

HEALTHCARE DECISIONS: FORMULATION AND
APPLICATION OF SEMIPARAMETRIC METHODS

by

CHAN SHEN

A dissertation submitted to the
Graduate School-New Brunswick
Rutgers, The State University of New Jersey

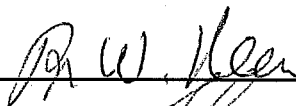
In partial fulfillment of the requirements

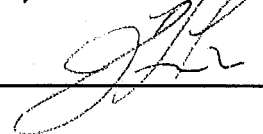
For the degree of Doctor of Philosophy

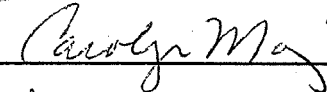
Graduate Program in Economics

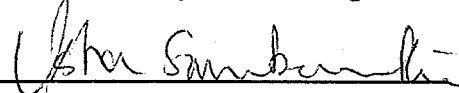
Written under the direction of Roger Klein

And approved by









New Brunswick, New Jersey

May, 2009

ABSTRACT OF THE DISSERTATION

HEALTHCARE DECISIONS: FORMULATION AND APPLICATION OF SEMIPARAMETRIC METHODS

By Chan Shen

Dissertation Director: Roger Klein

Currently, there appears to be a tradeoff between the performance of a semiparametric estimator in finite and large samples. In Chapter 1, we argue that this tradeoff occurs because of the nature of the bias reduction methods that are often employed in implementing these estimators. Accordingly, we develop a bias control mechanism that eliminates this tradeoff so as to ensure that the estimator performs well in finite samples while retaining desirable large sample properties.

Semiparametric models are commonly estimated under a single index assumption. In estimating these models, the consistency of the estimator critically depends on this assumption being correct. Therefore, in Chapter 2, we develop a test of this assumption. We formulate such a test and derive its large sample distribution under the null hypothesis of a single index. To ensure that the test statistic has good size and power properties in finite samples, we formulate a test whose form adapts to the model under the alternative hypothesis. Monte Carlo results confirm that the adaptive feature significantly improves the performance of the test statistic in finite samples.

Studying healthcare decisions poses many empirical challenges. Healthcare utilization and expenditures depend on health insurance and other health related variables. As insurance is a choice variable for the individual, there are potential endogeneity issues. Expenditures are only observed when utilization occurs and hence there is a selection problem. Furthermore, the decision to utilize healthcare and the decision about the level of treatment are determined by different decision makers. In Chapter 3, we study a system of three simultaneous equations: insurance, utilization, and expenditures. To avoid making traditional parametric distributional assumptions, we propose a semiparametric approach

based on the previous two chapters. Both parametric and semiparametric approaches are employed in an empirical study using the Medical Expenditure Panel Survey (MEPS) 2005 data. We find that insurance increases the likelihood of seeking healthcare by about 15% points (from about 80% to 95%). We also find that the parametric approach predicts insurance to increase the level of expenditures by 125%; while the semiparametric method predicts an increase of 51%, a number in accord with an important experimental study in the literature.

Acknowledgements

A great many people have contributed to my dissertation. I owe my gratitude to all those people who have made this dissertation possible.

My deepest gratitude is to my advisor, Prof. Roger Klein. I have benefitted greatly from his invaluable guidance, patience, and encouragement. The influence will remain with me.

I am extremely grateful to Prof. Carolyn Moehling and Prof. John Landon-Lane. I have been amazingly fortunate to have all the great comments and suggestions from them.

I am also indebted to Dr. Usha Sambamoorthi and Prof. Louise Russell for their insightful and inspiring comments on my dissertation. They helped me tremendously in understanding health economics research.

I would like to thank the members of the economics department, especially Dorothy Rinaldi, for making my graduate experience one that I will cherish forever.

I am also thankful to all my colleagues and friends. I greatly value their friendship and support.

Most importantly, none of this would have been possible without the love and patience of my family. I would especially like to express my heart-felt gratitude to my family in China, who has been a constant source of love, support, and strength in all these years.

Table of Contents

Abstract	ii
Acknowledgements	iv
Part I. Introduction	1
Part II. Chapter 1. A Semiparametric Estimator with Bias Corrections	7
1.1 Introduction	7
1.2 Moment Conditions and Bias Control	9
1.3 Assumptions, Definitions, and Results	13
1.4 Monte Carlo Designs and Results	20
1.4.1 Designs	20
1.4.2 Monte Carlo Results	21
1.5 Conclusions	22
Part II. Chapter 2. A Test of the Single Index Assumption in Semiparametric Models	23
2.1 Introduction	23
2.2 Moment Conditions and Bias Control	25
2.3 Assumptions, Definitions, and Results	28
2.4 Monte Carlo Designs and Results	34
3.4.1 Designs	35
3.4.2 Monte Carlo Results	37
2.5 Conclusions	39
Part II. Chapter 3. Determinants of Healthcare Decisions: Insurance, Utilization, and Expenditures	40
3.1 Introduction	40
3.2 The Model	41
3.2.1 Parametric Model	43

3.2.2 Semiparametric Model.....	46
3.3 Data.....	54
3.4 Results.....	58
3.5 Conclusions.....	63
Appendices	83
Appendix A.....	83
Appendix B.....	95
References	102
Vita	106

List of Tables

Table 1.1 Estimation Results	65
Table 2.1 Test Results	66
Table 2.2 Comparison of Fixed Weight and Adaptive Tests.....	68
Table 3.1 Description of Study Population	69
Table 3.2 Description of Study Population by Insurance Coverage	71
Table 3.3 Description of Study Population by Utilization.....	73
Table 3.4 Parametric and Semiparametric Estimation Results – Insurance Coverage	75
Table 3.5 Parametric and Semiparametric Estimation Results – Utilization	77
Table 3.6 Parametric and Semiparametric Estimation Results – Level of Expenditures..	79
Table 3.7 Marginal Effects across the Distribution of Select Variables of Interest.....	81

List of Illustrations

Figure 3.1 Estimated error distribution in the insurance equation.....	82
--	----

Part I

Introduction

1 A Semiparametric Estimator with Bias Corrections

Currently, there appears to be a tradeoff between the performance of a semiparametric estimator in finite and large samples. In Chapter 1, we argue that this tradeoff occurs because of the nature of the bias reduction methods that are often employed in implementing these estimators. Accordingly, we develop a bias control mechanism that eliminates this tradeoff so as to ensure that the estimator performs well in finite samples while retaining desirable large sample properties. As is typically done, we develop this estimator when the model satisfies an index structure. To explain this structure, let:

$$\begin{aligned} V_1 &= Z_1\theta_1 + Z_2\theta_2 + Z_3\theta_3 + c_1 \\ V_2 &= X_3\theta_3 + \theta_5(X_4^{\theta_4} - 1)/\theta_4 + c_2 \end{aligned}$$

When $E(Y|X) = G(V_1)$, we refer to this model as a single index model with linear index V_1 . Notice that this linear index permits interactions and higher order terms in that we can have $Z_1 = X_1$, $Z_2 = X_1X_2$, and $Z_3 = X_2^2$. When $E(Y|X) = G(V_2)$, this is a single index model with a nonlinear index. Finally, when $E(Y|X) = G(V_1, V_2)$, this is a double index model with indices V_1 and V_2 . As index parameters are at most identified up to location and scale in the cases above, it is common to normalize the parameters so that the coefficient on one of the explanatory variables is one and the constant term is zero.

The most commonly used index structure is a single index assumption (See, for example, Ahn (1997), Climov, Delecroix and Simar (2002), Fraga and Martins (2001), Gerfin (1996), Gorgens (2000), Gorgens and Horowitz (1999), Ichimura (1993), Klein and Sherman (2002), Klein and Spady (1993)). To explain the importance of this single index assumption, consider a nonparametric setting in which there is no parametric information on the form of the expectation of the dependent variable conditioned on the explanatory variables, X . In this case, it is difficult to obtain an accurate estimate of the conditional expectation when

the dimension of X is high. For example, with X a high dimensional vector of exogenous variables, consider the binary response model:

$$Y = \begin{cases} 1 : & f(X) > u \\ 0 : & \text{otherwise} \end{cases},$$

where f and the distribution of u are both unknown and u is independent of X . In this nonparametric case, it is well known that it is difficult to obtain a good estimate of $Pr(Y = 1|X) = E(Y|X)$ due to the high dimension of X . In contrast, a semiparametric single index model avoids this problem by imposing a parametric structure on f . For example, $f(X) = X\theta_0$ aggregates the information in X into a single linear index and reduces the dimension to one. This linear index is the most commonly used.

As in the example above, let θ_0 be a parameter vector of interest and denote $\hat{\theta}$ as its estimator. To insure that this estimator has desirable properties, in large and finite samples, it is critical to have a good estimator for the conditional expectation, $E(Y|V(\theta))$. There are methods to estimate this expectation (e.g., those based on higher order kernels) that deliver $\hat{\theta}$ with desirable large sample properties but often with poor finite sample performance. The problem here is that in implementing a bias control to obtain asymptotic normality, such methods can fail to satisfy restrictions on the true conditional expectation of interest in finite samples. For example, in the binary model above, such methods can deliver estimated probabilities that fall outside of the interval $[0,1]$. In contrast, methods that impose these restrictions (e.g., those based on regular kernels), perform well in finite samples. However, in failing to adequately implement a bias control, these methods do not achieve \sqrt{N} -asymptotic normality.

The objective of this chapter is to provide an estimator with desirable large sample properties and that also performs well in finite samples. To obtain these properties, we employ two alternative bias controls and regular kernels. The bias controls will insure normality, while the use of regular kernels allows us to impose known restrictions on the estimated conditional expectation. We find that the resulting estimator performs well in finite samples, an important feature for its use in applications. It should be remarked that

while this estimator was developed for single index models, it can be extended to the double index case (see Klein, Shen, and Vella, 2009). Such models are important and naturally arise when one or more of the explanatory variables is endogenous. For an application in the context of healthcare decisions, see Shen (2008).

2 A Test of the Single Index Assumption in Semiparametric Models

As discussed above, semiparametric models are commonly estimated under a single index assumption. In estimating these models, the consistency of the estimator critically depends on this assumption being correct. Therefore, in Chapter 2, we develop a test of this assumption. We formulate such a test and derive its large sample distribution under the null hypothesis of a single index. To ensure that the test statistic has good size and power properties in finite samples, we formulate a test whose form adapts to the model under the alternative hypothesis. Monte Carlo results confirm that the adaptive feature significantly improve the performance of the test statistic in finite samples.

The most commonly used index structure is a single index assumption (See, for example, Ahn (1997), Climov, Delecroix and Simar (2002), Fraga and Martins (2001), Gerfin (1996), Gorgens (2000), Gorgens and Horowitz (1999), Ichimura (1993), Klein and Sherman (2002), Klein and Spady (1993)). When the single index assumption does not hold, its imposition will result in an inconsistent estimator for the conditional expectation of interest. Given the sensitivity of the estimator to a single index assumption and given its wide use, a second objective of this paper is to formulate a test for this assumption.

There have been papers in the literature on testing parametric against semiparametric models (e.g., Härdle, Mammen, and Müller (1998), Hardle, Spokoiny, and Sperlich (1997), Horowitz and Hardle(1994)). Related tests of parametric models are given by Newey's (1985) paper on conditional moment tests and Bieren(1990) conditional moment test. This paper differs from those above in that it formulates a test for a main assumption in semiparametric models. We note that in a likelihood context with a parametric null hypothesis, Newey develops conditional moment tests that have optimal local power properties. It may be possible to extend these results to the present context, but this extension is beyond the scope of the current paper.

There have been some papers which focus on testing single index restrictions. Escanciano and Song (2007) provide a test focusing on average marginal effects and show that it has a minimax property; Andrews (1993) provides high level conditions for testing moment restrictions. Our paper differs from these in that we provide primitive conditions for a conditional moment test and for the estimator on which it is based. Tripathi and Kitimura (2001) employ an empirical likelihood approach for testing moment conditions of the form: $E[G(z, \theta)|X] = 0$. With G a known function and X continuous, they establish an optimality property for their test. The test proposed here is also an orthogonality test in that we test whether a function G is correlated with functions $M(X)$. However, unlike the above test, here the function G will be unknown as we set $G = Y - E(Y|V)$, where V is an index and the conditional expectation function $E(Y|V)$ is unknown. Further, the M -functions will be unknown and we will require nonparametric estimates of them. This feature is needed to insure that the form of test statistic adapts to the model under the alternative hypothesis. The proposed statistic also differs from those in the literature in the bias control mechanism that it employs. This mechanism is similar to that underlying the estimator, and results in a test statistic that has good size and power properties in finite samples.

3 Determinants of Healthcare Decisions: Insurance, Utilization, and Expenditures

A major healthcare policy issue in the U.S. today is the growing population without insurance. The key questions include: How does health insurance coverage affect the likelihood an individual seeks medical care? How does health insurance affect their healthcare expenditures? The purpose of this third chapter is to study these and other related policy questions.

There are many empirical challenges in studying people's healthcare decisions. An individual's decision about whether to utilize healthcare may depend on her insurance coverage. The level of utilization likely also depends on whether or not the individual has insurance. However, because insurance is a choice variable for the individual, we must allow for the possibility that this variable is endogenous. For example, people who have greater need for healthcare have more incentive to buy health insurance. Some papers in the literature deal

with this endogeneity by using instrumental variables (e.g., Vera-Hernandez, 1999; Holly et al., 2002; Wooldridge, 2002); others use experimental data to avoid this problem (e.g., Manning et al., 1987; Newhouse, 1993; Barros et al., 2008). However, instruments that are correlated with insurance coverage but not with utilization are difficult to find. Experimental data are scarce and often out of date. For example, the RAND Health Insurance Experiment, which remains to be the largest health policy study in U.S. history, started in 1971 and lasted for 15 years (RAND, 1974-1982). The structure, practice and philosophy of medicine has changed dramatically since the 1980s as has the insurance industry.

Another empirical challenge is the selection issue in the expenditure decision. Namely, expenditures are only observed for individuals who decide to see a doctor. One standard parametric approach deals with this problem by making distributional assumptions about error terms and then using a Heckman correction for sample selection (Heckman, 1976, 1979). An alternative approach is to use a two-part model (Duan et al., 1983, 1984, 1985). Both of these may be problematic as the Heckman correction approach can be sensitive to the distributional assumptions on error terms, while the two-part model approach also makes implicit distributional assumptions (Puhani, 2000). The literature in health economics or in economics in general does not provide a theoretical foundation or justification for these distributional assumptions. Moreover, if incorrect, they can result in incorrect inferences and policy conclusions with respect to healthcare decisions.

Yet another challenge is the complicated nature of the decision-making process. In healthcare, both the patient and the doctor are involved in making decisions. The patient decides whether to visit a doctor (or more generally a healthcare provider), and then the patient and doctor jointly decide what treatment the patient will have. These decisions are interrelated. Some papers deal with the two-part decision-making process in healthcare utilization (Newhouse, 1993; Mullahy, 1998), but none address the whole process of insurance choice, utilization, and expenditure level.

This chapter contributes to the current literature by taking into account the interrelated nature of healthcare decisions and using a semiparametric approach to address the empirical challenges. We study three components of obese people's healthcare decisions: insurance coverage, utilization, and the level of expenditures. Using the Medical Expenditure Panel

Survey (MEPS) 2005 data, we formulate and estimate a model for these three healthcare decisions. First, we estimate the model using a standard parametric approach. We estimate the health insurance and access decisions using bivariate probit-type estimators under the assumption that the errors are jointly normal. Then assuming that the error component in expenditures also has a normal distribution, we estimate the final equation making a traditional correction to control for both sample selection and the endogeneity of insurance.

As there is not a strong justification for the normality assumptions underlying the parametric formulation, we next employ a semiparametric approach in which these assumptions are not made. As an additional advantage to a semiparametric approach, it should be remarked that since marginal effects will in general not be constant in nonlinear models, we will report the impact of changing a variable of interest at several different points in the distribution of the variable of interest. The semiparametric approach will also allow greater flexibility in the pattern of these effects than in the parametric case. This approach is based on the estimator discussed in Chapter 1.

Both parametric and semiparametric approaches are employed in an empirical study using the Medical Expenditure Panel Survey (MEPS) 2005 data. We find that insurance increases the likelihood of seeking healthcare by about 15% points (from about 80% to 95%). We also find that the parametric approach predicts insurance to increase the level of expenditures by 125%; while the semiparametric method predicts an increase of 51%, a number in accord with an important experimental study in the literature.

Part II

Chapter 1

A Semiparametric Estimator with Bias Corrections

1.1 Introduction

Semiparametric models are typically estimated under an index structure. To explain this structure, let:

$$V_1 = Z_1\theta_1 + Z_2\theta_2 + Z_3\theta_3 + c_1$$

$$V_2 = X_3\theta_3 + \theta_5(X_4^{\theta_4} - 1)/\theta_4 + c_2$$

When $E(Y|X) = G(V_1)$, we refer to this model as a single index model with linear index V_1 . Notice that this linear index permits interactions and higher order terms in that we can have $Z_1 = X_1$, $Z_2 = X_1X_2$, and $Z_3 = X_2^2$. When $E(Y|X) = G(V_2)$, this is a single index model with a nonlinear index. Finally, when $E(Y|X) = G(V_1, V_2)$, this is a double index model with indices V_1 and V_2 . As index parameters are at most identified up to location and scale in the cases above, it is common to normalize the parameters so that the coefficient on one of the explanatory variables is one and the constant term is zero.

The most commonly used index structure is a single index assumption (See, for example, Ahn (1997), Climov, Delecroix and Simar (2002), Fraga and Martins (2001), Gerfin (1996), Gorgens (2000), Gorgens and Horowitz (1999), Ichimura (1993), Klein and Sherman (2002), Klein and Spady (1993)). To explain the importance of this single index assumption, consider a nonparametric setting in which there is no parametric information on the form of the expectation of the dependent variable conditioned on the explanatory variables, X . In this case, it is difficult to obtain an accurate estimate of the conditional expectation when the dimension of X is high. For example, with X a high dimensional vector of exogenous

variables, consider the binary response model:

$$Y = \begin{cases} 1 & : f(X) > u \\ 0 & : \text{otherwise} \end{cases},$$

where f and the distribution of u are both unknown and u is independent of X . In this nonparametric case, it is well known that it is difficult to obtain a good estimate of $Pr(Y = 1|X) = E(Y|X)$ due to the high dimension of X . In contrast, a semiparametric single index model avoids this problem by imposing a parametric structure on f . For example, $f(X) = X\theta_0$ aggregates the information in X into a single linear index and reduces the dimension to one. This linear index is the most commonly used.

As in the example above, let θ_0 be a parameter vector of interest and denote $\hat{\theta}$ as its estimator. To insure that this estimator has desirable properties, in large and finite samples, it is critical to have a good estimator for the conditional expectation, $E(Y|V(\theta))$. There are methods to estimate this expectation (e.g., those based on higher order kernels¹) that deliver $\hat{\theta}$ with desirable large sample properties but often with poor finite sample performance. The problem here is that in implementing a bias control to obtain asymptotic normality, such methods can fail to satisfy restrictions on the true conditional expectation of interest in finite samples. For example, in the binary model above, such methods can deliver estimated probabilities that fall outside of the interval $[0,1]$. In contrast, methods that impose these restrictions (e.g., those based on regular kernels), perform well in finite samples. However, in failing to adequately implement a bias control, these methods do not achieve \sqrt{N} -asymptotic normality.

¹With V_i i.i.d. distributed as g , a kernel density estimator for $g(t)$ is given as:

$$\hat{g}(t) = \sum \frac{1}{Nh} K[(t - V_i)/h].$$

When K is a density that is symmetric about zero (e.g., a standard normal), we refer to K as a regular kernel. In this case, it can be shown that the bias is $O(h^2)$, where h tends to zero at a rate given below. When K is a function that is symmetric about zero, integrates to one, and

$$\int z^{2p} K(z) dz = 0, \quad p = 1, 2, \dots,$$

then K is termed a higher order kernel. It can be shown that the bias in a density estimator based on this kernel is $O(h^{2(1+p)})$. Notice that unlike regular kernels, higher order kernels must take on negative values. Here, with a single index model, $p = 1$.

The objective of this chapter is to provide an estimator with desirable large sample properties and that also performs well in finite samples. To obtain these properties, we employ two alternative bias controls and regular kernels. The bias controls will insure normality, while the use of regular kernels allows us to impose known restrictions on the estimated conditional expectation.² We find that the resulting estimator performs well in finite samples, an important feature for its use in applications. It should be remarked that while this estimator was developed for single index models, it can be extended to the double index case (see Klein and Vella, 2008). Such models are important and naturally arise when one or more of the explanatory variables is endogenous. For an application in the context of healthcare decisions, see Shen (2008).

In organizing this chapter, we begin by discussing the moment conditions that characterize the estimator in Section 1.2. These conditions incorporate methods for controlling their bias using regular kernels. Section 1.3 contains assumptions and asymptotic results. Here, we will also outline the basic proof strategy, with the Appendix containing all formal and complete proofs. In section 1.5, we carry out Monte Carlo studies, where we evaluate the performance of the estimators in finite samples. To preview the results, we find that a bias corrected estimator based on regular kernels performs the best.

1.2 Moment Conditions and Bias Control

In describing these conditions and the nature of the bias controls, it will be useful to have simplified notation for sample averages of the quantities of interest. For this purpose, define:

$$\langle AB \rangle \equiv \sum_{i=1}^N [A_i B_i] / N; \quad \langle A/B \rangle \equiv \sum_{i=1}^N [A_i / B_i] / N$$

Further, we use the " Λ " symbol above a quantity of interest to indicate an estimator for it.

