health outcomes, and it has little to do with the health status of the rich, see Mellor & Milyo (2002) for more details. In our paper, we test both the strong and weak version of the relative income hypothesis, using CHNS data.

4.1.2 Inequality's Positive Effect

In view of the explanations given above, it seems reasonable to assume that income inequality has a negative effect on individuals' health. However, that is only one side of the coin. Judge & Paterson (2001) provided insights about the possibility of positive effects that income inequality might have on health. Consider the following scenario. There are two communities, both of which have the same average income level but different income distributions. The extreme case would be one community has no inequality that the income is uniformly distributed and the other community is polarized into a group of rich families and a group of impoverished ones. With a progressive tax rate system, the latter community generates higher tax revenue compared to the first one. The availability of funds for public investments is greater in the latter community. By elevating the levels of public goods, such as public education, transportation, medicaid, medicare, and health facilities, the latter community could enjoy a higher health status among its residents. Secondly, high income households tend to demand for more health promoting technological products than poor households. The latter community has stronger incentives to invent, promote, or simply introduce high tech medical equipments, which may eventually benefit the poor.

We would like to add another possible pathway. Namely, a moderate level of inequality is perhaps necessary to stimulate the economy to grow in per capita terms. Perfect equity constrains people's motivations, since then the system lacks the merit or compensation mechanism to recognize individuals' contributions and efforts. Economic growth could slow down with zero tolerance of inequality. In the long run, a community that allows a certain degree of inequality can enjoy higher average income level. The connection between higher income and better health has been well established. In our sample, we find that the average income is indeed rising as well as inequality. There could be a possible tradeoff between reducing inequality and raising income. In our paper, we use CHNS data to empirically test the relative income hypothesis as well as the likely constructive influences that inequality might have on health outcomes.

4.1.3 Health Selection Process

In order to properly estimate the effects of income and income inequality on health, we need to consider the situation that poor health leads to low income at the beginning time period, which is referred as the health selection process. Many medical conditions reduce people's working hours or even make them lose their jobs. People who are unable to work due to illness will suffer from financial difficulties. Families which have disabled members or live with severe medical conditions need to cope with the drop of the income that could otherwise increase the overall wellbeing of the household. When exploring the relationship between income, income inequality, and health, neglect of the health selection process can exaggerate the effects of income and inequality on health. In order to investigate the dual causal relationship between income and health, the time sequence of events regarding health and income needs to be observed and longitudinal data is necessary to produce convincing results.

4.1.4 A Dynamic Approach

Static cross-sectional studies are only able to pin down the association between income and health. The health selection process contaminates any results that study the effect of income on health. In the current paper, a dynamic framework controlling the initial condition problem, developed by Heckman(1981a, 1981b), is used for the following reasons. First, poverty itself is a dynamic process and people do move in and out of poverty from time to time. The time sequence of events, whether sickness happens first or poverty happens first, can help us to precisely estimate the relationship between income, income inequality, and health. The establishment of this temporal order can provide information of the causal direction between them rather than a simple association. Secondly, income level or income inequality at one specific point of time might be a poor marker for measuring a household's ability to access material resources. The results in this paper using CHNS data show that a household's average income level over time matters, and that the income variation from year to year does not affect people's health outcomes. That is to say, the long run income level is far more important than the short run income variations. Benzeval & Judge (2001) used the average income to approximate the true income level and investigated the time dimension between health and income. Most studies listed in their paper, with the aim at discovering the role of time dimension in the relationship between health and income, however, only examined developed countries, in particular by using longitudinal data sets from the USA, Canada, West Germany and Sweden. Half of the 16 papers listed in Benzeval & Judge (2001) used mortality rates and only two of them included the long run income level as well as the current income variation. Benzeval & Judge (2001) confirmed that the long run income level was more significant than any other temporary changes of income on health. In our paper, we use the Heckman(1981a, 1981b) framework to study the time dimension of health and income in China. The health selection problem is properly addressed by the initial condition problem involved in the dynamic process. The present of health selection process at the beginning time period can cause a low income. And low income reduce the probability of reporting a good health status in the next time period. The bad circle can carry over to the following years, which is captured by the health dynamic framework. In our results, the effect of income on health is attenuated but not diminished when we properly incorporate the health selection process into the health dynamic process. And the health selection process persistently affects our sample's time sequence of health outcomes. Moreover, we answer the question that whether income inequality itself is a detrimental factor to the sample's health. In order to reveal the effect of income inequality on health, we measure income inequality in both community level and individual level to provide a full picture of the connections between income inequality and health outcomes.

4.1.5 The "Open Door Policy" and Provincial Medical Resources.

The Chinese government introduced the economic reform called the "Open Door Policy" in December 1978, which was designed to transform the impoverished centrally planned economy into a market economy. Before that, the state-owned enterprizes covered all sectors of industry, and agriculture was practiced in collective farms. Everyone was employed by the government and the wage rate was fixed by the central government. There were neither financial markets nor private businesses. The reform of opening-up policy progressed through two stages. The first stage focused on agriculture, the opening-up of the country to foreign investments and giving permission to private businesses, leaving most industries state-owned. In the second stage, from the late 1980s, many stated-owned companies were privatized with only a few public monopolies left, and the economy experienced remarkable growth. The private sector seemed to be the most rapid in terms of the economic growth. The economic reform turned out to be successful in that it changed society considerably and reduced poverty massively. However, it resulted in a high level of income inequality, which was almost absent prior to the economic reform.

Since we restrict our sample to individuals aged 20 or above in the year 1997, all the individuals were born before 1977, prior to the time of the "Open Door Policy". The number of hospital beds per thousand of people of each province in 1979 had no association with the economic development. Before the economic reform, the allocation of medical resources was not market-oriented whatsoever. The number of hospital beds of each province was not tied with the people's income, but it reflected the availability of medical resources that was accessible for the public. In this sense, the number of hospital beds per thousand of people of each province can be used to signify the level of social investments in public health. The medical facilities were centrally planned for each province and equally accessible for all individuals. Therefore, it captures our sample's earlier lives' health information, and the initial endogeneity problem between individuals' health outcomes and their unobserved heterogeneities can be controlled.

The rest of the paper is organized as follows. Section 2 describes the data. Section 3 discusses the model. Section 4 presents the results. Section 5 contains conclusions and proposes future research directions.

4.2 The CHNS Data

In our paper, we use the data from China Health and Nutrition Survey. CHNS is an ongoing international collaborative project between the Carolina Population Center at the University of North Carolina at Chapel Hill and the National Institute of Nutrition and Food Safety at the Chinese Center for Disease Control and Prevention. Households from nine provinces were randomly collected from Liaoning, Hei Longjiang, Shandong, Jiangsu, Henan, Hubei, Hunan, Guangxi, and Guizhou. For each province, two cities and four counties were selected, and one county-town was chosen in each city and three villages from every county were chosen. From now on, we use community to refer county-town and villages. For each community, approximately 20 households were sampled. For each household, demographic characteristics as well as health related information were recorded, i.e., self-reported health status(SRHS), physical conditions, and health behaviors. The availability of SRHS was restricted to the years of 1997, 2000, 2004, and 2006. In our paper, the age of subjects in the sample is restricted to 20 or above before 1997, in that sense, all participants were born before the “Open Door Policy” economic reform. The reason to exclude children from our analysis is that their health status depends much on their parents’ income. Our sample consists of 2876 individuals. For each individual, there are four observations.