Then, letting $V(\theta_0) \equiv V(X; \theta_0)$ be a single index depending on explanatory variables, X ,

²There are other alternative methods that control for the bias under regular kernels. For example, Powell and Honore (2005) employ the following jackknife approach. Let $\hat{\beta}(h)$ be an estimator based on the window parameter h . Then, under this approach the final estimator is a linear combination of such estimators using different windows. In contrast, here we employ a two-stage approach that exploits a result due to Whitney Newey to insure asymptotic normality under regular kernels. In addition, we also implement a smoothing adjustment to the final estimator. We find that the resulting estimator performs quite well in finite samples. It is an open question as to whether or not a further improvement would be obtained if we jackknifed our estimator.

and on a vector of true parameter values, θ_0 , assume that we are interested in an extremum estimator for θ_0 . Let τ be a trimming function that controls for small denominators in a manner that we will make explicit below. In this section, for expositional simplicity, we take this trimming function as known. In the Appendix, we let this function depend on an estimated argument and show that it may be taken as known. Employing this trimming function, consider estimators whose gradients have the following structural form:

$$\hat{G}^*(\theta_0) \equiv \left\langle \left[Y - \hat{E}(Y|V(\theta_0)) \right] \tau \hat{W}^* \right\rangle,$$

where the weighting function, \hat{W}^* , has the form:

$$\hat{W}_i^* \equiv \hat{\alpha}(V_i(\theta_0)) \nabla \hat{E}_i$$

For SLS estimators (Ichimura (1993)), $\hat{\alpha}(V_i(\theta_0)) = 1$. For a QMLE estimator for binary response models (Klein and Spady (1993)), $\hat{\alpha}(V_i(\theta_0)) = 1/\hat{E}_i \left[(1 - \hat{E}_i) \right]$. For a QMLE estimator of ordered models (Klein and Sherman (2002)), the gradient consists of a number of components, all of which have the structure above. The weights differ above, but all consist of a function of the index and the derivative of a nonparametric expectation estimator. If the gradient, when normalized by \sqrt{N} is asymptotically distributed as normal, then it is not difficult to show that the underlying estimator of interest has an asymptotic normal distribution. Accordingly, in what follows, we focus on these gradient expressions.

For such estimators characterized by the gradient structure above, write the gradient as:

$$\begin{aligned} \sqrt{N} \hat{G}^*(\theta_0) &= \sqrt{N} \left[\hat{A}^*(\theta_0) - \hat{B}^*(\theta_0) \right] \\ \hat{A}^*(\theta_0) &\equiv \left\langle [Y - E(Y|V_0)] \tau \hat{W}^* \right\rangle \\ \hat{B}^*(\theta_0) &\equiv \left[\hat{E}(Y|V_0) - E(Y|V_0) \right] \tau \hat{W}^* \end{aligned}$$

For the first term, with the argument given in the Appendix, it can be shown that:

$$\sqrt{N} \left[\hat{A}^*(\theta_0) - A^*(\theta_0) \right] \xrightarrow{p} 0, \quad A^*(\theta_0) \equiv \langle [Y - E(Y|V_0)] \tau W^* \rangle$$

The second or B-component above contributes a bias to the estimator that we need to control in order to show that the gradient has an asymptotic normal distribution.

Below we will define $\hat{E}(Y|V_0)$ as a ratio of estimated functions: \hat{f}/\hat{g} , each of which converges to its true limiting value. We will be able to show that:³

$$\begin{aligned} \sqrt{N} \left[\hat{B}^*(\theta_0) - \hat{B}_S^* \right] &\xrightarrow{p} 0, \\ \hat{B}_S^* &= \left\langle \left[\hat{f}/\hat{g} - E(Y|V_0) \right] \tau \hat{W}^*(\hat{g}/g) \right\rangle \\ &= \left\langle \left[\hat{f} - \hat{g}E(Y|V_0) \right] \tau W^*/g \right\rangle + o_p(1). \end{aligned}$$

Since the above quantity is linear in the estimated components \hat{f} and \hat{g} , it is possible to control for the bias in \hat{B}_S^* by controlling for the bias in these estimated functions. Higher order kernels are commonly employed for this purpose. In this case a standard U-statistic projection argument, which we provide in the Appendix, immediately provides the result:

$$\begin{aligned} \sqrt{N} \left[\hat{B}_S^* - B_S^* \right] &\xrightarrow{p} 0, \\ B_S^* &= \langle [Y - E(Y|V_0)] \tau E[W^*|V_0] \rangle \end{aligned}$$

Asymptotic normality for the normalized gradient now follows from a standard central limit theorem. In using higher order kernels to control for the bias and deliver this result, it should be noted that such kernels can result in negative density estimates and (as is the case here) often do not perform as well as methods based on regular kernels that do not deliver the desired large sample properties. Here, we seek alternative bias controls that deliver the desired large sample results with regular kernels.

Recalling that the weight function contains the derivative of the expectation function, we exploit a property of this derivative due to Whitney Newey in the following theorem:⁴

³Note that:

$$\left\langle \left[\hat{f}/\hat{g} - E(Y|V_0) \right] \tau \hat{W}^* [(\hat{g}/g) - 1] \right\rangle \xrightarrow{p} 0$$

with the first and third term each converging to zero at a rate somewhat below $N^{-1/2}$. We will show in the Appendix that the overall or combined rate of the product is sufficient to provide the desired result.

⁴This result and its proof were provided to one of the authors in a private communication. The proof,

Theorem 0: With $V(\theta_0) \equiv V(X; \theta_0)$ as a single index, assume the following single index restriction holds:

$$E(Y|X) = E(Y|V(\theta_0)) \equiv F(V(\theta_0))$$

Then:

$$E[\nabla_{\theta} E(Y|V(\theta)) | V(\theta_0)]_{\theta=\theta_0} = 0.$$

Proof: Let $\delta(\theta) \equiv V(\theta_0) - V(\theta)$ and observe that $\delta(\theta_0) = 0$ and that $\nabla_{\theta} \delta(\theta) = -\nabla_{\theta} V(\theta)$. Then, employing the index restriction and using iterated expectations:

$$\begin{aligned} E(Y|V(\theta)) &= E_X[E(Y|V(\theta_0)) | V(\theta)] \\ &\equiv E_X[F[V(\theta_0)] | V(\theta)] \\ &= E_X[F[V(\theta) + \delta(\theta)] | V(\theta)] \\ &\equiv G(V(\theta), \delta(\theta)) \end{aligned}$$

Let G_k be the partial derivative of G taken w.r.t. θ in the k^{th} argument of G , $k = 1, 2$. From the chain rule:

$$\begin{aligned} \nabla_{\theta} G(V(\theta), \delta(\theta)) |_{\theta=\theta_0} &= G_1(V(\theta), 0) |_{\theta=\theta_0} + G_2(V(\theta_0), \delta(\theta)) |_{\theta=\theta_0} \\ &= \nabla_{\theta} F(V(\theta)) |_{\theta=\theta_0} - E[\nabla_{\theta} F(V(\theta)) | V(\theta_0)]_{\theta=\theta_0} \end{aligned}$$

The proof now follows.

From above, $\nabla_{\theta} E[Y|V(\theta)]_{\theta=\theta_0}$ behaves as an error component with conditional expectation 0. As this component enters multiplicatively into the gradient, we exploit its residual-like properties as a bias control. To utilize Newey's result, return to the gradient discussed above and let

$$H(V) \equiv E\left(\left[\hat{f} - \hat{g}E(Y|V_0)\right] | X\right)$$

which is very short and can be found in Klein and Sherman (2002), is also provided here.

Then, take an iterated expectation to obtain:

$$\begin{aligned}
 E \left[\hat{B}_S^* \right] &= \left\langle E_X E \left[\hat{f} - \hat{g} E(Y|V_0) \right] \tau W^* / g \mid X \right\rangle + o(1) \\
 &= \langle E_X (\tau H(V) W^* / g) \rangle \\
 &= \langle E_V \{ [H(V) / g] E[(\tau W^* | V)] \} \rangle
 \end{aligned}$$

If the trimming function, τ , depends on X , it is not possible to employ Newey's result and obtain 0 for this expectation. If the trimming depends on V , then this expectation would be zero by construction.

Based on the above observation, we consider a multi-stage estimation method. In the first stage, we trim on X and obtain consistent estimates for the index parameters. Using these parameter estimates, we construct an estimated index upon which to base trimming. In the second stage, we then trim on the basis of the (estimated) index rather than X . In so doing, the expected value of the gradient would be zero. However, such trimming upsets the consistency argument because it provides no protection for small denominators outside of a small neighborhood of the truth. To resolve this problem, we adjust expectations as follows. Recalling that $\hat{E} = \hat{f} / \hat{g}$, define an adjusted expectation as:

$$\hat{E}_a = \frac{\hat{f}}{\hat{g} + \Delta}$$

Below, we will define Δ such that it vanishes rapidly in regions where g is bounded away from zero. In regions where g tends to zero, Δ tends to zero very slowly. In this manner, we are able to preserve the consistency argument and establish asymptotic normality for the gradient.⁵ It is possible to further improve the performance of the estimator in finite samples under a smoothing adjustment, but we defer discussion of this issue until Section 1.3.

1.3 Assumptions, Definitions, and Results

To obtain the above results, we require standard assumptions on the data generating

⁵A similar strategy is employed in Klein and Spady (1993) so as to let trimming depend on an estimated density. That paper, however, relies on higher order kernels or local smoothing to obtain large sample results.

process, smoothness conditions on unknown densities, and given sets over which densities are positive. Assume:

(A1) Observations. With (Y_i, X_i) as the i^{th} observation on the dependent and explanatory variables, assume that (Y_i, X_i) is i.i.d. With X as the $N \times K$ matrix of observations on the explanatory variables (including a column vector of ones), assume that X has full column rank with probability 1.

(A2) Model. Under the null hypothesis $E(Y_i|X_i) = E(Y_i|V_i)$, $V_i \equiv X_{1i} + X_{2i}\theta_0$, where X_{1i} is continuous and θ_0 is in the interior of a compact parameter space, Θ . Furthermore, to simplify arguments we assume that X is bounded.⁶ In addition, $Var(Y_i|X_i)$ is bounded.

(A3) Estimator Characterization. Under the null hypothesis, with $-H_o$ positive definite and G_i being i.i.d., the estimator for θ_0 satisfies:

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta_0) &= -H_o^{-1}N^{-1/2}\sum_{i=1}^n G_i + o_p(1), \\ E(G_i) &= 0; \quad Var(G_i) = O(1). \end{aligned}$$

(A4) Continuous Variable Density. With X_k as any of the continuous X variables, denote $g_k(\bullet|y)$ as its density conditioned on $Y = y$. Denote $\nabla^d g_k(t|y)$ as the d^{th} partial derivative with respect to t , with $\nabla^o g_k(t|y) \equiv g_k(t|y)$. With g_k supported on $[a_k^*, b_k^*]$:

$$\begin{aligned} g_k &> 0 \text{ on } (a_k, b_k), \quad a_k^* < a_k < b_k < b_k^* \\ |\nabla^d g_k| &= O(1) \text{ on } [a_k, b_k], \quad d = 0, 1, 2, 3. \end{aligned}$$

⁶The assumption on X being bounded is not necessary, but simplifies several of the arguments.

(A5) Index Density. With $V_i \equiv X_{1i} + X_{2i}\theta_0$, let $g(x_1|y, x_2)$ be the indicated conditional density supported on $[a, b]$, Assume

$$\begin{aligned} g &> 0 \text{ on } (a, b) \\ |\nabla^d g| &= O(1) \text{ on } [a, b], \quad d = 0, 1, 2, 3. \end{aligned}$$

(A6) Tail Condition. With g_y as the density for the dependent variable, Y , assume that there exists T such that for $t > T$ and $df \geq 4$:

$$g_y(t) < 1/[(1 + t^2)^{(df+1)/2}].$$

The above assumptions are somewhat standard in the literature. Namely, the model must include a continuous variable (A2) and densities for continuous variables and the index must be sufficiently smooth, as implied by (A4-5). Notice that (A4-5) also specifies when density denominators become zero, which facilitate the trimming strategy. To establish uniform convergence results for estimated expectations, we require a tail condition on the density for the dependent variable, Y . While this assumption can be made in terms of the number of finite moments for Y , here we directly assume in (A6) that the density has tails that are no thinner than those for a t-distribution with $d \geq 4$. Additional window conditions will be required and are stated directly in the Theorems for which they are needed. To define the estimators, we will also require the definitions below.

(D1) Trimming. With Z_{ik} as the i^{th} observation on a continuous variable, Z_k , $k = 1, \dots, K$, let

$$\begin{aligned} \hat{\tau}_{ik} &\equiv \begin{cases} 1 & : \hat{a}_k < Z_{ik} < \hat{b}_k \\ 0 & : \text{otherwise,} \end{cases} \\ \hat{\tau}_i &\equiv \prod_k \hat{\tau}_{ik} \end{aligned}$$

where \hat{a}_k and \hat{b}_k are respectively lower and upper sample quantiles for Z_k . With X_k as an exogenous variable, when $Z_{ik} = X_{ik}$, we refer to $\hat{\tau}_i$ as X-trimming and write $\hat{\tau}_{ix} = \hat{\tau}_i$; with \hat{V} as the estimated index, when $Z_{ik} = \hat{V}$, $k = 1$, we refer to $\hat{\tau}_i$ as index-trimming and write $\hat{\tau}_{iv} = \hat{\tau}_i$.

In the case where a smooth trimming function is required, define:

$$\tau(z, \delta) \equiv [1 + \exp(-Ln(N)Ln(N)[z - \delta])]^{-1}$$

as a smoothed approximation to an indicator on $z \geq \delta$. A smoothed indicator on $z \in [a, b]$ is then defined as $\tau(z, a) - \tau(b, z)$.

(D2) Kernels. The kernel function $K(z)$ is termed regular if $K(z) \geq 0$, $\int K(z)dz = 1$, and $K(z) = K(-z)$. The function $K(z)$ will be termed a (normal) twicing kernel if $K(z) = 2\phi(z) - \phi(z/\sqrt{2})/\sqrt{2}$.

(D3) Expectations. With $h = O(N^{-r})$ and $K_{ij} \equiv K[(z_i - z_j)/h]$, the estimated conditional expectation with window parameter r is denoted as $\hat{E}_i \equiv \hat{E}(Y|Z = z_i)$ and is given by:

$$\hat{E}_i \equiv \left[\frac{1}{(N-1)h} \sum_{j \neq i} Y_j \hat{\tau}_j K_{ij} \right] / \left[\hat{\Delta}_i + \frac{1}{(N-1)h} \sum_{j \neq i} \hat{\tau}_j K_{ij} \right] \equiv \hat{f}_i / \hat{g}_i^*$$

The expectation is referred to as being:

- a) regular (\hat{E}) if $\hat{\tau}_j = 1$, $\hat{\Delta}_i = 0$, and K is a regular kernel.
- b) twicing if $\hat{\tau}_j = 1$, $\hat{\Delta}_i = 0$, and K is a (normal) twicing kernel (Newey, Hsieh, and Robins (2004)).
- c) adjusted (\hat{E}_a) if $\hat{\tau}_j = 1$, K is regular, and with \hat{q} as a lower sample quantile, (e.g., 0.01) of $\hat{g}(z_i)$, $i = 1, \dots, N$.

$$\hat{\Delta}_i \equiv h^\alpha \hat{q} \left[1 - \hat{\tau}_i(\hat{a}, \hat{b}) \right], 0 < \alpha < 1$$

(D4) First and Second Stage Estimators.⁷

$$\hat{\theta}_1 = \arg \max_{\theta} \hat{Q}_1, \quad \hat{Q}_1 \equiv -\frac{1}{2n} \sum_{i=1}^n \hat{\tau}_{ix} [Y_i - \tilde{E}(Y_i | v(X_i; \theta))]^2,$$

$$\hat{\theta}_2 = \arg \max_{\theta} \hat{Q}_2, \quad \hat{Q}_2 \equiv -\frac{1}{2n} \sum_{i=1}^n \hat{\tau}_{iv} [Y_i - \hat{E}_a(Y_i | v(X_i; \theta))]^2$$

(D5) Smoothing Adjustment. Letting $\hat{H}(\theta)$ be the Hessian w.r.t. \hat{Q}_2 , and \hat{E}^* be a regular expectation with window parameter $r^* = 1/5$, define:

$$\begin{aligned} \hat{B}(\hat{\theta}_2) &= \sum_{i=1}^n \hat{\tau}_{iv} (\hat{E}_i(\hat{\theta}_2) - E_i(\hat{\theta}_2)) \nabla \hat{E}_i(\hat{\theta}_2) \\ \hat{B}^*(\hat{\theta}_2) &= \sum_{i=1}^n \hat{\tau}_{iv} (\hat{E}_i^*(\hat{\theta}_2) - E_i(\hat{\theta}_2)) \nabla \hat{E}_i(\hat{\theta}_2) \end{aligned}$$

Then, define an adjusted estimator as:

$$\hat{\theta}^* = \hat{\theta}_2 - \hat{H}(\hat{\theta}_2)^{-1} [\hat{B}(\hat{\theta}_2) - \hat{B}^*(\hat{\theta}_2)]$$

As discussed earlier, we employ a two-stage estimator (D4) so as to utilize Newey's result as a bias control. The first stage of this estimator requires X-trimming (D1) and regular expectations (D3), while the second stage requires index-trimming (D1) and adjusted expectations (D3). We will compare results under regular and higher order kernels (D2). Notice that the twicing kernel in (D2) is a higher order kernel in that:

$$\begin{aligned} &\int z^2 \left[2\phi(z) - \phi\left(\frac{z}{\sqrt{2}}\right) / \sqrt{2} \right] dz \\ &= 2 - 2 \int [z/\sqrt{2}]^2 \phi\left(\frac{z}{\sqrt{2}}\right) / \sqrt{2} dz \\ &= 2 - 2 \int w^2 \phi(w) dw = 0, \quad w \equiv z/\sqrt{2} \end{aligned}$$

⁷As discussed earlier, there are many different estimators to which this chapter applies. We focus on variants of the SLS estimator so as to employ the same estimator over designs where the dependent variable is continuous or discrete.

In examining the second stage estimator for various designs, we had one design where the finite sample bias for the estimator was significantly larger than that for the other designs. In this case, the smoothing adjustment (D5) improved the properties of our estimator significantly in this case by reducing the bias in the estimator. To explain why this adjustment "works", recall the definition of $\hat{A}(\theta_0) - \hat{B}(\theta_0)$ in the previous section. Then, a standard Taylor expansion yields:

$$\hat{\theta}_2 - \theta_0 = -\hat{H}^{-1}(\theta^+)(\hat{A}(\theta_0) - \hat{B}(\theta_0)), \theta^+ \in [\hat{\theta}_2, \theta_0].$$

Defining an estimator with an infeasible adjustment as:

$$\hat{\theta}_I = \hat{\theta}_2 - \hat{H}^{-1}(\theta^+) [\hat{B}(\theta_0) - \hat{B}^*(\theta_0)],$$

then it immediately follows that

$$\hat{\theta}_I - \theta_0 = -\hat{H}^{-1}(\theta^+) [\hat{A}(\theta_0) - \hat{B}^*(\theta_0)].$$

This infeasible estimator is the same as $\hat{\theta}_2$, except the B -component now depends on an optimal expectation estimator. As a result, we would expect it to perform better in finite samples. Below, we show that this infeasible estimator can be approximated by the feasible estimator based on the adjustment in (D5) in that:

$$\sqrt{N} [\hat{\theta}_I - \hat{\theta}_2] \xrightarrow{p} 0.$$

Beginning with the estimator, Theorem A.1 below establishes consistency at both stages.

Theorem A.1: (Estimator Consistency). With $df \geq 4$ given in (A6), set $\lambda \equiv df / (1 - df)$. Denote $\hat{\theta}_1$ and $\hat{\theta}_2$ as the first and second stage estimators respectively and assume (A1-6). Base the first-stage estimator on a regular expectation (D3) with window r_1 :

$$1/8 < r_1 < 1/6; \quad 0 < r_1 < [1/2 - \delta] / [\lambda + \varepsilon]$$

Base the second-stage estimator on an adjusted expectation (D3) with adjustment parameter $0 < \alpha < 1$ and window r_2 :

$$1/8 < r_2 < 1/6; \text{ and } 0 < r_2 < [1/2 - \delta] / [\lambda(1 + a) + \varepsilon],$$

Then:

$$\left| \hat{\theta}_1 - \theta_o \right| = o_p(1); \quad \left| \hat{\theta}_2 - \theta_o \right| = o_p(1).$$

The normality arguments are based on moment conditions. After providing these results in Theorems A.2-3 below, we will outline the common structure of the argument.

Theorem A.2: (Estimator: Asymptotic Linearity and Normality). Assume (A1-6) and base the second stage estimator, $\hat{\theta}_2$, on an adjusted expectation (D3) with adjustment and window parameters as given in Theorem A.1. Letting $G(\theta_0) \equiv \nabla_{\theta'} Q_2(\theta_0)$, $H_0 \equiv \nabla_{\theta\theta'} Q_2(\theta_0)$, and $\Sigma \equiv H_0^{-1} E \left[\sqrt{N} G_0 G_0' \sqrt{N} \right] H_0^{-1}$:

$$\begin{aligned} a) & : \quad \left| \hat{\theta}_1 - \theta_o \right| = o_p \left(N^{-1/4} \right) \\ b) & : \quad \sqrt{N} \left(\hat{\theta}_2 - \theta_0 \right) = -H_0^{-1} \sqrt{N} G(\theta_0) + o_p(1) \\ c) & : \quad \sqrt{N} \left(\hat{\theta}_2^* - \hat{\theta}_2 \right) = o_p(1) \\ d) & : \quad \sqrt{N} \left(\hat{\theta}_2^* - \theta_0 \right) \xrightarrow{d} Z \sim N(0, \Sigma) \end{aligned}$$

In the special case when $Var(Y_i|X_i) = \sigma_o^2$ is constant, $\Sigma = -\sigma_o^2 H_0^{-1}$.