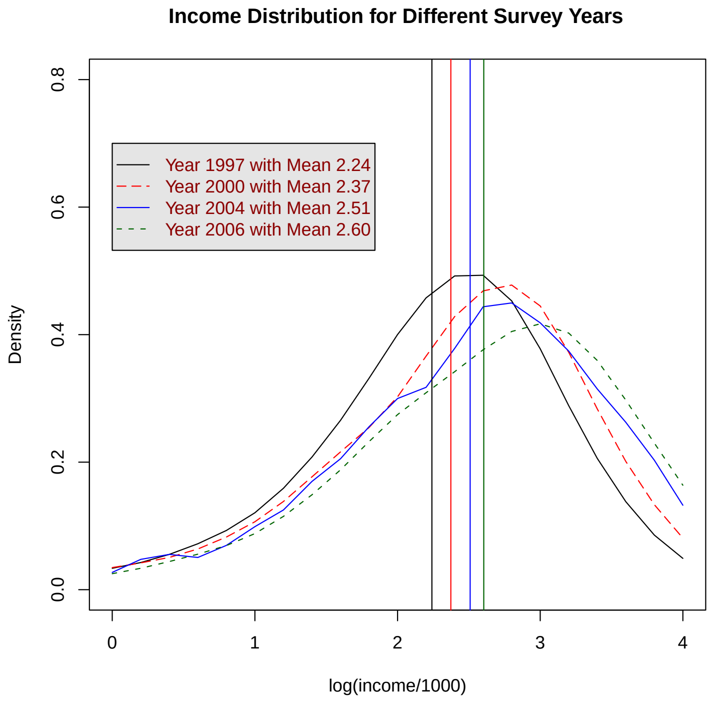


Figure 4.1: Increasing Income Inequality with Rising Income Level

Picture 1 shows income distributions of the sample for different survey years. The horizontal axis represents the log of household income in terms of thousands of Ren Min Bi(RMB). The vertical lines illustrate the mean income level of these distributions. They are matched by colors. It is clear that the average income in our sample was climbing over the years. Comparing the rich located at the higher percentiles of the income distributions to the poor positioned at the lower percentiles, the picture tells us that this increasing average income level was largely due to the improvements for the rich. The poor's absolute income remained still for the past 14 years. From another perspective, the poor were even poorer since their relative income measured by the share of the population's total income was shrinking. Also, the variance of the income distribution was certainly increasing since the right hand side tail became flatter. Clearly, income inequality was rising very fast.

It seems like the economic development favored the rich over the poor. Our question is whether this increasing income inequality in China would affect individuals' health as an independent source. If income inequality doesn't matter, the whole sample should not experience health disparity over the time, since their absolute income remained at the same level. Take a look at table 4.1 and 4.2, where we define the relevant variables and provide their summary statistics. SRHS is an indicator measuring health status with 1 being healthy and 0 otherwise. The probability of reporting a good health status in our sample dropped from 0.7211 in 1997 to 0.5549 in 2006. The biggest fall-off was from year 2000 to 2004. Also, the variance became greater from 0.4485 in 1997 to 0.4970 in 2006. Health status in our sample was also deteriorating over the time. Next we begin to explore their relationship.

4.2.1 Variables Involved

Our sample consists of 2876 individuals with 4 observations in 1997, 2000, 2004 and 2006 for each one of them. Gender is an indicator, with 1 being male. 46.38% of our sample are males. Since the gender ratio didn't change over the years, we simply put a dot for the following years, and we apply the same rule for URBAN. URBAN is another indicator for households' residence with 1 being urban and 0 being rural. 24.4% of our sample lived in urban. NumBH was the number of hospital beds per thousand of people for each province in 1979, measured in thousands. The group of variables measuring households income levels are HHINC[1], HHINC.MEAN, HHINC.DVI. They were all adjusted in terms of thousands of RMB according to the CPI in year 2000. For each year, we have the following equation,

$$HHINC_{it} = HHINC.MEAN_i + HHINC.DVI_{it}, \tag{4.1}$$

, where $i = 1, ..., N$, representing different individuals, and $t = 1, 2, 3, 4$, standing for a particular year. $HHINC_{it}$ represents income for individual i's household income at time t. $HHINC.MEAN_i$ is the average household income over the years[2]. $HHINC.DVI_{it}$ is the deviation of $HHINC_{it}$ from $HHINC.MEAN_i$. In this way, we decompose a household income into two parts, the long run household income level and the short run income variation. In our sample, the mean of $HHINC$ increased about three thousand RMB every survey year. During the same time, variance of $HHINC.DVI$ also got larger. Income inequality was really growing. The enlarging inequality in income distribution was also confirmed by an expanding Gini coefficient. Gini coefficient calculates community level inequality. In each community, 20 households were sampled. Gini coefficient represents the overall income inequality within a community. In 1997, the average Gini coefficient was 0.343. In 2006, it reached 0.41. The variance of Gini coefficient was relatively stable. The overall inequality on the community level increased. RDA and RANK are defined in the following. EDUC is years of education. HHSIZE is the number of family members. MARR is the marriage status with 1 for being married and 0 otherwise.

[1] The individual level income turns out to be insignificant for health outcomes. Households usually make inter-transfers when facing medical conditions. Therefore, it is more suitable to determine individuals health outcomes.

[2] For the mean income level of each household, we use the average of households income histories, which includes all survey years from 1989. It gives us better estimation of the real household income level.

4.2.2 Measuring Inequality.

Following Eibner & Evans (2005), we construct a relative deprivation index[3] that measures a household's relative income within a community. There were 20 households randomly selected in each community. Other households living in the same community can be regarded as neighbors and their living standards are directly compared to the household's own living standard. The overall level of inequality is picked up by the Gini coefficients within each community. The individual level inequality is calculated with respect to their "neighbors", in our case, it is based on the surrounding 19 households. If income inequality itself is a factor that deteriorate individuals health, the higher percentile of a household in the income distribution within its community, the better their health outcomes are. The relative deprivation of a household's absolute income, following Eibner & Evans (2005) and Yitzhaki (1979), is defined in the following

$$RDA_i = \frac{\Sigma_{j=1}^{N_k} (Y_j - Y_i)}{N_k}, \qquad \forall Y_j > Y_i, \tag{4.2}$$

where N_k is the number of households in community k, which equals to 20 in most cases. Y_i is the household i's income. RDA calculates the average of income differences between household i and other households who have higher incomes. Variable $RANK$ is defined to be the centile of a household's position in the income distribution within a community. $RANK$ ignores the magnitude of income differences, however, it picks up the order of households' socioeconomic status. Higher value of $RANK$ means less relative deprivation for a household and higher value of RDA means greater degree of relative deprivation. The lowest value for RDA is 0 and $RANK$ is within the range of 0 and 1.