To outline the proof for Theorem A.2b (other parts follow directly), note that the moment conditions underlying the estimator have the structure:

$$\sqrt{N} \left\langle \hat{\tau}(Y - \hat{M}) \hat{\omega} \right\rangle = \sqrt{N} \left\langle \hat{\tau}(Y - M) \hat{\omega} \right\rangle - \sqrt{N} \left\langle \hat{\tau}(\hat{M} - M) \hat{\omega} \right\rangle.$$

Here $\hat{\omega}$ is an estimated weight vector whose form is given above. Denote ω as the limiting value of the estimated weight. Then, for the first component above, a mean-squared convergence argument is employed in the Appendix together with a result from Pakes and

Pollard (1989) to show that :

$$\sqrt{N} \langle \hat{\tau}(Y - M)\hat{\omega} \rangle = \sqrt{N} \langle \tau(Y - M)\omega \rangle + o_p(1).$$

With $\hat{M} = \hat{f}/\hat{g}$, in the Appendix we show that the second component is within $o_p(1)$ of

$$\sqrt{N} \langle \tau(\hat{M} - M)\omega\hat{g}/g \rangle = \sqrt{N} \langle \tau(\hat{f} - \hat{g}M)\tau\omega/g \rangle.$$

As a U-statistic, this last term can be analyzed by conventional projection arguments. Provided that its expectation tends to zero, this term vanishes for the estimator. As discussed in section 1.2, the above expression will have expectation tending to zero if appropriate higher kernels are employed or when the trimming function, τ , depends only on the index. Regular kernels can be employed as this last condition holds.

1.4 Monte Carlo Designs and Results

In this section Monte Carlo experiments are used to investigate different estimators. The estimators examined here are defined in Section 1.3. Our Monte Carlo study shows that the two-stage normal kernel estimator with bias correction has the smallest root mean-square error (RMSE).

1.4.1 Designs

All of the designs have single index structures. In all the designs, we normalize such that $E(Y_i|X_i)$ has standard deviation 2.

In the first (basic) design, we use the following data generating process:

$$Y_i = M_{0i} + \varepsilon_i, \quad M_{0i} \propto (X_{1i} + X_{2i})^2,$$

where the X 's $\sim \chi^2(1)$ and $\varepsilon \sim N(0, 1)$.

The second design is a binary response design. With ε_i being i.i.d. $N(0, 1)$, under the null of a single index model:

$$Y_i = \begin{cases} 1 : & M_{0i} > \varepsilon_i \\ 0 : & \text{otherwise} \end{cases}, M_{0i} \propto X_{1i} + X_{2i} - 0.5$$

Unlike the previous two designs, here the two X 's are correlated. In particular, the X 's are linear combinations of the same χ^2 variable and different normal shocks.

Since discrete independent variables are very common in practice, the third design has a discrete regressor. The structure of the data generating process is the same as in the basic design, however, here X_{2i} is a binary variable.

In the fourth (general linear model) design, we generate data by:

$$Y_i = M_{0i} + \varepsilon_i, M_{0i} \propto (X_{1i} + X_{2i} + 2X_{3i})^2,$$

where the X 's $\sim \chi^2(1)$ and $\varepsilon \sim N(0, 1)$.

For all the designs, the sample size we use is $n=1000$, and the number of Monte Carlo replications is 1000.

There are a number of window and trimming choices that need to be specified. With windows having the form $h = O(N^{-r})$, for the stage1 and stage2 estimators, we set r at 1/6.1. Within the range of permissible values, the value gives the fastest point-wise convergence rate of the estimated expectation to the truth. For the smoothing adjustment, we select an optimal pointwise rate of 1/5. Finally, for the twicing kernel, we set this window at 1/7. In the case of trimming, all trimming is based on the .99 quantile for the relevant variables. Recall that in the second stage estimator, we adjust the denominator of estimated expectations. Here, we smoothly keep the index between the .005 and the .995 index quantiles. Recall also that this adjustment depends on a lower density quantile, and we select the .01 quantile for this purpose. Finally the adjustment depends multiplicatively on a window raised to the power of α , $0 < \alpha < 1$. In this case, we set α to be 1/2.

1.4.2 Monte Carlo Results

The estimators studied are SLS estimators using twicing kernels (SLS-TW), using X-trimming in the first stage (S1SLS), and using index-trimming in the second stage (S2SLS).

For each SLS variant there are two versions: having smoothing correction or not; the corrected ones have an extra "C" in front (e.g., CSLS-TW). Among the unadjusted estimators, in the general linear model and basic designs, S2SLS has an RMSE about 30% lower than that of the other estimators. The reduction in RMSE is smaller (8%) in the binary response case and close to zero in the discrete regressor design reported here. This last finding is design dependent and does not hold for other discrete designs.⁸

In terms of RMSE, bias adjusted estimators are quite close to uncorrected ones in all the designs except in the discrete regressor case, where it reduced RMSE by about 16% by cutting the bias in half. Essentially, the bias correction makes little difference when the bias in the uncorrected estimator is very small, but can have a large impact when this bias is large. Hence our conclusion would be that the bias corrected two-stage normal kernel estimator with bias reducing structure is the best choice. Detailed results can be found in Table 1.1 Estimation Results. Note that with exception of the discrete regressor design, in all the other designs the twicing kernel design are not reported because there are severe outliers resulting in misleading bias and variance values.

1.5 Conclusions

In summary, we have developed an estimator that has desirable large sample properties (consistency and asymptotic normality), and that also performs well in finite samples. We have obtained these properties by employing bias controls that make it possible to base the estimator on regular kernels. These finite and large sample properties are important in applied work.

⁸Specifically, we interchanged quadratic and cubic components so that the conditional mean function was cubic under the null. For this case, we found that the gain is substantial as is shown in the detailed table below:

An Alternative Discrete Regressor Design						
		SLS-TW	CSLS-TW	S1SLS	S2SLS	CS2SLS
Discrete Regressor (Flipped)	Bias	0.065	0.032	0.037	0.060	0.038
	Rvar	0.101	0.116	0.093	0.067	0.069
	Rmse	0.120	0.121	0.100	0.090	0.079

Chapter 2

A Test of the Single Index Assumption in Semiparametric Models

2.1 Introduction

In the previous chapter, we defined the index restrictions in semiparametric models. In particular, a single index assumption is commonly used in the semiparametric literature. For example, with

$$V \equiv X_1\beta_1 + \dots + X_K\beta_K$$

and with $X = [X_1, \dots, X_K]$ a single index assumption holds if: $E(Y|X) = E(Y|V)$. When, the single index assumption does not hold, the imposition of the single index assumption will result in an inconsistent estimator for the conditional expectation of interest.¹ Given the sensitivity of the estimator to a single index assumption and given its wide use, the objective of this chapter is to formulate a test for this assumption.

To test the single index assumption, it is critical to have a good estimator for the conditional expectation, $E(Y|V(\theta))$. To get a good estimator, we need a good parameter estimate and also a good method to estimate the expectation. Here, we employ the bias-corrected estimator discussed in the previous chapter, which has both good finite sample and large sample properties. For the expectation itself, there are methods to estimate this expectation (e.g., those based on higher order kernels²) that deliver desirable large sample properties but often with poor finite sample performance. The problem here is

¹As an alternative example of a double index model, return to the binary response model discussed earlier, and let the error term have a conditional variance that depends on another index. Namely, let $u = s(X\beta_o)\varepsilon$, where ε is independent of X .

²With V_i i.i.d. distributed as g , a kernel density estimator for $g(t)$ is given as:

$$\hat{g}(t) = \sum \frac{1}{Nh} K[(t - V_i)/h].$$

When K is a density that is symmetric about zero (e.g., a standard normal), we refer to K as a regular kernel. In this case, it can be shown that the bias is $O(h^2)$, where h tends to zero at a rate given below. When K is a function that is symmetric about zero, integrates to one, and

$$\int z^{2p} K(z) dz = 0, \quad p = 1, 2, \dots,$$

then K is termed a higher order kernel. It can be shown that the bias in a density estimator based on this kernel is $O(h^{2(1+p)})$. Notice that unlike regular kernels, higher order kernels must take on negative values. Here, with a single index model, $p = 1$.

that in implementing a bias control to obtain asymptotic normality, such methods can fail to satisfy restrictions on the true conditional expectation of interest in finite samples. For example, in the binary model above, such methods can deliver estimated probabilities that fall outside of the interval $[0,1]$. In contrast, methods that impose these restrictions (e.g., those based on regular kernels), perform well in finite samples. However, in failing to adequately implement a bias control, these methods do not achieve \sqrt{N} -asymptotic normality.

There have been papers in the literature on testing parametric against semiparametric models (e.g., Härdle, Mammen, and Müller (1998), Härdle, Spokoiny, and Sperlich (1997), Horowitz and Härdle(1994)). Related tests of parametric models are given by Newey's (1985) paper on conditional moment tests and Bieren(1990) conditional moment test. This chapter differs from those above in that it formulates a test for a main assumption in semiparametric models. We note that in a likelihood context with a parametric null hypothesis, Newey develops conditional moment tests that have optimal local power properties. It may be possible to extend these results to the present context, but this extension is beyond the scope of the current chapter.

There have been some papers which focus on testing single index restrictions. Escanciano and Song (2007) provide a test focusing on average marginal effects and show that it has a minimax property; Andrews (1993) provides high level conditions for testing moment restrictions. This chapter differs from these in that we provide primitive conditions for a conditional moment test and for the estimator on which it is based. Tripathi and Kitamura (2001) employ an empirical likelihood approach for testing moment conditions of the form: $E[G(z, \theta)|X] = 0$. With G a known function and X continuous, they establish an optimality property for their test. The test proposed here is also an orthogonality test in that we test whether a function G is correlated with functions $M(X)$. However, unlike the above test, here the function G will be unknown as we set $G = Y - E(Y|V)$, where V is an index and the conditional expectation function $E(Y|V)$ is unknown. Further, the M -functions will be unknown and we will require nonparametric estimates of them. This feature is needed to insure that the form of test statistic adapts to the model under the alternative hypothesis. The proposed statistic also differs from those in the literature in the bias control mechanism

that it employs. This mechanism is similar to that underlying the estimator proposed in the previous chapter, and results in a test statistic that has good size and power properties in finite samples.

In organizing this chapter, we begin by discussing the moment conditions that characterize the test statistic in Section 2.2. These conditions incorporate methods for controlling their bias using regular kernels. Section 2.3 contains assumptions and asymptotic results. Here, we will also outline the basic proof strategy, with the Appendix containing all formal and complete proofs. In section 2.5, we carry out Monte Carlo studies, where we evaluate the performance of the test statistics in finite samples. To preview the results, we find that the bias-corrected form of the test statistic has good size and power properties.

2.2 Moment Conditions and Bias Control

In what follows, we will first consider a general test of moment conditions and then specialize it to the test for the single index restriction. Consider test statistics based on moment conditions of the form:

$$G_{kT}(\theta_0) \equiv \langle [Y - E(Y|V(\theta_0))] W_{Tk} \rangle, k = 1, \dots, K$$

where $W_{Tk} = W_{Tk}(X_k)$ is a vector of observations on a function of the k^{th} exogenous variable, X_k . Define G_T as a column vector with G_{kT} as the k^{th} element. Under a null hypothesis of interest, H_0 , we assume that the following orthogonality condition holds:

$$E[G_T(\theta_0)] = 0,$$

In testing whether or not these conditions hold, we allow the conditional expectation $E(Y|V)$ and the weight W_k to be unknown functions that can be estimated nonparametrically. Accordingly, write the estimated k^{th} moment as:

$$\hat{G}_k(\hat{\theta}) \equiv \langle [Y - \hat{E}(Y|V(\hat{\theta}))] \hat{W}_{Tk} \rangle.$$

With a test statistic based on these estimated moments, we will need to show that $\sqrt{N}\hat{G}_{kT}(\hat{\theta})$

has an asymptotic normal distribution under the null hypothesis of interest. Employing a standard Taylor expansion and with $\hat{\theta}$ as a \sqrt{N} -consistent estimator that has an asymptotic linear representation³, we will be able to write

$$\sqrt{N}\hat{G}_{kT}(\hat{\theta}) = \sqrt{N}\hat{G}_{kT}(\theta_0) + \nabla E[G_{kT}(\theta_0)]\sqrt{N}[\hat{\theta} - \theta_0] + o_p(1).$$

As the second or parameter-uncertainty component poses no difficulty, here we focus on the first component and discuss the nature of the bias control that we employ.

With $V_o \equiv V(\theta_0)$, we will be able to decompose these moment conditions in the same form as the gradient for the estimator above and write

$$\begin{aligned} \sqrt{N}\hat{G}_{kT}(\theta_0) &= \sqrt{N}[A_T(\theta_0) - \hat{B}_T(\theta_0)] + o_p(1) \\ A_T(\theta_0) &\equiv [Y - E(Y|V_0)]W_{Tk} \\ \hat{B}_T(\theta_0) &\equiv \left\langle \left[\hat{E}(Y|V_0) - E(Y|V_0) \right] \hat{W}_{Tk} \right\rangle \end{aligned}$$

As for the estimator, here the second or B-component contributes a bias that we seek to control. As above, we will be able to show:

$$\begin{aligned} \sqrt{N}[\hat{B}_T(\theta_0) - \hat{B}_S] &\xrightarrow{p} 0, \\ \hat{B}_S &= \left\langle \left[\hat{f}/\hat{g} - E(Y|V_0) \right] \hat{W}_{Tk}(\hat{g}/g) \right\rangle \\ &= \left\langle \left[\hat{f} - \hat{g}E(Y|V_0) \right] W_{Tk}/g \right\rangle + o_p(1) \end{aligned}$$

Using higher order kernels to control for the bias in \hat{f} and \hat{g} , in a standard U-statistic argument, which is provided in the Appendix, we can show:

$$\begin{aligned} \sqrt{N}[\hat{B}_S - B] &\xrightarrow{p} 0, \\ B &= \langle [Y - E(Y|V_0)]E[W_k|V_0] \rangle \end{aligned}$$

³The estimators we consider are all of the form:

$$\sqrt{N}[\hat{\theta} - \theta_0] = -H_o^{-1}\sqrt{N}\langle G \rangle,$$

where H_o is the Hessian matrix, and $\sqrt{N}\langle G \rangle$ is asymptotically distributed as $N(0, \Sigma)$.

The moment condition in large samples now has a form to which a central limit theorem would apply under the null hypothesis. Namely:

$$\sqrt{N}\hat{G}_{kT}(\theta_0) = \sqrt{N}\langle [Y - E(Y|V_0)][W_k - E(W_k|V_0)] \rangle + o_p(1)$$

Taking estimation uncertainty into account, the "full" gradient has the form:

$$\sqrt{N}\hat{G}_{kT}(\hat{\theta}) = \sqrt{N}\hat{G}_{kT}(\theta_0) + \nabla E[G_{kT}(\theta_0)]\sqrt{N}[\hat{\theta} - \theta_0]$$

Letting $G(\hat{\theta})$ be the vector with k^{th} element $\hat{G}_{kT}(\hat{\theta})$, the test statistic is then given by a standard quadratic form:

$$T \equiv \sqrt{N}G(\hat{\theta})' \hat{\Sigma}^{-1} \sqrt{N}G(\hat{\theta}),$$

where $\hat{\Sigma}$ is a consistent estimator for the covariance matrix of $\sqrt{N}G(\hat{\theta})$. Various alternative estimators for this covariance matrix will be provided below and examined in the Monte Carlo section.

As in the case for the estimator, we find that the test statistic based on higher order kernels can be dominated by one based on an alternative bias control and regular kernels. Unfortunately, the weight need not and will not here have the residual property of the derivative weight entering the gradient for the estimator. Therefore, we propose to recenter the weight so that it has the same residual-like property as in the estimator case. Namely, with $\hat{V} \equiv V(\hat{\theta})$ define $\hat{G}^*(\hat{\theta})$ as a vector with the k^{th} element being:

$$\begin{aligned} \hat{G}_{kT}^*(\hat{\theta}) &\equiv \langle [Y - \hat{E}(Y|V(\hat{\theta}))] \hat{W}_k^* \rangle \\ \hat{W}_k^* &\equiv \hat{W}_k - E[\hat{W}_k^* | \hat{V}] \\ T^* &\equiv \sqrt{N}\hat{G}^*(\hat{\theta})' \hat{\Sigma}^{-1} \sqrt{N}\hat{G}^*(\hat{\theta}) \end{aligned}$$

For the test statistic proposed below, we will show that such recentering provides a bias control that makes it possible to employ regular kernels and still obtain the same large sample result obtained under higher order kernels. Namely, we will show that T^* is close in

probability to T , with T^* having a χ^2 distribution. We find below that T^* , which is based on this alternative bias control, has much better finite sample properties than T .

To specialize the above moment conditions and develop a corresponding test statistic (a quadratic form in the moments) for the single index assumption, we need to specify the weight function. A natural choice for this function would not be a function of any particular exogenous variable, but rather the full conditional expectation: $E(Y|X)$. In this case, the expected moment condition becomes:

$$\begin{aligned} & E ([Y - E(Y|V)] E(Y|X)) \\ &= E ([E(Y|X) - E(Y|V)] E(Y|X)) \\ &= E \left([E(Y|X) - E(Y|V)]^2 \right) \end{aligned}$$

Notice that this expected moment condition is zero iff $E(Y|X) = E(Y|V)$. The above weight would seem natural as the expected moment condition reduces to the distance between nonparametric and index expectations. However, it is difficult to obtain reasonable estimates of the full conditional expectation $W = E(Y|X)$ when the dimension of X is large. We are therefore motivated to seek low dimensional weights that are close to this full expectation. With "close" defined in a mean-squared error sense, low dimensional weights are given by:

$$W_k = \arg \min_{\omega} [E(W - \omega)^2 | X_k] = E(Y | X_k)$$

Notice that this weight depends on the actual form of the dependence of Y on X . In other words, it is adaptive to the alternative model. This property is desirable compared to fixed weights, because intuitively it yields better test power by being able to flexibly capture different violations of the null hypothesis. Our Monte Carlo study comparing one common fixed weight and our adaptive weight confirms the above observation. The fixed weight we use is the quadratic weight. Detailed discussions are in the Monte Carlo section.

2.3 Assumptions, Definitions, and Results

To obtain the above results, we require standard assumptions on the data generating

process, smoothness conditions on unknown densities, and given sets over which densities are positive. Assume:

(A1) Observations. With (Y_i, X_i) as the i^{th} observation on the dependent and explanatory variables, assume that (Y_i, X_i) is i.i.d. With X as the $N \times K$ matrix of observations on the explanatory variables (including a column vector of ones), assume that X has full column rank with probability 1.

(A2) Model. Under the null hypothesis $E(Y_i|X_i) = E(Y_i|V_i)$, $V_i \equiv X_{1i} + X_{2i}\theta_0$, where X_{1i} is continuous and θ_0 is in the interior of a compact parameter space, Θ . Furthermore, to simplify arguments we assume that X is bounded.⁴ In addition, $\text{Var}(Y_i|X_i)$ is bounded.

(A3) Estimator Characterization. Under the null hypothesis, with $-H_o$ positive definite and G_i being i.i.d., the estimator for θ_0 satisfies:

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta_0) &= -H_o^{-1}N^{-1/2}\sum_{i=1}^n G_i + o_p(1), \\ E(G_i) &= 0; \text{Var}(G_i) = O(1). \end{aligned}$$

(A4) Continuous Variable Density. With X_k as any of the continuous X variables, denote $g_k(\bullet|y)$ as its density conditioned on $Y = y$. Denote $\nabla^d g_k(t|y)$ as the d^{th} partial derivative with respect to t , with $\nabla^o g_k(t|y) \equiv g_k(t|y)$. With g_k supported on $[a_k^*, b_k^*]$:

$$\begin{aligned} g_k &> 0 \text{ on } (a_k, b_k), \quad a_k^* < a_k < b_k < b_k^* \\ |\nabla^d g_k| &= O(1) \text{ on } [a_k, b_k], \quad d = 0, 1, 2, 3. \end{aligned}$$

⁴The assumption on X being bounded is not necessary, but simplifies several of the arguments.

(A5) Index Density. With $V_i \equiv X_{1i} + X_{2i}\theta_0$, let $g(x_1|y, x_2)$ be the indicated conditional density supported on $[a, b]$, Assume

$$\begin{aligned} g &> 0 \text{ on } (a, b) \\ |\nabla^d g| &= O(1) \text{ on } [a, b], \quad d = 0, 1, 2, 3. \end{aligned}$$

(A6) Tail Condition. With g_y as the density for the dependent variable, Y , assume that there exists T such that for $t > T$ and $df \geq 4$:

$$g_y(t) < 1/[(1 + t^2)^{(df+1)/2}].$$

The above assumptions are somewhat standard in the literature. Namely, the model must include a continuous variable (A2) and densities for continuous variables and the index must be sufficiently smooth, as implied by (A4-5). Notice that (A4-5) also specifies when density denominators become zero, which facilitate the trimming strategy. To establish uniform convergence results for estimated expectations, we require a tail condition on the density for the dependent variable, Y . While this assumption can be made in terms of the number of finite moments for Y , here we directly assume in (A6) that the density has tails that are no thinner than those for a t-distribution with $d \geq 4$. Additional window conditions will be required and are stated directly in the Theorems for which they are needed. To define the test statistics, we will also require the definitions below.

(D1) Trimming. With Z_{ik} as the i^{th} observation on a continuous variable, Z_k , $k = 1, \dots, K$, let

$$\begin{aligned} \hat{\tau}_{ik} &\equiv \begin{cases} 1 & : \hat{a}_k < Z_{ik} < \hat{b}_k \\ 0 & : \text{otherwise,} \end{cases} \\ \hat{\tau}_i &\equiv \prod_k \hat{\tau}_{ik} \end{aligned}$$

where \hat{a}_k and \hat{b}_k are respectively lower and upper sample quantiles for Z_k . With X_k as an exogenous variable, when $Z_{ik} = X_{ik}$, we refer to $\hat{\tau}_i$ as X-trimming and write $\hat{\tau}_{ix} = \hat{\tau}_i$; with \hat{V} as the estimated index, when $Z_{ik} = \hat{V}$, $k = 1$, we refer to $\hat{\tau}_i$ as index-trimming and write $\hat{\tau}_{iv} = \hat{\tau}_i$.

In the case where a smooth trimming function is required, define:

$$\tau(z, \delta) \equiv [1 + \exp(-Ln(N)Ln(N)[z - \delta])]^{-1}$$

as a smoothed approximation to an indicator on $z \geq \delta$. A smoothed indicator on $z \in [a, b]$ is then defined as $\tau(z, a) - \tau(b, z)$.

(D2) Kernels. The kernel function $K(z)$ is termed regular if $K(z) \geq 0$, $\int K(z)dz = 1$, and $K(z) = K(-z)$. The function $K(z)$ will be termed a (normal) twicing kernel if $K(z) = 2\phi(z) - \phi(z/\sqrt{2})/\sqrt{2}$.

(D3) Expectations. With $h = O(N^{-r})$ and $K_{ij} \equiv K[(z_i - z_j)/h]$, the estimated conditional expectation with window parameter r is denoted as $\hat{E}_i \equiv \hat{E}(Y|Z = z_i)$ and is given by:

$$\hat{E}_i \equiv \left[\frac{1}{(N-1)h} \sum_{j \neq i} Y_j \hat{\tau}_j K_{ij} \right] / \left[\hat{\Delta}_i + \frac{1}{(N-1)h} \sum_{j \neq i} \hat{\tau}_j K_{ij} \right] \equiv \hat{f}_i / \hat{g}_i^*$$

The expectation is referred to as being:

- a) regular (\hat{E}) if $\hat{\tau}_j = 1$, $\hat{\Delta}_i = 0$, and K is a regular kernel.
- b) twicing if $\hat{\tau}_j = 1$, $\hat{\Delta}_i = 0$, and K is a (normal) twicing kernel (Newey, Hsieh, and Robins (2004)).
- c) adjusted (\hat{E}_a) if $\hat{\tau}_j = 1$, K is regular, and with \hat{q} as a lower sample quantile, (e.g., 0.01) of $\hat{g}(z_i)$, $i = 1, \dots, N$.

$$\hat{\Delta}_i \equiv h^\alpha \hat{q} \left[1 - \hat{\tau}_i(\hat{a}, \hat{b}) \right], 0 < \alpha < 1$$

(D4) Two Estimation Stages**(D5) Smoothing Adjustment for the estimator**

(D6) Test Statistics. The test statistics, T and T^* , are defined as above.

As discussed earlier, we employ a two-stage estimator (D4) so as to utilize Newey's result as a bias control. The first stage of this estimator requires X-trimming (D1) and regular expectations (D3), while the second stage requires index-trimming (D1) and adjusted expectations (D3). We will compare results under regular and higher order kernels (D2). Notice that the twicing kernel in (D2) is a higher order kernel in that:

$$\begin{aligned} & \int z^2 \left[2\phi(z) - \phi\left(z/\sqrt{2}\right) / \sqrt{2} \right] dz \\ &= 2 - 2 \int [z/\sqrt{2}]^2 \phi\left(z/\sqrt{2}\right) / \sqrt{2} dz \\ &= 2 - 2 \int w^2 \phi(w) dw = 0, \quad w \equiv z/\sqrt{2} \end{aligned}$$

In examining the second stage estimator and the test statistic for various designs, we had one design where the finite sample bias for the estimator was significantly larger than that for the other designs. As a result, we found that the test statistic had poor size properties in this case. The smoothing adjustment (D5) improved the size properties of our test statistic significantly in this case by reducing the bias in the estimator. To explain why this adjustment "works", recall the definition of $\hat{A}(\theta_0) - \hat{B}(\theta_0)$ in the previous section. Then, a standard Taylor expansion yields:

$$\hat{\theta}_2 - \theta_0 = -\hat{H}^{-1}(\theta^+) (\hat{A}(\theta_0) - \hat{B}(\theta_0)), \quad \theta^+ \in [\hat{\theta}_2, \theta_0].$$

Defining an estimator with an infeasible adjustment as:

$$\hat{\theta}_I = \hat{\theta}_2 - \hat{H}^{-1}(\theta^+) \left[\hat{B}(\theta_0) - \hat{B}^*(\theta_0) \right],$$

then it immediately follows that

$$\hat{\theta}_I - \theta_0 = -\hat{H}^{-1}(\theta^+) \left[\hat{A}(\theta_0) - \hat{B}^*(\theta_0) \right].$$

This infeasible estimator is the same as $\hat{\theta}_2$, except the B -component now depends on an optimal expectation estimator. As a result, we would expect it to perform better in finite samples. Below, we show that this infeasible estimator can be approximated by the feasible estimator based on the adjustment in (D5) in that:

$$\sqrt{N} \left[\hat{\theta}_I - \hat{\theta}_2^* \right] \xrightarrow{p} 0.$$

Recall that Theorems A.1-2 of the previous chapter established that the estimator we employed has the required form for the test statistic employed here, in Theorem A.3 below we examine the large sample properties of the test statistic.

Theorem A.3. (Test Statistic: Asymptotic Null-Distribution): Let

$$\hat{M} \equiv \hat{E}(Y|\hat{V}); \hat{M}_k \equiv \hat{E}(Y|X_k); \hat{M}_T \equiv \hat{E}_T(Y|\hat{V}),$$

where the first two expectations are regular with window parameter $r : 1/6 < r < 1/4$ and the third is twicing with window parameter $r_T : 1/8 < r_T < 1/6$. Define:

$$\hat{w}_k \equiv \hat{\tau}_k \hat{M}_k; \hat{w}_k^* \equiv \hat{w}_k - \hat{E}(\hat{w}_k|\hat{V}),$$

where the above expectation is a regular with window parameter $r^* : 1/6 < r^* < r < 1/4$. Denote \hat{T} and \hat{T}^* as the un-centered and centered moments with respective k^{th} elements:

$$\hat{T}_k(\hat{\theta}) = \langle \hat{\tau}_v (Y - \hat{M}_T) \hat{w}_k \rangle; \hat{T}_k^*(\hat{\theta}) = \langle (Y - \hat{M}) \hat{w}_k^* \rangle$$

Then, with $\varepsilon \equiv (Y - M)$, under the null hypothesis of a single index:

- a) : $\sqrt{N} \hat{T}_k^*(\hat{\theta}) = \sqrt{N} S_k + o_p(1)$, $S_k = \langle \varepsilon w_k^* \rangle - \langle \nabla_{\theta} w_k^* \rangle H_0^{-1} G_o$
- b) : $\sqrt{N} \left[\hat{T}_k^*(\hat{\theta}) - \hat{T}_k(\hat{\theta}) \right] = o_p(1)$
- c) : $\sqrt{N} T' \Sigma^{-1} T \xrightarrow{d} \mathcal{X} \sim \chi^2(K)$, $T = \hat{T}^*(\hat{\theta}), \hat{T}(\hat{\theta})$,

where with S_k as the k^{th} element of S : $\Sigma \equiv E[SS']$.

To outline the proof for A.3a (other parts follow directly), note that the moment conditions underlying the test statistic have the structure:

$$\sqrt{N} \langle \hat{\tau}(Y - \hat{M})\hat{\omega} \rangle = \sqrt{N} \langle \hat{\tau}(Y - M)\hat{\omega} \rangle - \sqrt{N} \langle \hat{\tau}(\hat{M} - M)\hat{\omega} \rangle.$$

Here $\hat{\omega}$ is an estimated weight vector whose form is given above, depending on whether the above moment conditions describe the estimator or the test statistic. Denote ω as the limiting value of the estimated weight. Then, for the first component above, a mean-squared convergence argument is employed in the Appendix together with a result from Pakes and Pollard (1989) to show that :

$$\sqrt{N} \langle \hat{\tau}(Y - M)\hat{\omega} \rangle = \sqrt{N} \langle \tau(Y - M)\omega \rangle + o_p(1).$$

With $\hat{M} = \hat{f}/\hat{g}$, in the Appendix we show that the second component is within $o_p(1)$ of

$$\sqrt{N} \langle \tau(\hat{M} - M)\omega\hat{g}/g \rangle = \sqrt{N} \langle \tau(\hat{f} - \hat{g}M)\tau\omega/g \rangle.$$

As a U-statistic, this last term can be analyzed by conventional projection arguments. Provided that its expectation tends to zero, this term vanishes for the centered test statistic. For the uncentered test statistic, it contributes precisely the term that makes it asymptotically close to the centered form. As discussed in section 2.2, the above expression will have expectation tending to zero if appropriate higher kernels are employed or when the trimming function, τ , depends only on the index. For both the centered test statistic, regular kernels can be employed as this last condition holds.

2.4 Monte Carlo Designs and Results

In evaluating the test statistics T and T^* , we will examine both bias-corrected and regular forms of the test statistics as defined in Section 2.3. Both test statistics depend on an estimated covariance matrix, and we provide results for two different estimates. With S

defined as in Theorem A.3, the covariance matrix is given by $\Sigma = E(SS')$, which may be estimated by a sample analogue. Alternatively, with $\varepsilon_i \equiv Y_i - E(Y_i|V_i)$, note that the k^{th} element of S has the form:

$$S_k = \sum_{i=1}^N w_{ik}^* \varepsilon_i - \langle \nabla_{\theta} w_k^* \rangle H_0^{-1} \sum_{i=1}^N \nabla_{\theta} E(Y_i|V_i) \varepsilon_i$$

Accordingly, elements of S will depend on ε_i^2 terms. Taking an iterated expectation and conditioning on X , write:

$$\Sigma = E[E(SS'|X)]$$

Given the form of S_k , it can be shown that the inner expectation depends on the variance of Y conditioned on the index. If this conditional variance is known to be constant, as it is in all but one of the designs below, then it can be factored out of the above expectation and directly estimated as an average of squared residuals. We will use the terms KCV (known constant conditional variance) and UCV (unknown conditional variance) to refer to these two covariance matrix estimates. Test statistics will be computed and compared under these two covariance matrix estimators.

Third, we compare the performance of our adaptive weight version of the test statistic and the fixed weight version. Recall that the test statistic depends on a weight function that depends on X_k , the k^{th} exogenous variable entering the model. The adaptive or predictive weight is given by $w(X_k) = E(Y|X_k)$, which is the optimal predictor of Y under quadratic loss. Notice that this weight has an unknown functional form that is model dependent. In contrast, a fixed weight has a known functional form that does not depend on the alternative. The fixed weight we use in our Monte Carlo study is the common quadratic weight $w(X_k) = X_k^2$. The Monte Carlo experiment confirms that the adaptive weight version of the test is robust. Namely, in some designs the two versions perform similarly, however, in other designs the adaptive weight strongly dominates the fixed one.

2.4.1 Designs

All of the designs have single index structures under the null hypothesis. For each design, the alternative does not satisfy a single index assumption. Under the alternative, the first

design is a double index model with continuous dependent variables; the second design is a binary response model with index heteroscedasticity; the third design is a double index model with discrete dependent variables; the last two designs are general linear models with no index structure. In all the designs, we normalize such that $E(Y_i|X_i)$ has standard deviation 2 under the null and alternative models.

In the first (basic) design, we use the following data generating method for the null hypothesis:

$$Y_i = M_{0i} + \varepsilon_i, \quad M_{0i} \propto (X_{1i} + X_{2i})^2,$$

where the X 's $\sim \chi^2(1)$ and $\varepsilon \sim N(0, 1)$. Under the alternative:

$$Y_i = M_{1i} + \varepsilon_i, \quad M_{1i} \propto [(X_{1i} + X_{2i})^2 + (X_{1i} - X_{2i})^3].$$

The second design is a binary response design. With ε_i being i.i.d. $N(0, 1)$, under the null of a single index model:

$$Y_i = \begin{cases} 1 : & M_{0i} > \varepsilon_i \\ 0 : & \text{otherwise} \end{cases}, \quad M_{0i} \propto X_{1i} + X_{2i} - 0.5$$

Unlike the previous two designs, here the two X 's are correlated. In particular, the X 's are linear combinations of the same χ^2 variable and different normal shocks.

The alternative model introduces heteroscedasticity, with $M_{1i}\varepsilon_i$ replacing ε_i above and with $M_{1i} \propto \sqrt{1 + (X_{1i} - X_{2i})^2}$. We normalize M_{0i} and M_{0i}/M_{1i} so that they have expectation zero and standard deviation 2. This design is the only one that does not have a constant conditional variance.

Since discrete independent variables are very common in practice, the third design has a discrete regressor. The structure of the null and alternative are the same as in the basic design, however, here X_{2i} is a binary variable.

In the fourth (general linear model) design, under the null hypothesis we generate data by:

$$Y_i = M_{0i} + \varepsilon_i, \quad M_{0i} \propto (X_{1i} + X_{2i} + 2X_{3i})^2,$$

where the X_i 's $\sim \chi^2(1)$ and $\varepsilon_i \sim N(0, 1)$. Under the alternative, which has no index structure:

$$Y_i = M_{1i} + \varepsilon_i, \quad M_{1i} \propto [3X_{1i}^2 + 2X_{2i}^2 + X_{3i}^2 + 3].$$

A fifth design is constructed to compare the properties of our adaptive weight version of the test statistic and the fixed weight version. This design is different from the other four in a way that will be explained below. Under the null hypothesis we generate data by:

$$Y_i = M_{0i} + \varepsilon_i, \quad M_{0i} \propto (X_{1i} + X_{2i} + X_{3i})^2,$$

where the X_i 's $\sim N(0, 1)$ and $\varepsilon_i \sim N(0, 1)$. Under the alternative, which has no index structure:

$$Y_i = M_{1i} + \varepsilon_i, \quad M_{1i} \propto [X_{1i}^3 + X_{2i}^3 + X_{3i}^3 + 1].$$

For all the designs, the sample size we use is $n=1000$, and the number of Monte Carlo replications is 1000. We provide results for theoretical sizes of 0.05 and 0.10. As discussed earlier, the estimator we employ is developed in the previous chapter, with all window and smoothing parameters set as discussed in Chapter I. For the test statistics, with one exception given below, we set the window parameter r to be $1/5$ for the expectations $E(Y|V)$ and $M_k = E(Y|X_k)$. The window parameter for $E(M_k|V)$ is $1/7$. The index-trimming is set at .95, while the X -trimming is set at .99.

3.4.2 Monte Carlo Results

We compare all the variants of test statistics we mentioned in our Monte Carlo study, involving known or unknown conditional variance (KCV or UCV) and different bias reducing mechanisms. The bias reducing mechanisms we employ are Twicing Kernel (TW), Regular Kernel using a window $r > \frac{1}{4}$ (BRR); and Recentering. We investigate the empirical size, power, and adjusted power, which is the empirical power using bootstrap critical value adjusting the empirical size to be equal to the theoretical size. For reasons discussed below, our Monte Carlo results recommend the centered BRR as the best among all those variants. In addition, in all cases where the conditional variance is constant, it is better to impose

this information.

In comparing different variants of the test statistic, note first that the recentered test statistics perform much better than uncentered ones in that the empirical sizes are much closer to theoretical value and power is also better. The uncentered test statistics in all the designs have highly inflated empirical sizes. For example, turn to the general linear model design. The recentered KCV BRR gives empirical sizes of 0.049 and 0.097 for 5% and 10% theoretical sizes; while the uncentered version gives 0.153 and 0.231 respectively. The recentered test also has better power properties. Similar results occur for the other designs.

Second, we compare results that depend on whether or not a known constant conditional variance (KCV) assumption is correctly imposed in estimating the covariance matrix. In all three designs where this assumption holds, the performance of the test statistic is improved. The sizes are reasonable and similar, but the power of KCV is higher than UCV. For example, in the basic design, the power of the UCV version gives adjusted power of only 0.7 and 0.792 for 5% and 10% theoretical sizes; while the KCV version gives powers of 0.878 and 0.914 respectively. Not surprisingly, a better test statistic results from imposing correct (constant conditional variance) information when estimating the covariance matrix.

As for kernel selection, the results are quite close to one another. However, BRR is the most stable over designs. For the discrete regressor design, the recentered KCV with BRR gives empirical sizes of 0.053 and 0.089 for 5% and 10% theoretical sizes; while the corresponding ones for simple expectation are 0.22 and 0.353; twicing kernel yields 0.177 and 0.282. The power is also slightly better than the other two. The difference among them in other designs is often small. For example in the general linear model design our recentered KCV under BRR gives size power combinations of (0.049, 0.817) and (0.097, 0.872) for 5% and 10% theoretical sizes; while corresponding expectation by index gives (0.045, 0.85) and (0.089, 0.889); twicing provides (0.045, 0.806) and (0.088, 0.863).

As a conclusion, the recentered test statistic using BRR stands out among all the variations we tried. It performs well under all the designs. Furthermore, when it is known that the conditional variance is constant, such information should be imposed.

To compare fixed with adaptive weights, recall that the fixed weights are the squares of

the exogenous variables that appear in the model, while the adaptive or predictive weights are the optimal (MSE) predictors of Y . With the exception of the fifth design, all of the other design have important quadratic elements. As a result, the fixed and adaptive weights explain a comparable proportions of the variation in the dependent variable in those designs. Not surprisingly, in these cases we find, but do not report, that fixed and adaptive weights perform similarly. In the fifth design, which is given above, quadratic elements are not important in the alternative model. Even collectively, such elements only explain 3.5% of the variation in Y . In contrast, collectively the adaptive weights explain 78.1%. Table 2.2 provides Monte Carlo results for the comparison between our adaptive weight version of the test and the fixed weight version. It is shown that our adaptive weight test statistic dominates the fixed weight version in this design by having much better power results. For example, at the 5% theoretical critical value panel, we find the adaptive weight version of the recentered BRR test with KCV has an empirical power of 0.962; while the number for the fixed weight version is much lower at 0.706.

2.5 Conclusions

In summary, we have formulated a test statistic for testing the frequently made single index assumption in semiparametric models. We establish the large sample distribution of the test statistic under the null hypothesis and show that it performs well across a variety of designs in Monte Carlo experiments. This performance is obtained by an embedded bias control mechanism, the adaptive nature of the test statistic, and also the estimator upon which it is based.

Chapter 3

Determinants of Healthcare Decisions: Insurance, Utilization, and Expenditures

3.1 Introduction

A major healthcare policy issue in the U.S. today is the growing population without insurance. The key questions include: How does health insurance coverage affect the likelihood an individual seeks medical care? How does health insurance affect healthcare expenditures?

There are many empirical challenges in studying people's healthcare decisions. An individual's decision about whether to utilize healthcare may depend on her insurance coverage. The level of utilization likely also depends on whether or not the individual has insurance. However, because insurance is a choice variable for the individual, we must allow for the possibility that this variable is endogenous. For example, people who have greater need for healthcare have more incentive to buy health insurance. Some papers in the literature deal with this endogeneity by using instrumental variables (e.g., Vera-Hernandez, 1999; Holly et al., 2002; Wooldridge, 2002); others use experimental data to avoid this problem (e.g., Manning et al., 1987; Newhouse, 1993; Barros et al., 2008). However, instruments that are correlated with insurance coverage but not with utilization are difficult to find. Experimental data are scarce and often out of date. For example, the RAND Health Insurance Experiment, which remains to be the largest health policy study in U.S. history, started in 1971 and lasted for 15 years (RAND, 1974-1982). The structure, practice, and philosophy of medicine have changed dramatically since the 1980s as has the insurance industry.

Another empirical challenge lies in the expenditure decision, where we only observe positive expenditures from individuals who decide to see a doctor. One standard parametric approach deals with this problem by making distributional assumptions about error terms and then using a Heckman correction for sample selection (Heckman, 1976, 1979). An alternative approach is to use a two-part model (Duan et al., 1983, 1984, 1985). Both of these may be problematic as the Heckman correction approach can be sensitive to the distributional assumptions on error terms, while the two-part model approach also makes

implicit distributional assumptions (Puhani, 2000). The literature in health economics or in economics in general does not provide a theoretical foundation or justification for these distributional assumptions. Moreover, if incorrect, they can result in incorrect inferences and policy conclusions with respect to healthcare decisions.

Yet another challenge is the complicated nature of the decision-making process. In healthcare, both the patient and the doctor are involved in making decisions. The patient decides whether to visit a doctor (or more generally a healthcare provider), and then the patient and doctor jointly decide what treatment the patient will have. These decisions are interrelated. Some papers deal with the two-part decision-making process in healthcare utilization (Newhouse, 1993; Mullahy, 1998), but none address the whole process of insurance choice, utilization, and expenditure level.

This chapter contributes to the current literature by taking into account the interrelated nature of healthcare decisions and using a semiparametric approach to address the empirical challenges. We study three healthcare decisions: insurance coverage, utilization, and the level of expenditures. Using the Medical Expenditure Panel Survey (MEPS) 2005 data, we formulate and estimate a model for these three healthcare decisions. As there is not a strong justification for normality assumptions underlying a traditional parametric formulation, we employ a semiparametric approach in which these assumptions are not made. As an additional advantage to a semiparametric approach, it should be remarked that since marginal effects will in general not be constant in nonlinear models, we will report the impact of changing a variable of interest at several different points in its distribution. The semiparametric approach will also allow greater flexibility in the pattern of these effects than in the parametric case. Nevertheless, as a convenient benchmark, we also estimate the model using a standard parametric approach.