The weak version of relative income hypothesis can be tested by the significance of RDA

[3] For other relative deprivation indices defined in Eibner & Evans (2005), we found they acted the same as the one we used in our paper.

and $RANK$. The significance of these two variables indicates that moving up in the income distribution will increase the probability of reporting a good health outcome. And the results from our sample confirm that individual level inequality plays a role on determining individuals health outcomes. On the other hand, Gini coefficient reflects the overall inequality within a community. It is the same for every household regardless of the levels of income. If the strong version of relative income hypothesis is true, then the coefficient of Gini index should be a negative number. Otherwise, the strong version doesn't hold for the data.

4.3 The Statistical Model

In this paper, we use Heckman(1981a,1981b) estimator, which is the dynamic probit model with the initial condition problem. The health status is a latent variable, denoted as H^*_{it},

$$H^*_{it} = \gamma H_{it-1} + X^{'}_{it}\beta + \alpha_i + \epsilon_{it}, \tag{4.3}$$

$$H_{it-1} = \begin{cases} 1 & \text{if } H^*_{it-1} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Here $i = 1, ..., N$, representing for different individuals; $t = 1, ..., T$, denoting the time dimension. X_{it} is a vector, including individual demographic characteristics, household characteristics, and community level variables. β is the vector of coefficients we are interested in. H_{it-1} is the preceding time period's health outcome, with 1 being healthy and 0 being unhealthy. α_i is the unobserved individual characteristic that might correlate with H_{it-1} and a subset of X_{it}. ϵ_{it} is assumed to be independent and identically distributed from a normal distribution. Equation 4.3 says that an individual health status depends on the health outcome in the last time period as well as a group of demographic variables.

The standard random intercept model assumes α_i is independent with all other explanatory variables. In our case, the unobserved individual characteristic can be considered as initial endowments for individual i, which correlates with long term income level and certainly affects health status.

Following Mundlak (1978), we assume that

$$\alpha_i = w_i a + \nu_i, \tag{4.4}$$

where w_i is the average of household income throughout time. ν_i is independent with w_i. Now equation 4.3 can be expressed as

$$Prob\left[H_{it} = 1 | H_{it-1}, X_{it}, \alpha_i\right] = \Phi\left(\gamma H_{it-1} + X'_{it}\beta + w_i a + \nu_i\right), \tag{4.5}$$

where Φ is the CDF of a standard normal distribution. However, w_i can be easily incorporated into X_{it}, then equation 4.5 looks just like equation 4.3. The only adjustment needed for replacing the household income with w_i is to add the short run income variations, which is the decomposition of income shown in equation 4.1. Therefore, we continue our discussion with functional form of equation 4.3. However, we run regressions in both cases to see the impacts on results by decomposing household income into the long run income level and the short run income variations.

The process of health dynamics is represented by γ . The previous health status affects current health outcome. The direct pathway is due to the nature that a medical condition can last for years and the recovery process can be long. With H_{it-1} and X_{it} together, the probability of reporting a poor health is broken down into two sources, the previous health outcome and other demographic variables, including household income level. The dynamics of health incorporates health history so that the estimates of coefficients of income and income inequality on health are more accurate. However, at the beginning time period where $t = 0$, the unobserved individual characteristic might correlate with initial health status. A bad draw of α_i can substantially increase the probability of reporting poor health at the beginning time and it continues to alter the path of health and income histories through γ, because health problem can substantially reduce individuals' ability to cope with work. Working hours are expected to be shorten and performances at work are presumed to be worse. Therefore, the health selection process is expressed by the endogeneity between α_i and H_{i0}. Heckman(1981a, 1981b) considered the initial condition problem and proposed the following solution, where

$$Prob\left[H_{i0} = 1|\alpha_i\right] = \Phi\left[Z'_{i0}\lambda + \theta\alpha_i\right]. \tag{4.6}$$

Z_{i0} typically includes X_{i0} and additional exogenous variables, which can capture population's health status in our case. We use $NumBT$ for the reasons listed above[4]. The endogeneity between α_i and H_{i0} can be tested by the null hypothesis of $\theta = 0$.

Now equations 4.3 and 4.6 together complete the model and the likelihood function for individual i is

$$L_i = \int \left(\Phi\left[(Z'_i\lambda + \theta\alpha_i)(2H_{i0} - 1)\right] \prod_{t=1}^{T} \Phi\left[(\gamma H_{it-1} + X'_{it}\beta + \alpha_i)(2H_{it} - 1)\right] \right) dF(\alpha_i), \tag{4.7}$$

where $F(\cdot)$ is the CDF of α_i. We assume the unobserved individual characteristic α_i has a normal distribution and equation 4.7 can be evaluated by Gaussian-Hermite quadrature,

[4] We also include other explanatory variables in equation 4.6. For instance, we use coastal indicator to show whether a household locates in the east costal area. The east coastal province demonstrated faster economic development compared to inland provinces. However, they are not significant in most of the regressions.

see Butler & Moffitt (1982). Hence, for a random sample of individuals, the maximized likelihood is given by the following

$$L = \Phi_i^N \int_{\alpha_i^*} \left(\Phi\left[(Z_i'\lambda + \theta\sigma_{\alpha_i}\alpha_i^*)\,(2H_{i0} - 1) \right] \prod_{t=1}^{T} \Phi\left[(\gamma H_{it-1} + X_{it}'\beta + \sigma_{\alpha_i}\alpha_i^*)\,(2H_{it} - 1) \right] \right) dF(\alpha_i^*), \tag{4.8}$$

where F is the cumulative distribution function of $\alpha_i = \sigma_{\alpha_i}\alpha_i^*$

4.4 Results & Discussion

In our analysis, we run three groups of probit regressions and compare the effects of income inequality on health. The first group goes through the traditional cross-sectional analysis which ignores the time dimension; The second group considers the longitudinal nature of our sample; however, it disregards the potential initial condition problem that we have discussed in the previous section; The last group utilizes the framework developed by Heckman(1981a, 1981b), in which the initial condition problem is properly accounted. Each group consists of ten regression functions, according to different measurements of income inequality.

4.4.1 The Role of Other Demographic Variables

Before we step into details of the effect of income and income inequality on the population's health, we notice that some variables are universally significant, such as age, gender and household size. In terms of the influence on people's health, age comes as the first important factor. The aging process weakens the samples' health. This is properly due to the fact that aging is highly associated with many medical conditions, especially chronic diseases. Gender is another variable that affects the health outcomes for all three groups of regressions.

Speaking of gender in China, males generally report better health outcomes than females. Lastly, household size is another important factor that affects health substantially. Unlike foundings in Li & Zhu (2004), results in our analysis indicate that a large family size has a negative effect on individuals' health. The possible explanation could be that the available resources for each family member are reduced due to a larger household size. Consider the case of two families with the same level of income, the one with more family members is inferior compared to the other in terms of their average level of consumption on health related products. As for other variables, such as years of education, urban indicator, and marriage status, they are generally insignificant for individuals health outcomes.

4.4.2 The Role of Income and Income Inequality

Table 3 and 4 show the results of the first group of regressions with 10 columns, which takes our sample as cross-sectional observations. Table 5 and table 6 present results for the probit panel model. We assume that there is no correlation between individual health and the unobserved characteristics at the beginning time period. Table 7, 8, 9 and 10 illustrate results for the dynamic probit panel model with the initial condition problems taken into consideration. First, column 1 in each group shows household income is always significant. The absolute income hypothesis is supported for the sample that an increasing income can improve health.

In column 2 of each group, we add RDA into regressions which measures individual level inequality. The presence of RDA reduces the magnitude of the coefficient of income on health by approximately 50% for all cases. It is found that RDA squared is also positively significant, and this indicates that shrinking income disparity at individual level can improve the population's health status, but at a decreasing rate. Also, the effect of RDA on health is equally important as income level. This is the evidence that income inequality is an independent source for individuals health outcomes and a redistribution system can improve the population's health. If 1000 RMB is transferred from the richest household to the poorest one without altering their income ranking, so that the richest household is still on the top of the income distribution and the poorest family stays where it is. The loss for the richest household is only coming from a lower income level, since its RDA is still 0. However, the gain for the poorest family comes from the increment of income level and the decreasing RDA. Therefore, the loss is less than the gain.

Column 3 of each group adds Gini coefficient, which is the community level of inequality, into the regression functions. We find that it turns out to be insignificant when we incorporate health dynamics into our analysis and control the initial condition problem. The significance of Gini and Gini squared in cases of cross-sectional probit regressions and probit panel regressions, saying that strong version of the relative income hypothesis holds when Gini coefficient is less than 0.4 approximately. After that threshold, the community level of inequality seems to have a positive effect on population's health. This is probably because the enlarging community level inequality in our sample is largely due to the fact that the rich is getting richer. The poor's real income level is untouched. The negative effect of inequality on the poor is already captured by the individual level inequality, RDA. The possible positive effects of community level income inequality, which come from the increasing tax revenue and expanding scale of investments on public goods, overcome the negative forces it might brought as we discussed above. Besides, the rising average income level will certainly increase the consumption of health related products in China, which leads to an increasing demand for health technologies. The gain exceeds the loss if a expanding community level inequality can result in a higher level of investments on public goods. However, when we look at the third group of regression functions, Gini and Gini squared are insignificant. We will discuss it when we talk about health dynamics.

The role of $SRHS_{it-1}$, which is the previous health outcome for a individual, is designed to take the health selection process into consideration. Poor health in the prior time period is likely to lead to a low income, and the consequence of low income could transfer into poor health in the present time period. In short, poor health leads to poor health. As shown in table 7 and 9, $SRHS_{it-1}$ is always significant and the presence of it makes Gini coefficient irrelevant. The possible pathway is that at the time of the first wave, the average age in our sample was already 44.75 years. Their initial health statuses, reported in 1997, were the results of long time accumulation of their earlier lifetime's health outcomes, which should be highly correlated with the level of public investments prior to 1979 due to the reasons discussed above. We use the number of hospital beds per thousand of people at the province level in 1979 to capture the level of public investments, since the economic reform did not start rolling. In this way, we control the initial health status by using $NumBT$. Gini coefficient becomes insignificant. The results in table 8 and 10 verify that $NumBP$ is indeed highly correlated with individuals' health.

Next, we take a look at the role of $RANK$. $RANK$ measures the order of a household's income within a community. In column 4 and 5 of each group, we see that $RANK$ is so powerful that $HHINC$ doesn't even matter anymore. It signifies the socio-economic status within a neighborhood very well so that health only depends on the order of their income. However, if we take a look from column 6 to 10 of each regression group, where income is decomposed into long run income level and short run variations, we can see that the household average income level is still significant with the presence of RANK. As a matter of fact, long run income level is always significant for all cases. The deviations usually don't play a role. The household's ability to deal with diseases is determined by the long run income level. The weak version of relative income hypothesis holds in China.

There are other variables, which measures the monetary and time cost of going to hospitals, are included in our original analysis. However, they turn out to be insignificant most of the time. The cost of transporting to medical facilities are so small in China that they don't change our results. Also variables of household environments are also excluded from our reports due to the same reason. The rising income makes them irrelevant.

4.4.3 Robust Check on Migrations

Throughout our empirical analysis, we find that the type of household registration (urban or rural) is insignificant most of the time. However, some family members move in and out from a household and their registration type might change over time. In China, policies are usually in favor of urban areas, where proficient doctors and high quality medical facilities are available, as well as better social infrastructures. Therefore, we replace the household registration type with individual registration type to check if we will find the results turned over. The results are not reported here, but what we find is that individual registration type is still insignificant in all settings. This is probably because rich families, regardless of their urban or rural status, can search for higher standards of medical attentions and therefore circumvent the disadvantages brought by registration type. Another possible explanation could be the role of SHRS. SHRS is subjective measurement of health status. People living in urban areas usually have higher expectations of their health conditions because all aspects of their daily life are improved by a biased government policy compared with rural residents. Albeit urban residents might have better health outcomes, objectively, they might not state that way. Lastly, we suspect that it is not the registration type but the migration status matters.

We construct an indicator variable where we include all cases of family members not living in the household, such as gone to school, military service, sought employment elsewhere and gone aboard. These samples are not currently living in the household but still accounted as household members. In our sample, only less than 2% of the observations belongs to this category, however, we find significant effects of migration status in cross-sectional models and probit panel models.[5] When we include this into dynamic probit models, it becomes insignificant, just like the case of Gini coefficient. We believe this is another evidence of health selection effects. Healthy individuals usually seek for better careers and better payments elsewhere. People with severe medical conditions have limited ability to leave their households, since they need to be taken care of by other family members. Once the health selection process is accounted, migration status becomes irrelevant, otherwise, it is a strong indictor for good health status of an individual.

4.5 Conclusions

In this paper, we provide evidences that in China individual level income inequality, measured by relative deprivation index, affects health outcomes. Health selection problem does exist. The absolute income hypothesis holds, however, its effect on health is attenuated when measurements of individual level inequality are present. The negatively significant effects of individual level inequality on health point to the weak version of relative income hypothesis. The strong version is rejected by using CHNS data. I also show

[5] The sample size was reduced from 2876 individuals to 1844 individuals due to the missing observations. The significant loss could be one of the factors that cause migration status to be insignificant. This is because if a family member is absent, his/her information might be forgone and these individuals are the ones involved in the question of migration status.

that a household's long run income level is more important for health than the short run income variation. The community level inequality's negative effect is offsett by its positive effect if a progressive tax rate system is adopted and public investments are improving. For a single household, improving its socio-economic status is equally important as raising its long run income level.