The chapter is organized as follows. Section 3.2 introduces the model and explains the parametric and semiparametric approaches in two subsections; Section 3.3 describes the dataset; Section 3.4 gives the main results; and Section 3.5 provides conclusions, discussions, and future research directions.

3.2 The Model

We study a set of three equations to examine the effects of different factors on healthcare decisions: health insurance, utilization, and expenditures. The first equation deals with the health insurance choice. Let I be an indicator of whether or not an individual selects private health insurance coverage. In the model below, an individual selects insurance if the net value to so doing, $V_I - \varepsilon_I$, is greater than zero. With V_I determined by a set of exogenous variables X_I , the model is as follows:

$$I = \begin{cases} 1 : & V_I > \varepsilon_I \text{ where } V_I = X_I \beta_I \\ 0 : & \text{otherwise} \end{cases},$$

The second equation describes the decision to seek healthcare. Let A be an indicator of whether or not an individual seeks access to healthcare from a doctor or other healthcare providers, and let X_A be a set of exogenous variables that determine the net value of utilizing healthcare. Then:

$$A = \begin{cases} 1 : & V_A + I\theta_A > \varepsilon_A \text{ where } V_A = X_A \beta_A \\ 0 : & \text{otherwise} \end{cases},$$

Notice that the insurance coverage enters this utilization (access) equation. There is a vast literature about the effects of moral hazard and adverse selection (see Arrow, 1963; Rothschild and Stiglitz, 1976; Chiappori and Salanie, 2001; Cardon and Hendel, 2001, for example). On the one hand, people who have insurance are much more likely to utilize healthcare than their uninsured counterparts. On the other hand, people who have greater demand for healthcare (e.g., those with high comorbidity levels) may have more incentive to obtain insurance coverage. Consequently in our estimations, we will use methods that deal with this endogeneity issue. To this end, exclusion restrictions will be needed. More specifically we need some variables in the insurance equation to be excluded from the utilization equation. In the data section, we will explain the variables used as exclusions.

The last equation explains the level of expenditures. Denote Y_E as the log of level of expenditures, and X_E as a set of exogenous variables that affects expenditures for those individuals who access healthcare services. Then, the model is given as:

$$Y_E = X_E\beta_E + I\theta_E + \varepsilon_E \quad : A = 1 \ .$$

An individual incurs positive expenditures only if a visit is made. The patient decides whether to visit a doctor, and then a joint decision is made by both the doctor and the patient. We address this two-part decision-making process by separating the two equations and allowing them to have different explanatory variables and parameters. Again, insurance, healthcare, and the individual's health status are interrelated. Insurance coverage is included in this model, because it may affect patient and doctor's joint decision about treatment plans. For example, insured people are much more likely to buy brand-name medications instead of their generic counterparts. There could also be an adverse selection problem here, because it is possible that people who are less healthy might have more incentive to purchase insurance. Hence our model will account for the interrelations between the above variables, and we will employ estimation methods that deal with both sample selection and endogeneity issues. Similar to the utilization equation, for identification purposes we will need some exogenous variables in the insurance and utilization models that are excluded from this expenditure equation. In the data section, we will discuss the particular variables that provide the required exclusion restrictions.

To avoid making strong distributional assumptions that are hard to justify, in this chapter, we employ a semiparametric method to estimate the three healthcare equations discussed above. Indeed, we will find that standard parametric distributional assumptions (e.g. joint normality) do not hold. Nevertheless, as a convenient benchmark, we also provide the parametric formulation and results. There are a variety of different methods for estimating the parametric model. To make the role of the parametric assumptions transparent, we estimate the parametric model in a manner that parallels the semiparametric approach. In the parametric case, we will make distributional assumptions on all three equations above. In contrast, we will relax these distributional as

3.2.1 Parametric Model

In the parametric model, we assume that the error terms in the above system of three

equations follow a trivariate normal distribution. A two-step estimation method is then employed to estimate the three equations. In the first step, the insurance and utilization decisions are jointly estimated by maximum likelihood. Assuming that the errors in the insurance and utilization equations are jointly distributed as bivariate normal, the likelihood has a bivariate probit form, which accounts for the possibility that insurance may be endogenous with respect to the access decision.

To identify the parameters without relying on nonlinearities, we require restrictions on the model. The insurance equation will depend on only exogenous variables, X_I , while the access decision will depend on exogenous variables, X_A , and whether or not the individual has insurance. In this triangular system of binary equations, the insurance equation is identified as it is essentially a reduced form. However, to identify the access equation, we impose exclusion restrictions on it. Namely, there are exogenous variables that affect the choice of insurance (i.e., variables in X_I) that are excluded from X_A . Such excluded exogenous variables do not affect the access decision once this decision already controls for insurance. Two exclusion variables are industry insurance rates and occupation. Industry and occupation have important impacts on insurance, because insurance plans are often provided by employers in the U.S. In the meantime, they are not expected to affect utilization or expenditures once the insurance decision is made. We will further discuss these exclusions in the data section.

In addition to the parameters in the joint model for the two decisions, the likelihood depends on the correlation between the errors. A non-zero correlation between the two error terms would indicate the endogeneity of insurance with respect to the utilization decision. As will be described below, in the empirical results we find this correlation to be small in absolute magnitude and not statistically different from zero.

In the second step, we estimate the expenditure equation by employing a Heckman correction (Heckman, 1976; Fische et al., 1981; Lee, 1982) that controls for both sample selection and the possibility that insurance is an endogenous variable with respect to expenditures. To simplify this correction or control term, we employ a form for it that is applicable when, as was found empirically, utilization and insurance errors are not correlated.¹ For individu-

¹As discussed in the next section, in a semiparametric formulation we will not need to make any assump-

als that utilize healthcare, recall the form of the expenditure model in the previous section. With ε_E as the error term in the log expenditure model, and denoting Z as the set of all the exogenous variables in the system of three equations, for $d \in \{1, 0\}$, define:

$$\lambda_d G_d(V_A, V_I) = E(\varepsilon_E | Z, A = 1, I = d).$$

In a parametric model with jointly normal errors, the G -functions above are known and the λ 's are parameters whose values are unknown. Typically, the above expectations are not zero and depend on the variables Z , A , and I . To estimate our model, we seek to remove the dependence of the errors on these conditioning variables. To this end, for $d \in \{1, 0\}$, define the recentered errors:

$$\varepsilon_d^* = \varepsilon_E - \lambda_d G_d(V_A, V_I),$$

where by construction,

$$E(\varepsilon_d^* | Z, A = 1, I = d) = 0.$$

Then, including the G -controls in the model and partitioning X_E into X_{E1} and X_{E0} according to whether $I = 1$ or 0 :

$$Y_E = \left\{ \begin{array}{ll} X_{E1}\beta_E + \theta_E + \lambda_1 G_1 + \varepsilon_1^* & : A = 1, I = 1 \\ X_{E0}\beta_E + \lambda_0 G_0 + \varepsilon_0^* & : A = 1, I = 0 \end{array} \right\}.$$

Here, the observations are stacked with $I = 1$ observations followed by $I = 0$ observations and the G functions are evaluated at the estimated indices. Provided that the above equation is identified and joint normality holds, OLS estimation provides consistent estimates.

To identify the above equation, without relying on nonlinearities in the G -controls, we impose exclusion restrictions on the exogenous variables X_E that enter this equation. Detailed discussions about these and other restrictions will be provided in the data section.

We conclude this discussion about the parametric model by emphasizing the importance of its restrictive parametric assumptions. Both the bivariate probit specification and the form of the Heckman correction term depend on the (joint) normality assumption. If this

tions on the functional form of this correction factor.

assumption is incorrectly imposed, the resulting estimator is typically inconsistent. In the next subsection, we propose a semiparametric approach that avoids the restrictive distributional assumptions. Furthermore, the parametric model assumes a threshold crossing structure, while the semiparametric model does not impose this structure. As discussed in the next section, this generalization has important implications for marginal effects.

3.2.2 Semiparametric Model

In the semiparametric model, no assumptions are made about the error terms $\varepsilon_I, \varepsilon_A$, and ε_E . In these types of models involving limited dependent variables, a parametric formulation requires an assumption about the distribution of these errors. However, if this assumption is incorrect, then the resulting estimator will not be consistent. Therefore, as there is no compelling arguments in the literature regarding error distributions, it is important to examine methods for which such distributional assumptions are not needed. Moreover, as will be pointed out below, in a semiparametric formulation it will also be possible to generalize the manner in which exogenous values and errors interact so as to result in a binary outcome (decision).

While the semiparametric model generalizes the parametric model, it does retain a parametric (index) restriction to insure that the estimator "works well" in moderately sized samples. To illustrate this restriction, return to the insurance model. In a commonly employed probit specification:

$$P(I = 1|X) = \Phi(X_I\beta_I),$$

where the function Φ is the cumulative distribution function for the model's standard normal error component, ε_I . In a semiparametric formulation, this function need not be specified and indeed can be estimated from the data along with parameters of interest. In such a formulation, the model is semiparametric because it makes no parametric assumptions on the error distribution, but does assume a parametric index, $V_I \equiv X_I\beta_I$. This index, V_I , need not be linear, but as discussed below it is important that it has a parametric form. In

a more general, nonparametric formulation, we might write:

$$P(I = 1|X) = F(X_1, X_2, \dots, X_K) = E(I|X).$$

However, when the dimension of X is large, it is difficult to "reliably" estimate the above probability (expectation).² Index restrictions serve to keep the relevant dimension of the problem small and thereby improve the finite sample behavior of the estimator. In general, a single index restriction takes the form:

$$E(I|X) = E(I|V_I) \equiv F_1(V_I).$$

In this form, not only is the function F_1 left unspecified, but the model also permits very flexible interactions between errors and values.

In some problems, a single index may not adequately describe the underlying behavior of interest. Given that the access model is not linear, when insurance is endogenous with respect to access, the access probability depends not only on its own index but also on the exogenous index driving the insurance decision. In this case, a double index model would be appropriate. Such a model would satisfy the following double index assumption:

$$E(Y|X) = E(Y|V_I, V_A) \equiv F_2(V_I, V_A),$$

where V_I, V_A are now two indices. Again, there are methods for reliably estimating the above expectation under this double index structure. As discussed below, estimators for both single and double index models will be employed here. Throughout, we use the notation $\hat{E}(Y|V)$ to denote an estimated conditional expectation for Y conditioned on V , where V may be a single index or a vector containing two indices. When this estimated expectation is evaluated at an estimate of V , as will be the case below, we will write $\hat{E}(Y|\hat{V})$.³

For the insurance and utilization decisions, we estimate the model by a method that is

²If X is continuous, then the convergence rate of the estimated expectation to the truth becomes slower as the dimension of X increases. If X is discrete, there may be few observations to estimate $E(Y|X)$ at each value of X .

³Here, we note that we employ an estimator (see Klein and Shen, 2008) that has desirable finite sample properties as well as traditional large sample properties of consistency and normality.

analogous to that for the parametric case. For that case, the form of the likelihood is known and the model is estimated by maximum likelihood. In contrast, here we do not make any distributional assumptions on error components, implying that the form of the likelihood is unknown. Nevertheless, it is possible to employ index assumptions above to develop an estimator for the likelihood.

We employ an estimator based on an extension of the approach in Klein and Shen (2008), where a bias correction mechanism was proposed to overcome finite sample performance issues of common semiparametric estimators in the literature. Monte Carlo studies in that paper show that this estimator dominates the others in terms of mean squared error. One component of the model below contains a triangular system of binary response equations. Klein, Shen, and Vella (2009) extend the bias-control mechanism discussed above to establish desirable large-sample properties for the estimator of this component. The estimator for these components of the model is then based on maximizing an "estimated log-likelihood". To define this function, for $r, s \in \{0, 1\}$, let

$$Y_{rs} = \begin{cases} 1 & : A = r, I = s \\ 0 & : \textit{Otherwise} \end{cases},$$

with the corresponding probabilities:

$$P_{rs} = \Pr(Y_{rs} = 1|V_A, V_I).$$

For $r = 1$ and $s = 1$ (other cases are analogous), notice that

$$\begin{aligned} P_{11} &\equiv \Pr(Y_{11} = 1|V_A, V_I) = \Pr(A = 1, I = 1|V_A, V_I) \\ &= \Pr(A = 1|I = 1, V_A, V_I) \Pr(I = 1|V_A, V_I) \\ &= \Pr(A = 1|I = 1, V_A) \Pr(I = 1|V_I) \\ &= E(A|I = 1, V_A) E(I|V_I). \end{aligned}$$

Hence, P_{11} can be estimated by estimating each of the above two expectations semiparametrically. The first expectation over A has a double index form, while the second one has

a single index form. The product of the above expectations (probabilities) then provides the joint probability of interest. It is important to impose a single index restriction on the I -model as it provides useful identifying information. Namely, in general double index models, identification requires that each index contains a continuous variable that is excluded from the other. This restriction is not required here. We do require, however, that there be at least one continuous variable (or a variable that can be viewed as "approximately continuous") in the model that enters access and insurance decisions.⁴ In addition to these continuity restrictions, we require and impose the same exclusion restrictions discussed in the previous section for the parametric model.

Given the estimated probabilities, we can now proceed as in Klein and Spady (1993) to estimate the model by maximizing the following estimated log likelihood:

$$\text{Log}\hat{L} = \sum_{r,s} Y_{rs} \text{Ln}(\hat{P}_{rs}).$$

When we assume that the above probabilities are known and have a bivariate normal structure, the estimator becomes bivariate probit. By estimating the probabilities using index assumptions as discussed above, we avoid assuming parametric functional forms.⁵

Turning to the expenditure equation, we again need a correction term that will enable us to deal with the sample selection and endogeneity problems. As above, with Z containing all of the exogenous variables in all three equations, and V_o referring to (V_A, V_I) , consider the control function:

$$\begin{aligned} G_d(V_o) &\equiv E(\varepsilon_E | A = 1, I = d, Z) \\ &= E(\varepsilon_E | A = 1, I = d, V_o) \end{aligned}$$

where $d \in \{1, 0\}$. Notice that this adjustment is similar to that in the parametric case, but

⁴In the insurance model, we treat the following variables: age, age², number of comorbidities, years of education, family size, and industry insurance rate as being approximately continuous; while in the access decision, these variables are: age, age², number of comorbidities, years of education, and family size.

⁵For technical reasons, and is standard in this literature, we trim out certain observations for which the probability is poorly estimated.

now we do not make any assumptions on its functional form here in the semiparametric formulation. With c as a constant:

$$X_E \beta_E = X_c \beta_c + c,$$

we can rewrite the expenditure equation as:

$$Y_E = \begin{cases} X_{c1} \beta_c + c + \theta_E + G_1 + u_1^* & : A = 1, I = 1 \\ X_{c0} \beta_c + c + G_0 + u_0^* & : A = 1, I = 0 \end{cases}$$

where $u_d^* = \varepsilon_E - G_d(V_o)$

$$E[u_d^* | A = 1, I = d, Z] = E[u_d^* | A = 1, I = d, V_o] = 0$$

Since the control functions are unknown, we employ an extension of Peter Robinson's differencing method (Robinson, 1988) to eliminate the unknown control functions:

$$Y_E - E(Y_E | A = 1, I = d, V_o) = \begin{cases} [X_{c1} - E(X_{c1} | A = 1, I = d, V_o)] \beta_c + u_1^* & : d = 1 \\ [X_{c0} - E(X_{c0} | A = 1, I = d, V_o)] \beta_c + u_0^* & : d = 0 \end{cases}$$

With “*” denoting a differenced variable:

$$Y^* = \begin{pmatrix} Y_E - E(Y_E | A = 1, I = 1, V_o) \\ Y_E - E(Y_E | A = 1, I = 0, V_o) \end{pmatrix},$$

$$X^* = \begin{pmatrix} X_{c1} - E(X_{c1} | A = 1, I = 1, V_o) \\ X_{c0} - E(X_{c0} | A = 1, I = 0, V_o) \end{pmatrix},$$

$$u^* = \begin{pmatrix} u_1^* \\ u_0^* \end{pmatrix},$$

we can rewrite the above equation as:

$$Y^* = X^* \beta_c + u^*.$$

Before proceeding to estimate the above differenced model, there are several identification issues that need to be discussed. First, it is clear that the constant term and the insurance variable disappear from the model. Second, as in the parametric model, we require additional identifying restrictions. To this end, we impose the same exclusion restrictions as in the parametric model discussed above. To see that these restrictions are needed, suppose that there are no variables excluded from X_c that appear in the indices V_I and V_A . Without such restrictions, it will be possible to take linear combinations of the X_c variables and reproduce one of the indices, say V_I . To illustrate the problem, for simplicity take $X_c = X_I$. (A similar argument holds when $X_c = X_A$.) Then, since V_I is a linear combination of the variables in X_I , there must exist a vector of coefficients C such that $X_c C = V_I$. It follows that:

$$\begin{aligned} X_c^* C &= [X_c - E(X_c|A = 1, V_o)]C = X_c C - E(X_c C|A = 1, V_o) \\ &= V_I - E(V_I|A = 1, V_o) = V_I - V_I = 0. \end{aligned}$$

Hence, there is perfect multicollinearity, which results in a lack of identification.

Replacing true expectations and index parameter values with their estimates, we can use OLS to estimate the expenditure equation and get consistent estimates. In this empirical study, we first use OLS to obtain consistent residuals. Second, employing squared residuals, in a semiparametric regression, we estimate the variance for the error conditioned on the X -variables through the two indices. We then employ these conditional variances in a GLS approach to obtain the final results.

Notice that in the above approach we can not directly estimate the impact of insurance coverage on expenditures (θ_E). Therefore, we next describe a strategy for indirectly obtaining this marginal effect. Recall from above that Y_{rs} is an indicator of the event $A = r$ and $I = s$, and P_{rs} is the corresponding probability. With ε as an error whose expectation is 0 when conditioning on only exogenous variables, and again with V_o referring to (V_A, V_I) ,

notice that:

$$\begin{aligned} E(\varepsilon_E|V_o) &= 0 \\ &= P_{rs}E(\varepsilon_E|Y_{rs} = 1, V_o) + (1 - P_{rs})E(\varepsilon_E|Y_{rs} = 0, V_o). \end{aligned}$$

For P_{rs} close to 1, we now have the following useful result:

$$R) : E(\varepsilon_E|Y_{rs} = 1, V_o) \doteq 0.$$

We exploit the above result to estimate θ_E , a marginal effect of interest. To this end, write the conditional expectations given insured/not:

$$\begin{aligned} M_1 &\equiv E(Y_E - X_c\beta_c|Y_{11} = 1, V_o) = c + \theta_E + E(\varepsilon_E|Y_{11} = 1, V_o) \\ M_0 &\equiv E(Y_E - X_c\beta_c|Y_{10} = 1, V_o) = c + E(\varepsilon_E|Y_{10} = 1, V_o). \end{aligned}$$

From property (R), for each individual j with P_{11} close to 1:

$$M_{1j} \doteq c + \theta_E.$$

Similarly, for each individual i with P_{10} close to 1:

$$M_{0i} \doteq c.$$

With \bar{M}_1 as an average over j and \bar{M}_0 as an average over i , we can recover θ_E by calculating:

$$\hat{\theta}_E = \bar{M}_1 - \bar{M}_0.$$

To get some sense as to how well the method described above works in practice, we conduct a small scale Monte Carlo experiment, where we find that this method performs very well. We generate data from the following design, which has the same structure as our model:

$$I = \begin{cases} 1 : & V_I > \varepsilon_I \text{ where } V_I = X_1 + X_2 + X_3 + 1 \\ 0 : & \text{otherwise} \end{cases},$$

$$A = \begin{cases} 1 : & V_A + 2I > \varepsilon_A \text{ where } V_A = X_1 + X_2 \\ 0 : & \text{otherwise} \end{cases},$$

$$Y_E = 4I + 2X_1 + 1 + \varepsilon_E \quad : A = 1 ,$$

where the X 's are all distributed as normal, and the errors are jointly normal with non-zero correlations between them. The sample size we use is $n=2000$, and the number of Monte Carlo replications is 1000. As we can see, the true $\theta_E = 4$. The average $\hat{\theta}_E$ from the Monte Carlo is 3.93, and the standard deviation is 0.16. In other words, the percentage bias is about 1.6%, and the variance is also small, taking into account that the truth is 4.

We want to point out another advantage of semiparametric estimation with respect to marginal effects here. Since marginal effects will in general not be constant in nonlinear models, more information can be revealed by examining the patterns of marginal effects at several different points in the distribution of a variable of interest. We will look at the patterns by changing this variable at several different points in its distribution. For instance, we report the marginal effects of education at different education level groups (less than high school, high school, and some college or higher). Because the semiparametric model is more flexible than the parametric model, the semiparametric case permits a richer pattern of marginal effects. As an example of such flexibility, in specifying the model for the insurance decision, a single index restriction would permit a model of the form:

$$I = 1 \text{ iff } f(V_I, \varepsilon_I) > 0,$$

where V_I is the index ($X_I\beta_I$ in the linear index case) and ε_I is the error. Here, f is an unknown function that may or may not be separable in the index and the error. Furthermore, the function f may or may not be monotonic in its arguments. Finally, the distribution of the error component is left unspecified. We will demonstrate how this flexibility yields

valuable information and policy implications in our result section.