4.6 Tables and Results

Table 4.1: Definition of Variables

Variables	Definition
SRHS	Indicator of Self Rated Health Status, 1 if good health
Age	Individuals' age(20 or above)
Gender	Indicator, 1 if male
HHINC	Household income(thousands of Yuan), adjusted to year 2000
HHINC.AVG	Household average income(Thousands of Yuan), adjusted to year 2000
HHINC.DVI	Household income minus average income(Thousands of Yuan), adjusted to Year 2000
Gini	Community level Gini coefficient
RDA	Relative deprivation index(Thousands of Yuan)
RANK	Centile Rank within a community
EDUC	Years of education
URBAN	Indicator for Urban, 1 if Urban, 0 if rural
HHSIZE	Household size, number of household's members
MARR	Indicator for Marriage status, 1 if married
NumBH	Number of hospital beds for each province (thousands)

Table 4.2: Summary Statistics of Variables

Year	Year 1997		Year 2000		Year 2004		Year 2006	
Variables	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
SRHS	0.7211	(0.4485)	0.6408	(0.4798)	0.5563	(0.4969)	0.5549	(0.4970)
Age	4.475	(1.237)	4.775	(1.237)	5.173	(1.237)	5.370	(1.237)
Gender	0.4638	(0.4987)	·	·	·	·	·	·
HHINC	13.769	(11.431)	16.896	(15.727)	20.222	(19.166)	22.952	(32.661)
HHINC.AVG	16.121	(10.610)	·	·	·	·	·	·
HHINC.DVI	−2.351	(10.867)	0.7757	(12.855)	4.101	(14.110)	6.831	(26.670)
Gini	0.3430	(0.079)	0.3809	(0.1021)	0.3894	(0.0846)	0.4106	(0.0912)
RDA	4.282	(3.979)	5.839	(5.086)	6.850	(6.384)	8.524	(8.5207)
RANK	0.5010	(0.2820)	0.4991	(0.2789)	0.5060	(0.2818)	0.4990	(0.2803)
EDUC	5.906	(4.128)	6.202	(4.424)	7.032	(5.578)	7.066	(6.403)
URBAN	0.2440	(0.4296)	·	·	·	·	·	·
HHSIZE	4.0824	(1.396)	3.976	(1.440)	3.6974	(1.501)	3.755	(1.612)
MARR	0.9085	(0.2882)	0.8960	(0.3052)	0.8936	(0.3083)	0.8911	(0.3114)
NumBH	9.630	(3.293)	·	·	·	·	·	·

Table 4.3: Probit Cross-sectional Model without Household Income Decomposition.

Variables	(1)	(2)	(3)	(4)	(5)
Age	$-.4210^{***}$	$-.4138^{***}$	$-.4112^{***}$	$-.4747^{***}$	$-.4269^{***}$
	(.0669)	(.0670)	(.0671)	(.0670)	(.0671)
Age2	$.0128^{**}$	$.0131^{*}$	$.0129^{**}$	$.0148^{**}$	$.0142^{**}$
	(.0064)	(.0064)	(.0064)	(.0064)	(.0064)
Gender	$.2119^{***}$	$.2084^{***}$	$.2089^{***}$	$.2115^{***}$	$.2128^{***}$
	(.0257)	(.0257)	(.0257)	(.0257)	(.0257)
HHINC	$.0043^{***}$	$.0024^{***}$	$.0025^{***}$	$-.0001$	$-.0002$
	(.0008)	(.0009)	(.0008)	(.0010)	(.0010)
HHINC Squared/1000	$.0059^{***}$	-0.0033^{*}	-0.0039^{**}	$-.0003$	$-.0002$
	(.0019)	(0.0019)	(0.0019)	(0.0021)	(.0022)
Gini			-1.913^{***}		-2.012^{***}
			(.7382)		(.7310)
Gini Squared			2.328^{***}		2.077
			(.8898)		(.8805)
RDA		$-.0223^{***}$	$-.0216^{***}$		
		(.0039)	(.0041)		
RDA Squared		$.0003^{***}$	$.0003^{***}$		
		(.0001)	(.0001)		
RANK				$.6289^{***}$	$.6424^{***}$
				(.1722)	(.2266)
RANK Squared				$-.2147$	$-.2060$
				(.1757)	(.2327)
EDUC	.0038	$.0053^{*}$	$.0051^{*}$	.0039	.0033
	(.0028)	(.0028)	(.0028)	(.0028)	(.0028)
URBAN	.0364	.0373	.0252	0374	.0229
	(.0318)	(.0319)	(.0322)	(.0318)	(.0322)
HHSIZE	$-.0187^{**}$	$-.0284^{***}$	$-.0282^{***}$	$-.0317^{***}$	$-.0308^{***}$
	(.0085)	.0086	(.0086)	(.0087)	(.0087)
MARR	$-.0408$	$-.0440$	$-.0472$	$-.0515$	$-.0555$
	(.0428)	(.0596)	(.0428)	(.0429)	(.0429)

Note: *,** and *** are the significant levels with respect to 10%, 5% and 1%.

Table 4.4: Probit Cross-sectional Model with Household Income Decomposition.

Variables	(6)	(7)	(8)	(9)	(10)
Age	$-.4084^{***}$	$-.4039^{***}$	$-.4033^{***}$	$-.4273^{***}$	$-.4219^{***}$
	(.0670)	(.0672)	(.0672)	(.0672)	(.0672)
Age2	$.0118^{*}$	$.0123^{*}$	$.0123^{*}$	$.0141^{**}$	$.0137^{**}$
	(.0064)	(.0064)	(.0064)	(.0096)	(.0096)
Gender	$.2182^{***}$	$.2156^{***}$	$.2156^{***}$	$.2184^{***}$	$.2190^{***}$
	(.0257)	(.0258)	(.0258)	(.0258)	(.0258)
HHINC MEAN	$.0186^{***}$	$.0168^{***}$	$.0175^{***}$	$.0117^{***}$	$.0112^{***}$
	(.0027)	(.0027)	(.0028)	(.0029)	(.0029)
HHINC Mean Squared/1000	$-.1546^{***}$	$-.1402^{***}$	$-.1498^{***}$	$-.085^{**}$	$-.0834^{*}$
	(.0406)	(.041)	(.0412)	(.0427)	(.0427)
HHINC Deviation	.0002	−.0011	$-.0014^{*}$	$-.0020^{**}$	$.00390^{**}$
	(.0007)	(.0007)	(.0009)	(.0007)	(.0010)
Gini			-1.711^{**}		-1.923^{***}
			(.7346)		(.7284)
Gini Squared			2.331^{***}		2.146^{**}
			(.8851)		(.8757)
RDA		$-.0231^{***}$	$-.0241^{***}$		
		(.0039)	(.0040)		
RDA Squared		$.0003^{***}$	$.0003^{***}$		
		(.0001)	(0.0001)		
RANK				$.5560^{***}$	$.5636^{***}$
				(.1729)	(.1730)
RANK Squared				−.1833	−.1838
				(.1743)	(.1744)
EDUC	.0007	.0017	.0019	.0005	.0003
	(.0029)	(.0029)	(.0029)	(.0029)	(.0029)
URBAN	.0047	.0037	−.0070	.0085	−.0032
	(.0322)	(.0323)	(.0326)	(.0322)	(.0325)
HHSIZE	$-.0236^{***}$	$-.0351^{***}$	$-.0363^{***}$	$-.0363^{***}$	$-.0355^{***}$
	(.0085)	(.0087)	(.0087)	(.0087)	(.0087)
MARR	−.0526	−.0571	−.0604	−.0623	−.0654
	(.0429)	(.0429)	(.0430)	(.0430)	(.0430)

Table 4.5: Probit Panel Model without Household Income Decomposition.