3.3 Data

The Medical Expenditure Panel Survey (MEPS) is an on-going nationally representative survey of U.S. civilian non-institutionalized population started in 1996 by U.S. Department of Health and Human Services. Surveys of households, employers, and medical providers are conducted to collect information on healthcare expenditures and health insurance coverage as well as demographic and socioeconomic characteristics.⁶

We consider the subsample of obese adults between the ages of 22 and 64, who are employed. People who have body mass index (BMI) greater than 30 are considered obese (CDC, 1985-2007). We focus on the obese population, because this is a growing population that might have different healthcare needs and patterns than other groups. We also focus on individuals who are employed, because in the United States, insurance is often linked with employment. In fact, health insurance plans are often offered by employers. We exclude individuals who have public insurance, because having public insurance is not expected to be a consumer's choice for working adults between the ages of 22 and 64. The final sample consists of 2,771 individuals.⁷

The key endogenous variables that we seek to explain are insurance coverage, utilization of the healthcare system, and the level of expenditures. The insurance variable here is an indicator of whether the individual has private health insurance coverage. The expenditures are the total amount paid for healthcare services, including both out-of-pocket payments and payments by insurance; but not including payments for over-the-counter drugs. Note that the expenditures are derived from the MEPS Household and Medical Provider Components. Since both the healthcare providers and the consumers are surveyed, it is more reliable than typical surveys. We define utilization of healthcare system as having positive healthcare expenditures.⁸

⁶We note that the semiparametric model can be less sensitive to reporting errors than parametric models (see, for example, Hausman et al., 1998).

⁷Other exclusion criteria included: individuals who died during the year, missing values on the exogenous variables used. Various robustness checks indicate that there are no selection issues in this sample.

⁸We use this indicator instead of the self-reported healthcare utilization, because the self-reported utilization may suffer from recall errors, whereas the expenditure data were collected by both sides and hence more reliable.

The explanatory variables include demographics, socioeconomic status, and health related characteristics. The demographics include age, gender, race/ethnicity (white, non-white), marital status (married, other), family size, and region (northeast, midwest, south, west). Years of education, income, occupation class, and industry insurance rates are included as socioeconomic characteristics. We use an indicator for white-collar jobs (professional, management, business and financial operations) to reflect the impact of occupation, and the percentage of people having insurance in each industry in the Kaiser study as a variable to reflect the impact of industry (Kaiser Family Foundation, 2006). The health related characteristics include number of comorbidities, presence of mental illnesses, and whether they are current smokers. Each individual was asked whether or not they had any of a number of conditions. The comorbidity variable then counts the following health problems: Alzheimer’s disease, asthma, arthritis, cancer, emphysema, diabetes, heart disease, high blood pressure, osteoarthritis, and stroke. This variable is included to capture differences in people’s physical health status and is often employed in health studies (e.g., Klabunde, 2000). Presence of mental illnesses is an indicator of whether an individual has depression, anxiety, or schizophrenia. In the following paragraphs, we discuss the set of exogenous variables to be included in each equation and the exclusion variables.

The set of explanatory variables in the insurance decision equation includes: age, age², gender, race/ethnicity, marital status, family size, region, education, income, occupation, industry insurance rate, number of comorbidities, presence of mental illnesses, and whether they are current smokers. It is conceivable that older people have greater incentive to obtain insurance coverage because they often have more health issues and concerns. Females may have different health needs than males. It would be interesting to know whether race has an impact as it may shed some light upon racial disparities in health. Married people might have more incentive to obtain insurance coverage. The same reason applies to family size. Different regions might have different healthcare policies and plans as well as different availabilities of healthcare services.⁹ Consequently region may also have an impact on insurance. The education level is often used as a proxy for health literacy (e.g., Lindau et al., 2006; Steinvil et al., 2008). We expect people with more education to be more health

⁹No detailed information about state of residence is available in the MEPS dataset.

conscious and willing to spend more time and money on health. Hence those people are more likely to have insurance. As is known in the literature, occupation and industry have important effects on people's insurance (Kaiser Family Foundation, 2006). In the United States, insurance plans for working adults often come as a part of the compensation package. When good insurance packages are offered, people are more likely to have insurance. We expect physical and mental illnesses to increase people's incentive to buy insurance. The impact of smoking is unclear. On the one hand, current smoking might result in greater healthcare demands because it has a negative impact on health; on the other hand, it might indicate a lower health consciousness.

The exogenous variables in the utilization equation are: age, age², gender, race/ethnicity, marital status, family size, region, education, income, number of comorbidities, presence of mental illnesses, and an indicator of current smoking. Older people are expected to have higher probabilities of utilizing healthcare. Again there could be gender differences in utilization. For example, females might utilize healthcare regularly because of their annual papsmear checkup. It would be revealing to know whether there are racial differences in utilization after controlling for insurance. The impact of marriage and family size on utilization is mixed. These variables might have a positive impact (e.g., spousal pressures); while there could also be a negative impact because family duties take up time and increase the opportunity cost of visiting a doctor. Different regions may have varied healthcare availabilities, and hence may affect utilization. For the same reason as in the insurance equation, more education can have a positive effect on utilization by improving health literacy. Physical and mental illnesses increase the likelihood of utilizing healthcare, because they increase the need for health services. The impact of being a current smoker is still unclear.

The expenditure equation includes the following variables as exogenous explanatory variables: age, age², gender, race/ethnicity, family size, education, income, number of comorbidities, presence of mental illnesses, and an indicator of current smoking. Older people may incur more healthcare expenditures. Males and females may have different healthcare needs. Racial differences in healthcare expenditures would be interesting to know. Physical and mental illnesses are expected to increase the level of expenditures, and the magnitude

of these effects would be interesting to know. The impact of smoking is again unclear. One would expect a long term negative effect on health from smoking and increased healthcare expenditures in the long run. However, there is no information about smoking history, and even if available, it might not be accurate because of the nature of a survey. Moreover, current smoking might indicate a lower health consciousness.

Recalling the exclusion restrictions discussed in the previous section, we want to emphasize the exclusion variables we use here. It should be noted that exclusion restrictions for this particular type of model are difficult to find. For example, we have argued that the access decision is separate and distinct from the expenditure decision, which makes exclusion restrictions necessary to identify them. However, one would think that many of the variables that affect healthcare utilization would also affect healthcare expenditures. In this chapter, we use the following exclusion restrictions. The industry insurance rate and occupation are excluded from both utilization and expenditure equations; while marital status and region are excluded from the expenditure equation. In the United States, health insurance is often included in the compensation package offered by an employer. Different jobs might offer varied choices of insurance packages at different prices. Hence it affects the insurance decision by affecting the cost of buying insurance. However, once the insurance coverage decision is made, it is plausible to assume that the industry insurance rate and occupation class would not affect the benefit or the cost of utilization and expenditures after controlling for income and education. Recall that the patient makes decisions about insurance and utilization, while the doctor and patient jointly decide on the level of treatment, with the doctor being the main decision maker. Once a patient decides to visit a healthcare provider, we assume that the prescribed treatment does not depend on marriage or region. Hence the level of expenditures may not depend on these variables. We recognize the difficulty in finding appropriate restrictions for the type of model that we estimate, but view the exclusion restrictions discussed above as being plausible.

Some summary statistics with preliminary bivariate analysis of the data are provided in Tables 3.1-3.3. Note that the continuous variables are categorized into groups to show the distribution of those variables. However, they remain continuous in estimating the model. Table 3.1 provides detailed summary statistics for the variables discussed above.

Of the 2,771 individuals in our dataset, 488 (18%) are uninsured and 262 (10%) have no utilization. The level of expenditures for those that utilize healthcare is very skewed. About 40% of them have expenditures of less than \$1,000, while 8% of them incur more than \$10,000 in healthcare expenditures. Less than half of the study population (49%) have none of the physical illnesses mentioned above and 19% have two or more comorbidities. In this sample, 62% are married. Table 3.2 describes the population by insurance coverage (insured/uninsured), and also provides preliminary chi-square test results. It shows that 14% of the people who utilize healthcare are uninsured, while 48% of those who have no utilization are uninsured. Age and comorbidities increase the likelihood of having insurance probably by increasing the incentive to get insurance. White people and married people are more likely to be insured. It also shows that socioeconomic characteristics play an important role in people's insurance coverage. More income and more education both increase the probability of having insurance. Industry and occupation are significant factors that affect insurance coverage. People in highly insured industries are more likely to be insured. Current smokers are less likely to have insurance coverage. There is not much difference between males and females in insurance coverage. Table 3.3 provides a similar cross tabulation by utilization: 94% of those insured utilize healthcare, while only 74% of those uninsured utilize healthcare. There is significantly less utilization among those uninsured. With the exception of gender, most of the characteristics affect utilization and insurance decisions similarly. The insurance decisions are not very different for males and females, while females are much more likely to utilize healthcare.

3.4 Results

In this section, we discuss both parametric and semiparametric results of estimating the three equations. Before we discuss these results, there is an important normalization issue for the semiparametric case that affects how we present results for insurance and utilization. For simplicity, we illustrate the issue for the insurance decision. As discussed earlier, the estimates are based on an estimate of the probability:

$$Pr(I = 1|X_I\beta_I) = Pr(I = 1|a + b(X_I\beta_I)) , \text{ where } a \text{ and } b \text{ are constants.}$$

The above probability does not depend on a or b . Hence one of the variables will have its parameter normalized to a fixed value. After estimation, we normalize the parameter of education to the corresponding parametric estimate for presentation purposes.¹⁰ Nevertheless, the marginal effects of all variables are retrievable and invariant to normalization. The same logic applies to the utilization (access) equation.

Below we examine both parametric and semiparametric results for the three decisions. We compare not only the normalized estimates and average marginal effects but also patterns of marginal effects calculated at different levels of certain continuous variables of interest. Most of the normalized estimates and average marginal effects are close between the two approaches for insurance and utilization decisions. However, the two estimation methods yield very different estimated effects of insurance on expenditures. Furthermore, the semiparametric approach gives richer patterns of marginal effects. Detailed results are provided in Tables 3.4-3.7.

First, we look at the insurance coverage decision. As shown in Table 3.4, both the normalized estimates and the average marginal effects are similar for parametric and semiparametric approaches. The biggest marginal effect on the probability of having insurance comes from marital status, with the p-values of the coefficient on married in both approaches being less than 0.01. Marriage increases the probability of having private insurance coverage by more than 7% points. Region also has a significant effect on the insurance coverage, with the northeast indicator having coefficient p-values of 0.06 and 0.02 in parametric and semiparametric models respectively. People in the northeast region are 4-5% points more likely to have insurance compared to people living in the west. White people are 4% points more likely to have insurance than non-whites (coefficient p-value < 0.01). Education and income level both have significant positive impacts on insurance coverage. Industry insurance rate, which is one of the exclusions, has a substantial impact on the insurance decision. Increasing the industry insurance rate by 5% increases the probability of having insurance by more than 2% points on average. Occupation class is marginally significant. The number of comorbidities also has a significant positive effect on insurance. When the number of

¹⁰The choice of variable on which to normalize does not affect estimation results (provided that the variable belongs in the model).

comorbidities increases by one, the average increase in the probability of having insurance is 2% points.

With respect to the normalized parameter estimates and averaged marginal effects, the parametric and semiparametric results for the utilization decision are also similar. These results are presented in Table 3.5. One of the most important questions here is how insurance coverage affects utilization, and parametric and semiparametric estimations provide very similar results. The average marginal effect of insurance coverage is 14-15% points (the probabilities move from 78-80% to 93-94%,) meaning if we move everyone in the sample from uninsured to insured, the average gain in the probability of visiting a doctor is 14-15% points, a large number. Both the number of comorbidities and the presence of mental illnesses have very significant positive impacts on utilization (coefficient p-value < 0.05). For the number of comorbidities, parametric estimation gives a higher marginal effect of 6% points compared to the 3% points of the semiparametric approach; while for the presence of mental illnesses, both approaches give an average marginal effect of 4% points. One interesting finding here is that females are much more likely to visit a doctor. Parametric and semiparametric estimations yield average marginal effects of 6% points and 4% points respectively. Another interesting finding is that income does not have a significant impact on utilization. Once the insurance coverage decision is fixed, income does not matter. Marital status, which is one of the exclusions, has a highly significant impact on utilization. Married people are 2-3% points more likely to utilize healthcare. Region, as an additional exclusion, is marginally significant. Another important finding here is that the correlation factor in the parametric estimation is very small in absolute magnitude (-0.09), and it is statistically insignificant with a p-value of 0.61. As discussed earlier, this finding has implications for the form of an adjustment factor in estimating the expenditure equation.¹¹

The final equation deals with the level of healthcare expenditures. Note that most of the marginal effects are the same as the coefficient estimates here. With the exception of the impact of insurance, estimates in the two approaches are similar. The number of comorbidities and the presence of mental illnesses both have very significant effects in this

¹¹The variables excluded from the expenditure equation are: marital status and region. Marital status is highly significant, and region is marginally significant.

equation. Both have p-values of less than 0.01. Having one more physical disease can increase the level of expenditures by about 35% on average; while having a mental illness increases it by more than 45%. Income again does not have a big impact on the level of healthcare expenditures.

For insurance coverage, which is the factor of most interest in this study, the semiparametric approach estimates the marginal impact to be 51%.¹² This impact would seem to be credible as it is very close to the number in a previous study by Newhouse and the Insurance Experiment Group (Newhouse et al., 1993). Their study based on the RAND Health Insurance Experiment shows that mean predicted expenditure in the 0% coinsurance (free-care) plan is 46 percent higher than in the 95% coinsurance plan. We want to keep in mind that the relevance of the study may be lowered by the fact that it was done more than a decade ago. Nevertheless, this study based on experimental data confirms our semiparametric result. In contrast, parametric estimation gives a marginal effect of 125%. We note that there are many other parametric studies that have also found an insurance impact of this magnitude (e.g., Hadley and Holahan, 2003; Miller et. al., 2003). These studies treat insurance as exogenous and state that in so doing the marginal impact of insurance has an upward bias. However, none of these studies have quantified the extent of this bias.

To understand the large difference between semiparametric and parametric results, we performed several different checks. First, we examined the normality assumption in the insurance equation by using semiparametric methods to estimate the density of the error. In particular, we obtained the semiparametric estimate of the expectation of the insurance dummy conditioned on the index. In a traditional threshold-crossing model, this estimated expectation is the estimate of the distribution function for the error term. Taking a numerical derivative then produces its density. As shown in Figure 3.1, the density estimator, which we re-centered to have median zero, is remarkably non-normal for the insurance error. It should be noted that other components of the model (access and expenditures) depend on the insurance decision. Therefore, misspecification errors in the insurance equation will be transmitted to these other components of the model.

¹²The 90 percent confidence interval for the marginal effect is approximately [.29, .72], which is based on the asymptotic distribution of the estimator as given in Klein, Shen, and Vella (2009).

To evaluate the implications of parametric distributional assumptions not holding, we performed the following experiment. Recall that in the parametric model, the G-functions that control for selection and endogeneity are known under normality. The parametric results were then obtained in an OLS estimation of the expenditure equation with these G-functions included. Given the failure of parametric distributional assumptions to hold, it would seem that these parametric G-functions are incorrect. Accordingly, we semiparametrically estimated these functions without making any assumptions on their functional forms. Recall that the semiparametric estimates of the expenditure equation were obtained by differencing out the G-functions as their form was not known. However, once all of the parameters of the expenditure model have been estimated, it is possible to obtain the semiparametric estimates of these functions. With the subscript s indicating a semiparametric estimator:

$$\begin{aligned}\hat{G}_{1s} &= \hat{E} \left[Y_E - X_c \hat{\beta}_s - \hat{c}_s - \hat{\theta}_s I \mid A = 1, I = 1, \hat{V}_I, \hat{V}_A \right] \\ \hat{G}_{0s} &= \hat{E} \left[Y_E - X_c \hat{\beta}_s - \hat{c}_s - \hat{\theta}_s I \mid A = 1, I = 0, \hat{V}_I, \hat{V}_A \right]\end{aligned}$$

Replacing the parametric G-functions with the flexibly estimated semiparametric functions above, we then re-estimated the parametric expenditure equation. The marginal impact of insurance was found to be .50, which confirms the finding that the parametric marginal effect has an upward bias by a factor of two.

Besides the difference in marginal effect of insurance coverage on the level of expenditures, Table 3.7 shows that parametric and semiparametric approaches also give very different marginal effects for different population groups. Here, the parametric approach restricts the marginal effects of the groups to be monotonic, while the semiparametric approach does not have this restriction and hence can provide more accurate results. In the parametric estimation, the marginal effects of education on insurance for the three groups (less than high school, high school, and some college or more) are 3.22% points, 2.00% points, and 1.05% points, which shows a strong monotonic relation; in the semiparametric estimation, without this restriction, the largest marginal effect is also in the "less than high school" population, and it is 1.37% points. However, the marginal effects in the other two

groups are at a similar level of 1.0-1.1% points. The same pattern happens for the marginal effects of education on utilization. The parametric estimation yields marginal effects of 1.21% points, 0.58% points, and 0.38% points respectively for the three groups, while semiparametric estimation suggests again that the marginal effects are close in the three groups (3.25% points, 2.81% points, and 2.96% points). This result suggests that it is important to improve health literacy in all groups, with probably more of the effort placed on people having less than high school education. Another interesting observation concerns the industry insurance rate. In the semiparametric case, the biggest marginal effect (1.71% points compared to 1.55% points and 1.37% points) is in the middle group where the industry insurance rate is 75-90%. The people in those industries have the greatest marginal benefit of getting into a more insured industry. In contrast, in the parametric case, marginal effects are again monotonic.

3.5 Conclusions

This chapter studies the determinants of three healthcare decisions: insurance, utilization, and expenditures. We study the above interrelated healthcare decisions by analyzing a system of three simultaneous equations. Both parametric and semiparametric methods are employed to estimate the model. The merit of our semiparametric approach compared to a parametric approach is that it avoids distributional and functional form assumptions, which are not well justified. Indeed, while there are many similarities, parametric and semiparametric approaches yield some very different results, which can lead to different policy implications.

Without repeating all the empirical results, we want to summarize some important findings and their policy implications. We find that insurance has a substantial effect on both utilization and expenditures. Both methods suggest that having private insurance coverage increases the likelihood of seeking healthcare by about 15% points. However, the estimated magnitude of the effect on expenditure diverge. The parametric estimation predicts the level of expenditures to increase by 125% if universal insurance is given; while semiparametric estimation predicts an increase of 51%, a number close to that found in a Rand experimental study (Newhouse et al., 1993). Because the parametric assumptions are

incorrect, the parametrically estimated impact of insurance on expenditures has an upward bias on the order of 100%. The policy relevance of this finding is that the cost of extending universal healthcare is much lower than predicted by traditional parametric methods.

Other marginal effects are also worth noting. Education is an important factor in every healthcare decision, and hence improving health literacy is an important issue in the obese population. Given the pattern in the marginal effects, parametric results suggest that most, if not all, of the emphasis be placed on improving health literacy of the low education group (below high school). In contrast, results from the semiparametric case suggest that it is important to improve health literacy among all education groups (with the low group somewhat favored). Finally both physical and mental illnesses increase expenditures dramatically. Physical illnesses increase the level of expenditures by about 35%, and mental illnesses increase it even more (45%+). This suggests that the obese population with physical and mental illnesses is a very challenging population. More prevention and treatment of physical and mental illnesses should be provided to this population.

There are some limitations and consequently some future research directions that we want to point out. First, this study is based on the obese ($BMI > 30$) population. It would be interesting to investigate the magnitude of marginal effects for different BMI categories. Second, we do not have information to distinguish the type of healthcare encounters, for example, whether it is a preventive checkup with a physician or an acute episode of some disease. It would be useful to distinguish different types of healthcare use, so that we can study the effects on different types of healthcare. Third, since this is a cross-sectional dataset, we do not know the temporal effects. It would be interesting to know, for example, how the use of preventive care in the previous periods affect inpatient care use in subsequent time periods.