Variables	(1)	(2)	(3)	(4)	(5)
Age	$-.4705^{***}$	$-.4540^{***}$	$-.4462^{***}$	$-.4809^{***}$	$-.4637^{***}$
	(.0869)	(.0871)	(.0873)	(.0868)	(.0871)
Age2	.0114	.0113	.0107	.0134	.0121
	(.0083)	(.0083)	(.0084)	(.0083)	(.0083)
Gender	$.2528^{***}$	$.2469^{***}$	$.2476^{***}$	$.2509^{***}$	$.2520^{***}$
	(.0357)	(.0357)	(.0357)	(.0356)	(.0357)
HHINC	$.0039^{***}$	$.0014^{***}$	$.0016^{***}$	$-.0009$	$-.0009$
	(.0009)	(.0010)	(.0010)	(.0013)	(.0013)
HHINC Squared/1000	$.0059^{***}$	-0.0022	-0.0031	$-.0001$	$-.0001$
	(.0026)	(0.0023)	(0.0023)	(0.0025)	(.0026)
Gini			-2.587^{***}		-2.675^{***}
			(.8449)		(.8358)
Gini Squared			3.039^{***}		2.700
			(1.016)		(1.005)
RDA		$-.0256^{***}$	$-.0240^{***}$		
		(.0046)	(.0049)		
RDA Squared		$.0003^{***}$	$.0003^{**}$		
		(.0001)	(.0001)		
RANK				$.6624^{***}$	$.6735^{***}$
				(.1973)	(.1976)
RANK Squared				$-.2161$	$-.2286$
				(.2015)	(.2019)
EDUC	.0022	.0046	.0044	.0027	.0022
	(.0036)	(.0036)	(.0036)	(.0036)	(.0036)
URBAN	.0658	.0640	.0456	.0656	.0431
	(.0436)	(.0437)	(.0442)	(.0435)	(.0440)
HHSIZE	$-.0149$	$-.0255^{**}$	$-.0282^{**}$	$-.0281^{***}$	$-.0274^{**}$
	(.0107)	(.0109)	(.0109)	(.0109)	(.0109)
MARR	$-.0626$	$-.0658$	$-.0707$	$-.0724$	$-.0788$
	(.0541)	(.0542)	(.0543)	(.0541)	(.0543)

Table 4.6: Probit Panel Model with Household Income Decomposition.

Variables	(6)	(7)	(8)	(9)	(10)
Age	$-.4534^{***}$	$-.4414^{***}$	$-.4360^{***}$	$-.4724^{***}$	$-.4586^{***}$
	(.0866)	(.0868)	(.0869)	(.0865)	(.0869)
Age2	.0104	.0107	.0103	.0128	.0118
	(.0083)	(.0083)	(.0083)	(.0083)	(.0083)
Gender	$.2576^{***}$	$.2529^{***}$	$.2528^{***}$	$.2570^{***}$	$.2574^{***}$
	(.0355)	(.0355)	(.0355)	(.0354)	(.0355)
HHINC MEAN	$.0205^{***}$	$.0185^{***}$	$.0175^{***}$	$.0132^{***}$	$.0122^{***}$
	(.0037)	(.0038)	(.0028)	(.0039)	(.0040)
HHINC Mean Squared/1000	$-.1689^{***}$	$-.1526^{***}$	$-.00016^{***}$	$-.00009^{**}$	−.00008
	(.0563)	(.0566)	(.00005)	(.00005)	(.00005)
HHINC Deviation	.0003	−.0012	$-.0015^{*}$	$-.0021^{**}$	$-.0021^{**}$
	(.0007)	(.0008)	(.0008)	(.0009)	(.0009)
Gini			-2.384^{**}		-2.602^{***}
			(.8381)		(.8309)
Gini Squared			3.025^{***}		2.763^{**}
			(1.008)		(.9976)
RDA		$-.0263^{***}$	$-.0266^{***}$		
		(.0045)	(.0048)		
RDA Squared		$.0003^{***}$	$.0003^{***}$		
		(.0001)	(0.0001)		
RANK				$.5973^{***}$	$.6109^{***}$
				(.1975)	(.1978)
RANK Squared				−.1968	−.2043
				(.1992)	(.1996)
EDUC	−.0010	.0007	.0009	−.0009	.0003
	(.0036)	(.0037)	(.0037)	(.0036)	(.0036)
URBAN	.0249	.0203	−.0044	.0281	.0100
	(.0440)	(.0440)	(.0444)	(.0439)	(.0443)
HHSIZE	$-.0200^{*}$	$-.0329^{***}$	$-.0335^{***}$	$-.0334^{***}$	$-.0325^{***}$
	(.0107)	(.0109)	(.0109)	(.0109)	(.0109)
MARR	−.0735	−.0784	−.0828	−.0831	−.0883
	(.0541)	(.0541)	(.0542)	(.0541)	(.0542)

Table 4.7: Random-Effects Dynamic Probit Model without Household Income Decomposition(Regression for $t > 1$).

SRHS($t > 1$)	(1)	(2)	(3)	(4)	(5)
$SRHS_{t-1}$	.1109**	.0999*	..1004**	.1107**	.1093**
	(.0519)	(.0522)	(.0522)	(.0519)	(.0521)
Age	−.4638***	−.4622***	−.4659***	−.4743***	−.4753***
	(.1033)	(0.1038)	(.1038)	(.1033)	(.1034)
Age2	0.0136	(.0142)	.0147	.0157	.0157
	(0.0097)	(0.0097)	(.0097)	(.0096)	(.0096)
Gender	.2460***	.2457***	.2441***	.2463***	.2464***
	(0.0390)	(.0392)	(.0392)	(.0390)	(.0390)
HHINC	.0054***	.0037***	.0037***	.0008	.0009
	(.0010)	(.0011)	(.0011)	(.0013)	(.0014)
HHINC Squared/1000	.0074***	−0.0052**	−0.0058**	−.0011	−.0017
	(.0023)	(0.0024)	(0.0024)	(0.0026)	(.0026)
Gini			−.8091		−.8743
			(.9794)		(.9678)
Gini Squared			1.504		1.227
			(1.146)		(1.129)
RDA		−.0161***	−.0186***		
		(.0052)	(.0054)		
RDA Squared		.0002	.0002*		
		(.0002)	(0.0001)		
RANK				.6446***	.6424***
				(.2266)	(.2266)
RANK Squared				−.2022	−.2060
				(.2327)	(.2327)
EDUC	.0063*	.0072*	.0081	.0063*	.0065*
	(0.0038)	(.0038)	(.0039)	(.0038)	(.0038)
URBAN	.0245	.0319	.0279	.0308	.0249
	(.0475)	(.0479)	(.0483)	(.0475)	(.0479)
HHSIZE	−.0330***	−.0395***	−.0426***	−.0453***	−.0460***
	(.0118)	0.0119	(.0120)	(.0120)	(.0120)
MARR	−.0824	−.0840	.0120	−.0965	−.0988*
	(.0593)	(.0596)	(.0596)	(.0594)	(.0594)

Table 4.8: Random-Effects Dynamic Probit Model without Household Income Decomposition(Regression for $t = 1$).

SRHS($t = 1$)	(1)	(2)	(3)	(4)	(5)
Age	$-.3758^{**}$	$-.3700^{**}$	$-.3598^{**}$	$-.3863^{**}$	$-.3726^{**}$
	(.1514)	(.1511)	(.1518)	(.1512)	(.1519)
Age2	.0123	.0126	.0114	.0142	.0130
	(.0157)	(.0157)	(.0114)	(.0157)	(.0158)
Gender	$.1745^{***}$	$.1700^{***}$	$.1728^{***}$	$.1720^{***}$	$.1742^{***}$
	(.0592)	(.0591)	(.0593)	(.0591)	(.0594)
HHINC	$.0083^{*}$	$.0016^{***}$	.0012	$-.0013$	$-.0055$
	(.0048)	(.0054)	(.0054)	(.0071)	(.0074)
HHINC Squared/1000	$-.0967$	$-.0485$	$-.0372$	0.0037	.0512
	(.0723)	(.0761)	(.0773)	(.0917)	(.0985)
Gini			-1.409		-1.688
			(2.387)		(2.378)
Gini Squared			.4304		.4054
			(3.326)		(3.331)
RDA		$-.0442^{**}$	$-.0322^{*}$		
		(.0188)	(.0193)		
RDA Squared		.0015	.0010		
		(.0010)	(.0011)		
RANK				$.9617^{**}$	1.055^{***}
				(.3953)	(.3995)
RANK Squared				$-.6829^{*}$	$-.7236^{*}$
				(.3893)	(.3923)
EDUC	.0111	.0123	.0102	.0119	.0096
	(.0084)	(.0083)	(.0084)	(.0083)	(.0084)
URBAN	$-.0168$	$-.0234$	$-.0384$	$-.0194$	$-.0403$
	(.0720)	.0719	(.0733)	(.0719)	(.0733)
HHSIZE	.0165	.0109	.01603	.0106	.0143
	(.0211)	(.0212)	(.01196)	(.0212)	(.0213)
MARR	$-.0138$	$-.0176$	$-.0193$	$-.0174$	$-.0222$
	(.1004)	(.1003)	(.1006)	(.1004)	(.1008)
NumBH	$.0154^{**}$	$.0186^{**}$	$.0164^{*}$	$.0176^{**}$	$.0166^{*}$
	(.0085)	(.0086)	(.0087)	(.0086)	(.0087)
θ	$.7321^{***}$	$.7077^{***}$	$.7267^{***}$	$.7234^{***}$	$.7414^{***}$
	(.1150)	(.1106)	(.1133)	(.1152)	(.1174)

Table 4.9: Random-Effects Dynamic Probit Model with Household Income Mean-Deviation Decomposition(Regression for $t > 1$).

SRHS($t > 1$)	(6)	(7)	(8)	(9)	(10)
$SRHS_{t-1}$	.1055**	.0918*	.0947*	.1030**	.1041**
	(.0519)	(.0522)	(.0522)	(.0519)	(.0520)
Age	−.4564***	−.4577***	−.4626***	−.4735***	−.4748***
	(.1033)	(.1041)	(.1039)	(.1035)	(.1034)
Age2	.0135	.0140	.0146	.0156	.0157
	(0.0096)	(0.0097)	(.0096)	(.0096)	(.0096)
Gender	.2517***	.2526***	.2503***	.2530***	.2524***
	(0.0390)	(.0393)	(.0392)	(.0391)	(.0390)
HHINC MEAN	.0211***	.0196***	.0213***	.0129***	.0139***
	(.0041)	(.0042)	(.0042)	(.0043)	(.0044)
HHINC Mean Squared/1000	−.1698***	−.1559***	−.1727***	−.0817	−.0921
	(.0615)	(.0621)	(0622)	(.064)	(.064)
HHINC Deviation	.0011	−.0001	−.0007	−.0014	.0390
	(.0008)	(.0009)	(.0009)	(.0009)	(.0010)
Gini			−.4496		−.6589
			(.9765)		(.9665)
Gini Squared			1.316		1.120
			(1.141)		(1.124)
RDA		−.0186***	−.0229***		
		(.0050)	(.0053)		
RDA Squared		.0002$*$	.0003**		
		(.0001)	(.0001)		
RANK				.5850***	.5777**
				(.2274)	(.2273)
RANK Squared				−.1488	(−.1427)
				(.2302)	(.2301)
EDUC	.0034	.0039	.0048	.0031	.0035
	(.0039)	(.0038)	(.0039)	(.0039)	(.0039)
URBAN	−.0048	−.0000	−.0047	.0019	−.0036
	(−.0048)	(.0484)	(.0486)	(.0480)	(.0483)
HHSIZE	−.0360***	−.0451***	−.0501***	−.0498***	−.0510***
	(.0118)	.0120	(.0121)	(.0120)	(.0120)
MARR	.0118	−.0941	−.0971	−.1058*	−.1080*
	(.0594)	(.0598)	(.0597)	(.0595)	(.0595)

Table 4.10: Random-Effects Dynamic Probit Model with Household Income Mean-Deviation Decomposition(Regression for $t = 1$).

SRHS($t = 1$)	(6)	(7)	(8)	(9)	(10)
Age	$-.3489^{**}$	$-.3488^{**}$	$-.3441^{**}$	$-.3748^{**}$	$-.3688^{**}$
	(.1509)	(.1507)	(.1512)	(.1510)	(.1517)
Age2	.0097	.0107	.0102	.0132	.0127
	(.0156)	(.0156)	(.0157)	(.0157)	(.0157)
Gender	$.1808^{***}$	$.1780^{***}$	$.1792^{***}$	$.1794^{***}$	$.1804^{***}$
	(.0592)	(.0591)	(.0593)	(.0591)	(.0594)
HHINC MEAN	$.0156^{**}$	$.0137^{**}$	$(.0119)^{*}$	$-.0111^{*}$	.0079
	(.0062)	(.0063)	(.0063)	(.0066)	(.0067)
HHINC Mean Squared/1000	$-.1593^{*}$	$-.1916^{**}$	$-.1608^{*}$	0.1567^{*}	$-.0002^{**}$
	(.0946)	(.0953)	(.0960)	(.0954)	(.0001)
HHINC Deviation	−.0443	−.0055	−.0044	−.0045	−.0045
	(.0031)	(.0034)	(.0034)	(.0040)	(.0040)
Gini			−1.412		−1.829
			(2.386)		(2.377)
Gini Squared			.6116		.7317
			(3.326)		(3.329)
RDA		$-.0490^{***}$	$-.0378^{*}$		
		(.0188)	(.0193)		
RDA Squared		.0015	$-.0011^{***}$		
		(.0011)	(.0011)		
RANK				$.8842^{*}$	$.9315^{**}$
				(.3792)	(.3808)
RANK Squared				−.5870	−.6220
				(.3792)	(.3915)
EDUC	.0076	.0075	.0062	.0076	.0058
	(.0085)	(.0085)	(.0086)	(.0085)	(.0086)
URBAN	−.0448	−.0640	−.0728	−.0537	−.0705
	(.0732)	(.0733)	(.0745)	(.0732)	(.0746)
HHSIZE	.0136	.0041	.0095	.0047	.0089
	(.0212)	(.0213)	(.0215)	(.0214)	(.0215)
MARR	−.0267	−.0314	−.0306	−.0272	−.0284
	(.1004)	(.1003)	(−.0306)	(.1005)	(.1009)
NumBH	.0137	$.0167^{*}$	$.0149^{*}$	$.0157^{*}$	$.0147^{*}$
	(.0085)	(.0086)	(.0087)	(.0086)	(.0087)
θ	$.7205^{***}$	$.6922^{***}$	$.7114^{***}$	$.7141^{***}$	$.7342^{***}$
	(.1140)	(.1090)	(.1123)	(.1139)	(.1169)