Table 1.1 Estimation Results

Basic Design			
	S1SLS	S2SLS	CS2SLS
Bias	0.000	0.000	0.000
Rvar	0.042	0.031	0.031
Rmse	0.042	0.031	0.031

Binary Response Design			
	S1SLS	S2SLS	CS2SLS
Bias	-0.010	-0.002	0.000
Rvar	0.063	0.060	0.059
Rmse	0.064	0.060	0.059

Discrete Regressor Design					
	SLS-TW	CSLS-TW	S1SLS	S2SLS	CS2SLS
Bias	-0.019	-0.017	-0.023	-0.036	-0.019
Rvar	0.045	0.046	0.041	0.036	0.037
Rmse	0.049	0.049	0.047	0.050	0.042

General Linear Model Design			
	S1SLS	S2SLS	CS2SLS
Bias	-0.005	0.000	0.001
	0.043	-0.003	-0.004
Rvar	0.089	0.069	0.069
	0.132	0.109	0.109
Rmse	0.089	0.069	0.069
	0.139	0.109	0.109

Table 2.1 Test Results Continued

Discrete Regressor Design									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
R	size			0.206	0.220			0.338	0.353
	power			0.973	0.989			0.989	0.992
	adjusted power			0.911	0.967			0.943	0.978
TW	size	0.772	0.770	0.167	0.177	0.809	0.813	0.273	0.282
	power	0.996	0.996	0.99	0.994	0.996	0.998	0.994	0.995
	adjusted power	0.007	0.069	0.96	0.986	0.027	0.218	0.977	0.992
BRR	size	0.864	0.874	0.051	0.053	0.909	0.901	0.096	0.089
	power	0.996	0.999	0.996	0.997	1.000	1.000	0.996	0.997
	adjusted power	0.014	0.144	0.996	0.997	0.048	0.322	0.996	0.998

General Linear Model Design									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
R	size			0.034	0.045			0.082	0.089
	power			0.572	0.850			0.684	0.889
	adjusted power			0.624	0.856			0.717	0.892
TW	size	0.136	0.132	0.032	0.045	0.207	0.203	0.087	0.088
	power	0.565	0.822	0.511	0.806	0.675	0.861	0.629	0.863
	adjusted power	0.301	0.697	0.572	0.817	0.492	0.789	0.655	0.868
BRR	size	0.152	0.153	0.038	0.049	0.227	0.231	0.090	0.097
	power	0.695	0.909	0.522	0.817	0.800	0.931	0.647	0.872
	adjusted power	0.417	0.806	0.585	0.819	0.606	0.874	0.665	0.875

Table 2.2 Comparison of Fixed Weight and Adaptive Tests

Quadratic Weight Test									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
R	size			0.036	0.043			0.076	0.091
	power			0.045	0.698			0.094	0.774
	adjusted power			0.062	0.713			0.119	0.783
TW	size	0.139	0.144	0.034	0.040	0.219	0.226	0.075	0.088
	power	0.055	0.716	0.043	0.695	0.109	0.780	0.094	0.774
	adjusted power	0.012	0.559	0.061	0.708	0.020	0.662	0.120	0.783
BRR	size	0.049	0.057	0.038	0.044	0.099	0.106	0.073	0.086
	power	0.044	0.704	0.044	0.697	0.094	0.774	0.095	0.771
	adjusted power	0.045	0.681	0.061	0.706	0.099	0.768	0.126	0.786

Adaptive Test									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
R	size			0.043	0.048			0.089	0.095
	power			0.768	0.968			0.845	0.976
	adjusted power			0.786	0.968			0.852	0.976
TW	size	0.466	0.460	0.045	0.049	0.565	0.565	0.093	0.095
	power	0.813	0.975	0.757	0.966	0.877	0.981	0.833	0.976
	adjusted power	0.223	0.822	0.770	0.966	0.349	0.881	0.846	0.976
BRR	size	0.094	0.090	0.045	0.049	0.164	0.159	0.093	0.099
	power	0.796	0.971	0.757	0.962	0.860	0.978	0.830	0.975
	adjusted power	0.727	0.959	0.763	0.962	0.806	0.973	0.842	0.975

Table 3.1
Description of Study Population

	N	%
All	2771	100.0
Insurance Coverage		
Insured	2283	82.4
Uninsured	488	17.6
Utilization		
yes	2509	90.5
no	262	9.5
Expenditures		
no expenditures	262	9.5
<1,000	990	35.7
1,000-2,000	443	16.0
2,000-5,000	607	21.9
5,000-10,000	265	9.6
10,000+	204	7.4
Education		
Less than high school	471	17.0
High school	946	34.1
College or higher	1354	48.9
Age		
<40	1022	36.9
40-49	856	30.9
50+	893	32.2
Income		
<20,000	781	28.2
20,000-30,000	569	20.5
30,000-50,000	794	28.7
50,000+	627	22.6
Gender		
Female	1460	52.7
Male	1311	47.3
Race		
White	1551	56.0
Non-white	1220	44.0
Number of Comorbidities		
Zero	1352	48.8
One	881	31.8
Two plus	538	19.4
Mental Illnesses		
yes	540	19.5
no	2231	80.5
Current Smoker		
yes	542	19.6
no	2229	80.4
Marital Status		

Table 3.1
Description of Study Population

		N	%
Family Size	Married	1714	61.9
	Other	1057	38.1
Family Size	One-Two	1219	44.0
	Three-Four	1069	38.6
	Five plus	483	17.4
Region	Northeast	381	13.7
	Midwest	611	22.0
	South	1206	43.5
	West	573	20.7
Industry Insurance Rate	<75% insured	519	18.7
	75-90% insured	1326	47.9
	90%+ insured	926	33.4
Occupation	White-collar	830	30.0
	Other	1941	70.0

Table 3.2
Description of Study Population by Insurance Coverage

	Insured		Uninsured		p-value
	N	%	N	%	
All	2283	82.4	488	17.6	
Utilization					<.01
yes	2147	85.6	362	14.4	
no	136	51.9	126	48.1	
Expenditures					<.01
no expenditures	136	51.9	126	48.1	
<1,000	765	77.3	225	22.7	
1,000-2,000	388	87.6	55	12.4	
2,000-5,000	549	90.4	58	9.6	
5,000-10,000	248	93.6	17	6.4	
10,000+	197	96.6	7	3.4	
Education					<.01
Less than high school	283	60.1	188	39.9	
High school	778	82.2	168	17.8	
College or higher	1222	90.3	132	9.7	
Age					<.01
<40	788	77.1	234	22.9	
40-49	727	84.9	129	15.1	
50+	768	86.0	125	14.0	
Income					<.01
<20,000	470	60.2	311	39.8	
20,000-30,000	461	81.0	108	19.0	
30,000-50,000	745	93.8	49	6.2	
50,000+	607	96.8	20	3.2	
Gender					0.77
Female	1200	82.2	260	17.8	
Male	1083	82.6	228	17.4	
Race					<.01
White	1379	88.9	172	11.1	
Non-white	904	74.1	316	25.9	
Number of Comorbidities					<.01
Zero	1052	77.8	300	22.2	
One	761	86.4	120	13.6	
Two plus	470	87.4	68	12.6	
Mental Illnesses					<.01
yes	471	87.2	69	12.8	
no	1812	81.2	419	18.8	
Current Smoker					<.01
yes	423	78.0	119	22.0	
no	1860	83.4	369	16.6	
Marital Status					<.01
Married	1470	85.8	244	14.2	
Other	813	76.9	244	23.1	

Table 3.2
Description of Study Population by Insurance Coverage

	Insured		Uninsured		p-value
	N	%	N	%	
Family Size					<.01
One-Two	1025	84.1	194	15.9	
Three-Four	888	83.1	181	16.9	
Five plus	370	76.6	113	23.4	
Region					<.01
Northeast	346	90.8	35	9.2	
Midwest	531	86.9	80	13.1	
South	943	78.2	263	21.8	
West	463	80.8	110	19.2	
Industry Insurance Rate					<.01
<75% insured	343	66.1	176	33.9	
75-90% insured	1121	84.5	205	15.5	
90%+ insured	819	88.4	107	11.6	
Occupation					<.01
White-collar	754	90.8	76	9.2	
Other	1529	78.8	412	21.2	

Table 3.3
Description of Study Population by Utilization

	Utilization		No Utilization		p-value
	N	%	N	%	
All	2509	90.5	262	9.5	
Insurance Coverage					<.01
Insured	2147	94.0	136	6.0	
Uninsured	362	74.2	126	25.8	
Education					<.01
Less than high school	380	80.7	91	19.3	
High school	849	89.7	97	10.3	
College or higher	1280	94.5	74	5.5	
Age					<.01
<40	872	85.3	150	14.7	
40-49	785	91.7	71	8.3	
50+	852	95.4	41	4.6	
Income					<.01
<20,000	668	85.5	113	14.5	
20,000-30,000	495	87.0	74	13.0	
30,000-50,000	749	94.3	45	5.7	
50,000+	597	95.2	30	4.8	
Gender					<.01
Female	1369	93.8	91	6.2	
Male	1140	87.0	171	13.0	
Race					<.01
White	1449	93.4	102	6.6	
Non-white	1060	86.9	160	13.1	
Number of Comorbidities					<.01
Zero	1138	84.2	214	15.8	
One	834	94.7	47	5.3	
Two plus	537	99.8	1	0.2	
Mental Illnesses					<.01
yes	518	95.9	22	4.1	
no	1991	89.2	240	10.8	
Current Smoker					0.02
yes	477	88.0	65	12.0	
no	2032	91.2	197	8.8	
Marital Status					0.01
Married	1571	91.7	143	8.3	
Other	938	88.7	119	11.3	
Family Size					<.01
One-Two	1129	92.6	90	7.4	
Three-Four	972	90.9	97	9.1	
Five plus	408	84.5	75	15.5	
Region					0.07
Northeast	354	92.9	27	7.1	

Table 3.3
Description of Study Population by Utilization

	Utilization		No Utilization		p-value
	N	%	N	%	
Midwest	562	92.0	49	8.0	
South	1086	90.0	120	10.0	
West	507	88.5	66	11.5	

Table 3.4
Parametric and Semiparametric Estimation Results -- Insurance Coverage

	<i>Parametric Estimation</i>			ME (% pts.)
	Estimate	(SE)	p-value	
Intercept	-6.47	(0.57)	<.01	
Age	0.04	(0.02)	0.09	0.10
Age ²	-3.86E-04	(2.62E-04)	0.14	
Number of Comorbidities	0.10	(0.04)	0.01	1.90
Mental Illnesses	0.14	(0.09)	0.12	2.68
Female	-0.01	(0.07)	0.83	-0.30
White	0.28	(0.07)	<.01	5.74
Income	0.29	(0.03)	<.01	0.58
Current Smoker	-0.07	(0.08)	0.36	-1.45
Years of Education	0.09	(0.01)	<.01	1.85
Married	0.36	(0.07)	<.01	7.54
Family Size	0.01	(0.02)	0.72	0.16
Region-Northeast	0.23	(0.12)	0.06	4.41
Region-Midwest	0.12	(0.10)	0.23	2.44
Region-South	-0.10	(0.08)	0.23	-2.03
Industry Insurance Rate	2.61	(0.30)	<.01	2.54
White-collar	-0.04	(0.08)	0.61	-0.88

Table 3.4 Continued
Parametric and Semiparametric Estimation Results -- Insurance Coverage

	<i>Semiparametric Estimation</i>			
	Estimate	(SE)	p-value	ME (% pts.)
Intercept				
Age	-0.01	(0.02)	0.56	0.00
Age ²	1.65E-04	(2.91E-04)	0.57	
Number of Comorbidities	0.12	(0.05)	0.01	1.84
Mental Illnesses	0.07	(0.09)	0.47	1.00
Female	0.09	(0.08)	0.25	1.42
White	0.27	(0.09)	<.01	4.32
Income	0.74	(0.13)	<.01	3.31
Current Smoker	-0.07	(0.09)	0.45	-1.03
Years of Education	0.09			1.40
Married	0.45	(0.10)	<.01	7.22
Family Size	0.00	(0.02)	0.85	0.06
Region-Northeast	0.00	(0.14)	0.02	4.83
Region-Midwest	0.04	(0.10)	0.68	0.64
Region-South	-0.03	(0.09)	0.75	-0.44
Industry Insurance Rate	2.81	(0.56)	<.01	2.06
White-collar	-0.12	(0.08)	0.13	-1.86

Estimate=parameter estimate; SE=standard error; ME (% pts.)=average marginal effect in percentage points.

Expenditure and income are in \$1,000 and are logged.

Reference group for region = West.

Marginal effects of continuous variables are calculated by moving everyone in the sample above by 1 unit, except income and industry insurance rate which were moved by 10% and 5% respectively.

Marginal effects of discrete variables are calculated by moving everyone in the sample from zero to one.

Table 3.5
Parametric and Semiparametric Estimation Results – Utilization

	<i>Parametric Estimation</i>			
	Estimate	(SE)	p-value	ME (% pts.)
Intercept	0.03	(0.67)	0.97	
Age	-0.04	(0.03)	0.20	0.10
Age ²	5.64E-04	(3.56E-04)	0.11	
Number of Comorbidities	0.56	(0.07)	<.01	5.55
Mental Illnesses	0.31	(0.12)	0.01	3.69
Female	0.43	(0.08)	<.01	5.77
White	0.05	(0.09)	0.57	0.66
Income	-0.01	(0.04)	0.86	-0.01
Current Smoker	-0.10	(0.09)	0.28	-1.36
Years of Education	0.05	(0.02)	<.01	0.70
Married	0.25	(0.09)	0.01	3.38
Family Size	-0.04	(0.03)	0.12	-0.55
Region-Northeast	0.02	(0.14)	0.88	0.28
Region-Midwest	0.04	(0.12)	0.76	0.47
Region-South	0.08	(0.10)	0.39	1.11
Insurance Coverage	0.88	(0.29)	<.01	15.50
Correlation Factor	-0.09	(0.17)	0.61	

Table 3.5 Continued
Parametric and Semiparametric Estimation Results – Utilization

	<i>Semiparametric Estimation</i>			
	Estimate	(SE)	p-value	ME (% pts.)
Intercept				
Age	0.01	(0.04)	0.86	0.12
Age ²	1.25E-04	(4.81E-04)	0.79	
Number of Comorbidities	0.55	(0.26)	0.03	3.39
Mental Illnesses	0.64	(0.33)	0.05	4.18
Female	0.51	(0.27)	0.05	3.52
White	-0.35	(0.22)	0.10	-2.27
Income	-0.28	(0.21)	0.19	-1.40
Current Smoker	0.13	(0.12)	0.27	0.88
Years of Education	0.05			0.36
Married	0.36	(0.18)	0.05	2.40
Family Size	-0.10	(0.07)	0.15	-0.66
Region-Northeast	0.00	(0.19)	0.41	-1.05
Region-Midwest	0.19	(0.18)	0.27	1.31
Region-South	0.21	(0.18)	0.22	1.44
Insurance Coverage				13.70
Correlation Factor				

Estimate=parameter estimate; SE=standard error; ME (% pts.)=average marginal effect in percentage points.

Expenditure and income are in \$1,000 and are logged.

Reference group for region = West.

Marginal effects of continuous variables are calculated by moving everyone in the sample above by 1 unit, except income and industry insurance rate which were moved by 10% and 5% respectively.

Marginal effects of discrete variables are calculated by moving everyone in the sample from zero to one.

Table 3.6
Parametric and Semiparametric Estimation Results -- Level of Expenditures

	<i>Parametric Estimation</i>			
	Estimate	(SE)	p-value	ME (%)
Intercept	5.00	(0.52)	<.01	
Age	-1.23E-03	(0.02)	0.95	1.40
Age ²	1.72E-04	(2.23E-04)	0.45	
Number of Comorbidities	0.40	(0.04)	<.01	40.40
Mental Illnesses	0.55	(0.07)	<.01	54.99
Female	0.15	(0.06)	0.01	15.45
White	0.25	(0.06)	<.01	25.17
Income	-0.01	(0.04)	0.78	-0.10
Current Smoker	-0.18	(0.07)	0.01	-17.73
Years of Education	0.03	(0.01)	0.01	3.28
Family Size	-0.04	(0.02)	0.05	-3.56
Insurance Coverage	1.25	(0.32)	<.01	124.85
Correction Term wrt Visit	-0.05	(0.33)	0.89	
Correction Term wrt Insurance	-0.32	(0.16)	0.05	

Table 3.6 Continued
Parametric and Semiparametric Estimation Results -- Level of Expenditures

	<i>Semiparametric Estimation</i>			
	Estimate	(SE)	p-value	ME (%)
Intercept				
Age	1.60E-04	(0.02)	0.99	1.00
Age ²	1.15E-04	(2.36E-04)	0.63	
Number of Comorbidities	0.35	(0.08)	<.01	34.96
Mental Illnesses	0.45	(0.10)	<.01	45.32
Female	0.08	(0.08)	0.33	8.06
White	0.36	(0.07)	<.01	36.37
Income	0.09	(0.05)	0.06	0.93
Current Smoker	-0.20	(0.07)	0.01	-19.75
Years of Education	0.03	(0.01)	0.02	3.23
Family Size	-0.03	(0.02)	0.13	-3.07
Insurance Coverage				50.63
Correction Term wrt Visit				
Correction Term wrt Insurance				

Estimate=parameter estimate; SE=standard error; ME (%) =average marginal effect in percentages.

Expenditure and income are in \$1,000 and are logged.

Reference group for region = West.

Marginal effects of continuous variables are calculated by moving everyone in the sample above by 1 unit, except income and industry insurance rate which were moved by 10% and 5% respectively.

Marginal effects of discrete variables are calculated by moving everyone in the sample from zero to one.

Table 3.7
Marginal Effects across the Distribution of Select Variables of Interest

		ME on Insurance (% pts.)	
		parametric	semiparametric
Education			
	Less than high school	3.22	1.37
	High school	2.00	1.00
	College or higher	1.05	1.11
Industry Insurance Rate			
	<75% insured	4.28	1.55
	75-90% insured	2.32	1.71
	90%+ insured	1.41	1.37
		ME on Utilization (% pts.)	
		parametric	semiparametric
Education			
	Less than high school	1.21	3.25
	High school	0.58	2.81
	College or higher	0.38	2.96
Industry Insurance Rate			
	<75% insured	--	--
	75-90% insured	--	--
	90%+ insured	--	--

ME on Insurance (% pts.)=median marginal effect on insurance in percentage points.

ME on Utilization (% pts.)=median marginal effect on utilization in percentage points.

Marginal effects of education are calculated by moving everyone in the sample above by one year.

Marginal effects of industry insurance rate are calculated by moving everyone in the sample above by 5%.

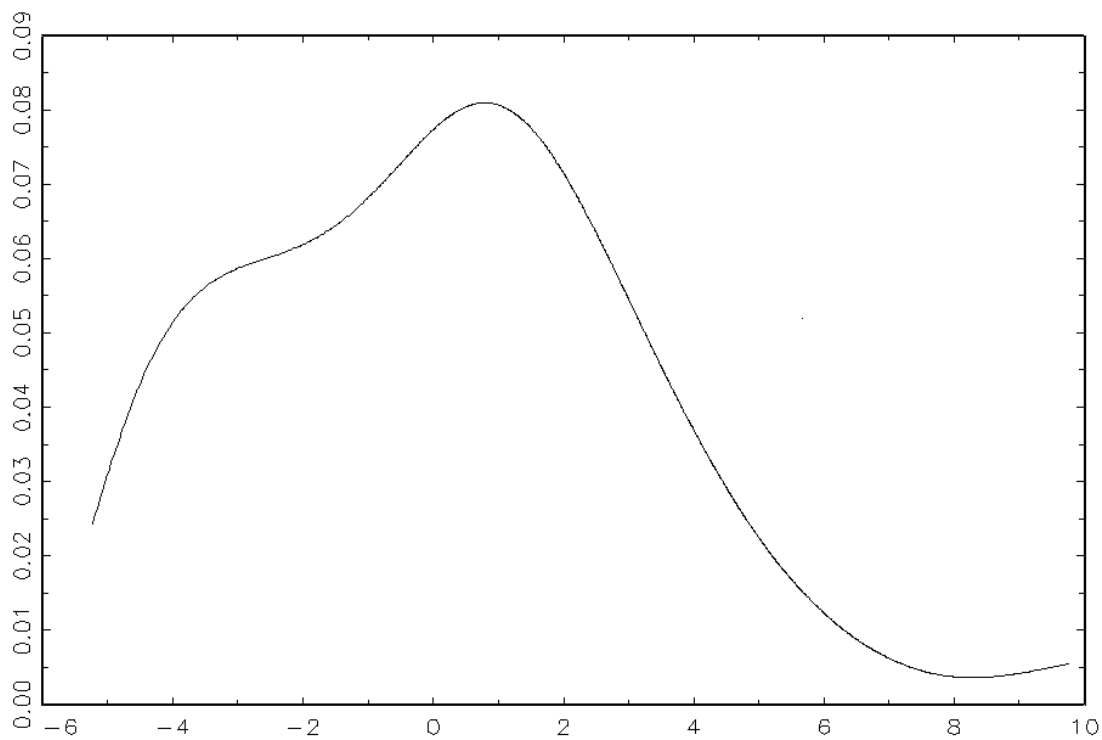


Figure 3.1
Estimated error distribution in the insurance equation

Appendix A

1 Main Results

In the proofs of Theorems 1-2 below we provide proofs for the large sample properties of the second stage estimator, with the argument for the first-stage estimator being similar but shorter as it is based on a regular expectation. In so doing, we simplify notation by not subscripting objective functions, gradients, and hessian expressions.

Proof of Theorem 1. (Consistency: $\hat{\theta}_2$). Define:

$$\hat{Q}(\theta) \equiv \left\langle \hat{\tau}_v \left[Y - \hat{f}/\hat{g}^* \right]^2 \right\rangle; \quad Q(\theta) \equiv \left\langle \tau_v [Y - f/g]^2 \right\rangle$$

Then, recalling (D3), letting $\delta_i \equiv \hat{\tau}_v \left| \hat{f}_i/\hat{g}_i^* - f_i/g_i \right|$, and $\varepsilon_i \equiv Y_i - f_i/g_i$:

$$\left| \hat{Q}(\theta) - Q(\theta) \right| \leq C + S + T$$

$$C \equiv 2 \langle |Y| \delta \rangle; \quad S \equiv \left\langle \left(\hat{f}/\hat{g}^* + f/g \right) \delta \right\rangle; \quad T \equiv \frac{1}{N} \sum |\hat{\tau}_{vi} - \tau_{vi}| \varepsilon_i^2$$

For C , with $\hat{C}_1^2 \equiv 4 \langle \hat{\tau}_v Y^2 \rangle = O_p(1)$, from Cauchy's inequality and Lemma 5:

$$C \leq \hat{C}_1 \langle \delta^2 \rangle^{1/2} = O_p(1) \langle \delta^2 \rangle^{1/2} = o_p(1)$$

With a similar argument holding for S and T , $\hat{Q}(\theta)$ converges uniformly in θ to $Q(\theta)$ in probability. From standard arguments, $Q(\theta)$ converges uniformly to $E[Q(\theta)]$ in probability. Therefore:

$$\sup_{\theta} \left| \hat{Q}(\theta) - E[Q(\theta)] \right| \xrightarrow{p} 0.$$

From Ichimura (1993), $E[Q(\theta)]$ is uniquely maximized at θ_0 , which completes the proof.

We provide the proof for the asymptotic linear characterization in Theorem 2(b); other results are similarly obtained or follow directly.