Bibliography

Ai, C. & Norton, E. C. (2003), 'Interaction terms in logit and probit models', *Economics Letters* **80**(1), 123 – 129.

Aitchison, J. & Aitken, C. G. G. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**(3), pp. 413–420. *http://www.jstor.org/stable/2335719

Benzeval, M. & Judge, K. (2001), 'Income and health: the time dimension', *Social Science & Medicine* **52**(9), 1371–1390.

Braun, W. J. & Hall, P. (2001), 'Data sharpening for nonparametric inference subject to constraints', *Journal of Computational and Graphical Statistics* **10**(4), pp. 786–806.

Butler, J. S. & Moffitt, R. (1982), 'A computationally efficient quadrature procedure for the one-factor multinomial probit model', *Econometrica* **50**(3), 761–64.

Choi, E. & Hall, P. (1999), 'Data sharpening as a prelude to density estimation', *Biometrika* **86**(4), 941–947. *http://www.jstor.org/stable/2673598

Choi, E., Hall, P. & Rousson, V. (2000), 'Data sharpening methods for bias reduction in nonparametric regression', *The Annals of Statistics* **28**(5), 1339–1355.

Eibner, C. E. & Evans, W. N. (2005), 'Relative deprivation, poor health habits and mortality', *Journal of Human Resources* .

Gu, J., Li, D. & Liu, D. (2007), 'Bootstrap non-parametric significance test', *Journal of Nonparametric Statistics* **19**(6), 215–230.

Hall, P. & Huang, L.-S. (2001), 'Nonparametric kernel regression subject to monotonicity constraints', *The Annals of Statistics* **29**(3), pp. 624–647. *http://www.jstor.org/stable/2673965

Hall, P., Huang, L.-S., Gifford, J. A. & Gijbels, I. (2001), 'Nonparametric estimation of hazard rate under the constraint of monotonicity', *Journal of Computational and Graphical Statistics* **10**(3), 592–614.

Hall, P., Li, Q. & Racine, J. S. (2007), 'Nonparametric estimation of regression functions in the presence of irrelevant regressors', *The Review of Economics and Statistics* **89**(4), 784–789.

Hall, P. & Presnell, B. (1999), 'Biased bootstrap methods for reducing the effects of contamination', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**(3), pp. 661–680. *http://www.jstor.org/stable/2680729

Heckman, J. (1981*a*), 'Heterogeneity and state dependence', pp. 91–140.

Heckman, J. (1981*b*), 'The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process', pp. 114–178.

Judge, K. & Paterson, I. (2001), 'Poverty, income inequality and health', (01/29).

Kawachi, I. & Kennedy, B. P. (1999), 'Income inequality and health: pathways and mechanisms', *Health services research* **34**, 215 – 227.

Krugman, P. (1996), 'The spiral of inequality'.

Lavergne, P. & Patilea, V. (2008), 'Breaking the curse of dimensionality in nonparametric testing', *Journal of Econometrics* **143**(1), 103 – 122.

Lavergne, P. & Vuong, Q. (2000), 'Nonparametric significance testing', *Econometric Theory* **16**(04), 576–601.

Li, H. & Zhu, Y. (2004), 'Income, income inequality, and health: Evidence from china'.

Li, Q. & Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Lynch, J., Smith, G., Kaplan, G. & House, J. (2009), 'Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions', *BMJ. British medical journal* **320**, 1200.

Masry, E. (1996*a*), 'Multivariate local polynomial regression for time series: Uniform strong consistency and rates', *J. Time Ser. Anal* **17**, 571–599.

Masry, E. (1996*b*), 'Multivariate regression estimation: Local polynomial fitting for time series', pp. 81–101.

Mellor, J. M. & Milyo, J. (2002), 'Income inequality and health status in the united states: Evidence from the current population survey', *The Journal of Human Resources* **37**(3), pp. 510–539.

Mundlak, Y. (1978), 'On the pooling of time series and cross section data', *Econometrica* **46**(1), 69–85.

Nadaraya, E. A. (1965), 'On non-parametric estimates of density functions and regression curves', *Theory of Probability and its Applications* **10**(1), 186–190.
*http://link.aip.org/link/?TPR/10/186/1

Pagan, A. & Ullah, A. (1999), 'Nonparametric econometrics / adrian pagan, aman ullah', pp. xviii, 424 p. :.
*http://www.loc.gov/catdir/toc/cam026/98037218.html

Racine, J. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99 – 130.
*http://www.sciencedirect.com/science/article/B6VC0-48N30VY-1/2/b935a0c6531e8ad7cbb7e571aa5a21a1

Racine, J. S. (1997), 'Consistent significance testing for nonparametric regression', *Journal of Business and Economic Statistics* **15**, 369–378.

Racine, J. S., Hart, J. & Li, Q. (2006), 'Testing the significance of categorical predictor variables in nonparametric regression models', **25**(4), 523 – 544.

Racine, S. J., Parmeter, F. C. & Du, P. (2008), 'Constrained nonparametric kernel regression: Estimation and inference', *Submitted to Journal of Econometrics* .

Sperlich, S., Tjostheim, D. & Yang, L. (2002), 'Nonparametric estimation and testing of interaction in additive models', *Econometric Theory* **18**(2), 197–251.

Stone, C. J. (1980), 'Optimal rates of convergence for nonparametric estimators', *The Annals of Statistics* **8**(6), 1348–1360.

Wang, M.-C. & Ryzin, J. v. (1981), 'A class of smooth estimators for discrete distributions', *Biometrika* **68**(1), pp. 301–309.
*http://www.jstor.org/stable/2335831

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhya-: The Indian Journal of Statistics, Series A* **26**(4), pp. 359–372.
*http://www.jstor.org/stable/25049340

Wilkinson, R. (1999), 'Putting the picture together: prosperity, redistribution, health, and welfare', *Social determinants of health, Oxford University Press* pp. 256 – 274.

Yitzhaki, S. (1979), 'Relative deprivation and the gini coefficient', *The Quarterly Journal of Economics* **93**(2), 321–24.