Proof of Theorem 2. (Asymptotic Normality: $\hat{\theta}_2$). With $\hat{H}(\theta) \equiv \nabla_{\theta\theta'} \hat{Q}(\theta)$ and $\hat{G}(\theta) \equiv \nabla_{\theta'} \hat{Q}(\theta)$, from a Taylor series expansion:

$$\sqrt{N} \left(\hat{\theta}_2 - \theta_0 \right) = - \left[\hat{H}(\theta^+)^{-1} \right] \left[\sqrt{N} \hat{G}(\theta_0) \right], \quad \theta^+ \in \left[\hat{\theta}_2, \theta_0 \right].$$

For the Hessian, with $H(\theta) \equiv \nabla_{\theta\theta'} Q(\theta)$:

$$\sup_{\theta} \left| \hat{H}(\theta) - EH(\theta) \right| \leq \sup_{\theta} \left| \hat{H}(\theta) - H(\theta) \right| + \sup_{\theta} |H(\theta) - EH(\theta)|$$

From Lemma 5, the first term converges in probability to 0. From standard arguments, the second term also converges in probability to zero. Therefore, as $\theta^+ \xrightarrow{p} \theta_0 : \hat{H}(\theta^+) \xrightarrow{p} EH(\theta_0) \equiv H_0$

For the gradient, with $\hat{w} \equiv \nabla_{\theta} \hat{M}$:

$$\sqrt{N} \hat{G}(\theta_0) = \sqrt{N} \left[\langle [Y - M] \hat{\tau} \hat{w} \rangle - \left\langle \left[\hat{M} - M \right] \hat{\tau} \hat{w} \right\rangle \right] \equiv \sqrt{N} \left[\hat{G}_A - \hat{G}_B \right],$$

For \hat{G}_A , with $\varepsilon \equiv Y - M$, and $G_A \equiv \langle [Y - M] \tau w \rangle$,

$$\sqrt{N} \left[\hat{G}_A - G_A \right] = \sqrt{N} \left[\Delta_1 + \Delta_2 + \Delta_3 \right],$$

$$\Delta_1 \equiv \langle \varepsilon \tau [\hat{w} - w] \rangle; \quad \Delta_2 \equiv \langle \varepsilon [\hat{\tau} - \tau] w \rangle; \quad \Delta_3 \equiv \langle \varepsilon [\hat{\tau} - \tau] [\hat{w} - w] \rangle$$

From Lemma 9, $\Delta_1 \xrightarrow{p} 0$. For Δ_2 , let

$$\tau_i \equiv 1 \{c_{1o} < v_i(\theta_o) < c_{2o}\} \equiv \tau_i(\alpha_o), \alpha_o \equiv [\theta_o, c_{1o}, c_{2o}]$$

Employing a similar strategy to that in Klein (1993), let $N_\varepsilon \equiv \langle \alpha : |\alpha - \alpha_o| < \varepsilon \rangle$, $\varepsilon = o(1)$. Then, $\sqrt{N} \Delta_2 = o_p(1)$ if

$$\sup_{N_\varepsilon} N^{1/2} \sum [\tau_i(\alpha) - \tau_i(\alpha_o)] \varepsilon_i w_i / N = o_p(1)$$

for all $\varepsilon = o(1)$.¹ The result then follows from Pakes and Pollard (1989, Lemma 2.17, p. 1037).

Turning to Δ_3 , let $\tau^*(\hat{\alpha})$ be an indicator on the union of the sets over which $\tau(\hat{\alpha})$ and $\tau(\alpha_o)$ are defined. Then:

$$\sqrt{N} |\Delta_3| \leq \sqrt{N} \langle |\varepsilon| |\tau(\hat{\alpha}) - \tau(\alpha_o)| \tau^*(\hat{\alpha}) |\hat{w} - w| \rangle \leq \sqrt{N} |\Delta_{31}| |\Delta_{32}|,$$

$$|\Delta_{31}| \equiv \left[\sum \varepsilon_i^2 [\tau_i(\hat{\alpha}) - \tau_i(\alpha_o)]^2 / N \right]^{1/2}, \quad |\Delta_{32}| = \left[\sup_{N_\varepsilon} \sum \tau_i^*(\alpha) [\hat{w}_i - w_i]^2 / N \right]^{1/2}$$

To analyze $|\Delta_{31}|$, for $k = 1, 2$ let:

$$S(z) \equiv \left\{ 1 + \exp \left[- \left(N^{-(s-\varepsilon)} + z \right) / N^{-(s-\varepsilon)/2} \right] \right\}^{-1}, \quad 0 < \varepsilon < s$$

$$S_k^* \equiv S(|v(\theta) - v_0| + |c_k - c_{k0}| - |v_0 - c_{k0}|) + 1 - S(0), \quad k = 1, 2.$$

Then, from Klein (1993, Lemma A.1):

$$|\tau_i(\alpha_o) - \tau(\hat{\alpha})| \leq S_1^* + S_2^*.$$

Let $\delta_N \equiv |v - v_0| + |c_1 - c_{10}|$, $w_k \equiv |v_0 - c_{k0}|$, and write $S_k^* \equiv S(\delta_N - w_k) + 1 - S(0)$. Note that $|\hat{\alpha} - \alpha_o| = N^{-(s+\varepsilon)}$, $s > 1/4$ and that $1 - S(0) = o_p(N^{-s})$. As in Klein (1993, Lemma A.2), Taylor expand $S(\delta_N - w_k)$ in δ_N about $\delta_N = 0$. Assuming that $E(\varepsilon_i^2 | X_i)$ is bounded, it can then be shown that $|\Delta_{31}|^2 = O_p(N^{-1/2})$. Since $|\Delta_{32}|^2 = O_p(N^{-1/2})$, the result follows.

For \hat{G}_B , from Lemma 9:

$$\sqrt{N} \left[\hat{G}_B - \hat{G}_B^* \right] \xrightarrow{p} 0, \quad \hat{G}_B^* \equiv \left\langle \left[\hat{M} - M \right] \tau w \right\rangle$$

¹If uniformity holds for $q \in \mathcal{N}_\varepsilon$ for all $\varepsilon = o(1)$, then uniformity holds over $o_p(1)$ neighborhoods of q_o .

Next, noting that $\hat{M}_i \equiv \hat{f}_i/\hat{g}_i$, recalling the definition of K_{ij} in (D3), letting

$$\rho_{ij} \equiv \frac{1}{h} [Y_j K_{ij} - M K_{ij}] [\tau_i w_i] / g_i,$$

and employing Lemma 10:

$$\begin{aligned} & \sqrt{N} [\hat{G}_B^* - U_N] \xrightarrow{p} 0, \\ U_N & \equiv \left\langle \left(\hat{f}/\hat{g} - M \right) w \frac{\hat{g}}{g} \right\rangle = \left\langle \left(\hat{f} - \hat{g}M \right) \frac{w}{g} \right\rangle = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \rho_{ij} / N. \end{aligned}$$

A U-statistic has the form:

$$\binom{N}{2}^{-1} \sum_i \sum_{j > i} \rho_{ij}^*, \rho_{ij}^* = \rho_{ji}^*.$$

For U_N above:

$$\begin{aligned} N(N-1)U_N &= \sum_i \sum_{j > i} \rho_{ij} + \sum_i \sum_{j < i} \rho_{ij} = \sum_i \sum_{j > i} \rho_{ij} + \sum_j \sum_{i > j} \rho_{ij} \\ &= \sum_i \sum_{j > i} \rho_{ij} + \sum_j \sum_{i > j} \rho_{ij} = \sum_i \sum_{j > i} \rho_{ij} + \sum_i \sum_{j > i} \rho_{ji}. \end{aligned}$$

Therefore, with $\rho_{ij}^* = [\rho_{ij} + \rho_{ji}] / 2$, U_N has the conventional U-statistic form.

As discussed in section 2.1, with $U_N = B_S^*$, $E(\rho_{ij}) = 0 \implies E(U_N) = 0$ under index trimming and the residual property of w_i . Since U_N is a centered U-statistic and since it can be shown that $E(\rho_{ij}^* \rho_{ij}^*) = o(N)$, then (see Serfling(1980) and Powell, Stock, and Stoker(1989)): $\sqrt{N}[U_N - \hat{U}_N] = o_p(1)$, where:

$$\sqrt{N}\hat{U}_N \equiv N^{-1/2} \sum_i [E(\rho_{ij} | X_i, Y_i) + E(\rho_{ji} | X_i, Y_i)] \equiv T_1 + T_2$$

For T_1 : $E(\rho_{ij}) = 0 \implies E(T_1) = 0$. As $E(\rho_{ij} | X_i, Y_i) = O(h^2)$, it may be shown that $Var(T_1) \rightarrow 0$. It follows that $T_1 = o_p(1)$. The T_2 -term vanishes because $E(w_j | V_j) = 0$.

Therefore, $\sqrt{N}\hat{G}_B \xrightarrow{p} 0$, from which it follows that

$$\sqrt{N}(\hat{\theta}_2 - \theta_0) = -H_0^{-1} \left[\sqrt{N} \langle [Y - M] \tau w \rangle \right]$$

Asymptotic normality now follows from a standard central limit theorem.

Below we provide the proof for Theorem 3(a) which characterizes the k^{th} centered moment underlying the test statistic. Parts (b-c) of the theorem are either immediate or have arguments similar to (a).

Proof of Theorem 3. (Test Statistic: Asymptotic Null-Distribution) Define:

$$\sqrt{N}\hat{T}_k^*(\hat{\theta}) = \sqrt{N} \left\langle \left(Y - \hat{M}(\hat{\theta}) \right) \left(\hat{M}_k - \hat{E}(\hat{M}_k | V(\hat{\theta})) \right) \right\rangle$$

From a Taylor series expansion, Theorem 2, and Lemma 3:

$$\sqrt{N}\hat{T}_k^* (\hat{\theta}) = \sqrt{N}\hat{T}_k^* (\theta_0) - R_k, \quad R_k \equiv \langle \nabla_{\theta} w_k^* \rangle H_0^{-1} G_o + o_p(1)$$

With $\hat{w}_k^* \equiv \hat{M}_k - \hat{E}(\hat{M}_k|V)$, write $\hat{T}_k^* (\theta_0) = \hat{T}_{Ak}^* + \hat{T}_{Bk}^*$, where:

$$\hat{T}_{Ak}^* \equiv \langle (Y - M) \rangle \hat{w}_k^*; \quad \hat{T}_{Bk}^* \equiv - \left\langle \left(\hat{M} - M \right) \hat{w}_k^* \right\rangle$$

Analogous to the proof of Theorem 2, we show that $\sqrt{N}\hat{T}_{Ak}^* = \sqrt{N}T_k^* + o_p(1)$ and $\sqrt{N}\hat{T}_{Bk}^* = o_p(1)$, where:

$$T_k^* (\theta_0) = \langle (Y - M) (M_k - E(M_k|V)) \rangle$$

Write $\sqrt{N} \left[\hat{T}_{Ak}^* - T_k^* \right]$ as:

$$\sqrt{N} \left\langle (Y_i - M_i) \left\{ \left[\hat{M}_k - M_k \right] - \left[\hat{E}(\hat{M}_k|V) - E(M|V) \right] \right\} \right\rangle$$

Convergence in probability to zero then follows for the first component from Lemma 7 and for the second component from Lemma 8. For \hat{T}_{Bk}^* , from lemmas 9 -10, $\sqrt{N}\hat{T}_B = o_p(1)$ by an argument similar to that for \hat{G}_B in Theorem 2. Hence:

$$\sqrt{N}\hat{T}_k^* (\hat{\theta}) = \sqrt{N}T_k^* (\theta_0) - R_k(\theta_0) + o_p(1)$$

2 Intermediate Lemmas:

2.1 Convergence Rates

The proof of the following Lemma is due to Bhattacharaya (1967) and relies on an exponential bound due to Hoeffding (1963). A version of the proof is also contained in Klein (1993).

Lemma 1. (Uniform Convergence Rates for Bounded Functions). With z_i i.i.d., Let $m_i \equiv m(t; z_i, \theta)$ be random variables such that:

$$|m_i N^{-s}| = O(1)$$

Then, for θ and t in compact sets, m as the vector with i^{th} element m_i , and $\delta > 0$:

$$\sup_{t, \theta} |\langle m \rangle - E[\langle m \rangle]| = o_p \left(N^{-(1/2)+s+\delta} \right)$$

Lemma 2. Assume:

$$\langle \hat{a}\hat{a} \rangle = O_p(N^{-1}h^{-s}), \quad \langle \hat{b}\hat{b} \rangle = O_p(N^{-1}h^{-t}),$$

where $s + t < 6$. Then, with $h = O(N^{-r})$, $r < 1/6$: $\sqrt{N} \langle \hat{a}\hat{b} \rangle = o_p(1)$

Proof. The proof follows directly from Cauchy's inequality:

$$\left[\sqrt{N} \langle \hat{a}\hat{b} \rangle \right]^2 \leq N \langle \hat{a}\hat{a} \rangle \langle \hat{b}\hat{b} \rangle$$

Lemma 3. (Convergence Rates) For V a continuous random variable with density g_v , let $\nabla_{\theta}^d(g_v)$ be the d^{th} partial derivative of g with respect to θ , $\nabla_{\theta}^0(\hat{g}_v) \equiv \hat{g}_v$. Let $\hat{\psi}(t; \theta)$

refer to either $\hat{g}_v(t; v)$ or $\hat{f}_v(t; v)$ and let $\psi(t; \theta)$ refer to the corresponding true functions, g_v or f_v . Then, for θ in a compact set and t in a compact subset of the support of V , the following rates hold for $d = 0, 1, 2$:

$$\begin{aligned} a) & : \sup_{t, \theta} E \left\{ \left[\nabla_{\theta}^d \left(\hat{\psi}(t; \theta) \right) - E \left(\nabla_{\theta}^d \left(\hat{\psi}(t; \theta) \right) \right) \right]^2 \right\} = O \left(\frac{1}{Nh^{2d+1}} \right) \\ b) & : \sup_{t, \theta} \left| E \left(\nabla_{\theta}^d \left(\hat{\psi}(t; \theta) \right) \right) - \nabla_{\theta}^d \left(\psi(t; \theta) \right) \right| = O(h^2) \end{aligned}$$

Proof. As the proof is standard (e.g., see Klein, 1993), we outline it below. When $\psi(t; \theta) = g$ and $d = 1$. The variance calculation in (a) is immediate². For the bias calculation (b), write $E[\nabla_{\theta}^1(\hat{g}(t; \theta))]$ as:

$$\begin{aligned} \frac{1}{h} \int \nabla_{\theta}^1 (K[(t-v)/h]) g_x(x) dx &= \frac{1}{h} \nabla_{\theta}^1 \int K[(t-v)/h] g_x(x) dx = \\ \frac{1}{h} \nabla_{\theta}^1 \int K[(t-v)/h] g_v(v) dv &= \nabla_{\theta}^1 \int K(z) g_v(t+hz) dz \end{aligned}$$

The result now follows from a standard Taylor expansion in h , with t restricted to be away from the support boundary for V .

The test statistic depends on the marginal expectation of Y conditioned separately on each variable in the index. For a discrete variable Z , $E(Y|Z = t)$ can be estimated as the sample mean of Y for those observations at the support point or by using the same kernel representation employed for continuous random variables. Delgado and Mora (1995) provide a similar result using the nearest neighbor estimator. As the argument for kernels is very short, we provide it below.

Lemma 4. (Discrete Regressors). Let Z be a discrete random variable with support points $t_k : Pr(Z = t_k) > 0$. With t as one of these points, define the sample mean:

$$\bar{Y}(t) \equiv \sum_{Z_j = t} Y_j / N(t),$$

where $N(t)$ is the number of sample observations for which the random variable $Z = t$. Assuming $E|Y_j|$ is bounded, and that \hat{E} is a regular expectation with window parameter $r > 0$ (D3), then:

$$\left| \hat{E}(Y|Z = t) - \bar{Y}(t) \right| = O_p(1/N)$$

Proof. With $\{\bullet\}$ as an indicator on the indicated set, by definition $\hat{E}(Y|Z = t)$ is

²The estimator has the form:

$$\sum \frac{1}{h^2} k[(t-w_i)/h] / N$$

With the bias term vanishing faster than the second moment term, the order of the variance is given by:

$$\frac{E(k^2[(t-w_i)/h])}{h^4 N}$$

Letting $z = (w-t)/h$, a factor of h disappears in the Jacobian; the result follows.

given as:

$$\frac{\sum \{Z_j = t\} Y_j K(0) + \sum \{Z_j \neq t\} Y_j K [(t - Z_j) / h]}{\sum \{Z_j = t\} K(0) + \sum \{Z_j \neq t\} K [(t - Z_j) / h]} \equiv \frac{\bar{Y}(t) + \Delta_1}{1 + \Delta_0},$$

$$\Delta_d \equiv \sum_{Z_j \neq t} Y_j^d K [(t - Z_j) / h] / [N(t) K(0)], \quad d = 0, 1$$

Then,

$$\left| \bar{Y}(t) - \hat{E}(Y|Z=t) \right| = |[\Delta_0 \bar{Y}(t) - \Delta_1]| / [1 + \Delta_0] \leq |\Delta_1| + |\bar{Y}(t)| |\Delta_0|$$

For $|t - Z_j| > c$, a fixed positive and finite constant, $K [(t - Z_j) / h] = o(1/N^2)$. For the normal-kernel case, this term vanishes at an exponential rate. The result follows by taking expectations of both sides.

To establish consistency for the estimator, we require the relative convergence results below.

Lemma 5. (Adjusted Expectations) Recalling that X is bounded and that θ lies in a compact set, assume that $E \equiv E(Y|V)$ is bounded, where V is the index. From the tail condition (A6), Y has tails thinner than a t-distribution with $df \geq 4$ degrees of freedom. Define $\lambda \equiv df / (df - 1)$ and let $\varepsilon, \delta > 0$. Recalling the adjustment parameter α in (D3), let \hat{E}_A be an adjusted expectation with adjustment parameter $\alpha : 0 < \alpha < 1/2$ and window parameter r :

$$0 < r < \frac{1/2 - \delta}{\lambda(1 + \alpha) + \varepsilon}$$

Then, with ∇_{θ}^k as the partial derivative operator as defined above and recalling the definition of $\hat{g}^*(t)$:

$$(1) : \sup_{\theta} \left\langle \left[\hat{E}_A - E \right]^2 \right\rangle = o_p(1)$$

From (D3), recall that $\hat{E}_a \equiv \hat{f}(x\theta; \theta) / \hat{g}^*(x\theta; \theta)$. Assume that $\nabla_{\theta}^k E$ and $\nabla_{\theta}^k g$ are $O(1), k = 0, 1, 2$. From (D3), recall that $\hat{E}_a \equiv \hat{f}(x\theta; \theta) / \hat{g}^*(x\theta; \theta)$. Let the window parameter satisfy:

$$0 < r < \frac{1/2 - \delta}{\lambda(1 + k) + \varepsilon}$$

Then, for θ in an $o_p(1)$ neighborhood of $\theta_0, k = 0, 1, 2$, and for $D = \nabla_{\theta}^k [\hat{f}(x\theta; \theta) - f(x\theta; \theta)]$ or $\nabla_{\theta}^k [\hat{g}^*(x\theta; \theta) - \hat{g}^*(x\theta; \theta)]$:

$$(2) \sup_{x, \theta} \hat{\tau}(x\hat{\theta}) D / \hat{g}^*(x\theta; \theta)^a = o_p(1)$$

Proof. For (1), since $f(t)/g(t)$ is by assumption bounded, it suffices to show $T_f, T_g = o_p(1)$:

$$T_f \equiv \sup_{\theta} \left\langle \left[(\hat{f} - f) / \hat{g}^* \right]^2 \right\rangle; \quad T_g \equiv \sup_{\theta} \left\langle \left[(\hat{g} - g) / \hat{g}^* \right]^2 \right\rangle.$$

For T_f (the proof for T_g is similar), write $T_f \leq A + B$, where:

$$A \equiv \sup_{\theta} \left\langle \left| \left(\hat{f} - E\hat{f} \right) / \hat{g}^* \right| \right\rangle; \quad B \equiv \sup_{\theta} \left\langle \left[\left(E\hat{f} - f \right) / \hat{g}^* \right]^2 \right\rangle$$

Each of these terms is examined below.

A: Relative Convergence to Expectation

With $b > 0$, let $b_j \equiv 1$ if $|Y_j| < h^{-b}$ and 0 otherwise. Following a strategy employed by Ichimura (1993), consider separately bounded and unbounded regions for Y_j . Letting $K_j \equiv K [(t - v_j(\theta)) / h]$, define:

$$\hat{f}_b(t) \equiv \sum_{j=1}^N \frac{b_j Y_j}{hN} K_j; \quad \hat{f}_u(t) \equiv \sum_{j=1}^N \frac{(1 - b_j) Y_j}{hN} K_j$$

Then, $A \leq A_b + A_u$, where:

$$A_b \equiv \sup_{\theta, t} \left| \left(\hat{f}_b(t) - E\hat{f}_b(t) \right) / \hat{g}^*(t) \right|$$

$$A_u \equiv \sup_{\theta, t} \left| \left(\hat{f}_u(t) - E\hat{f}_u(t) \right) / \hat{g}^*(t) \right|.$$

Recall that $\hat{g}^*(t) \equiv \hat{g}(t) + h^a \hat{q}(1 - \hat{\tau})$, where $(1 - \hat{\tau})$ is a smoothed indicator that depends on lower and upper sample quantiles denoted by \hat{q}_a and \hat{q}_b . With q_a and q_b as the corresponding population quantiles, let $\mathcal{A}^* \equiv \{g : q_a^* < t < q_b^*\}$ be a fixed subset of the support for V that contains $\mathcal{A} \equiv \{t : q_a < t < q_b\}$. Define $\tau^*(t)$ as the indicator on \mathcal{A}^* , then letting

$$\Delta_b \equiv h^{-a} \sup_{\theta, t} \left| \left(\hat{f}_b(t) - E\hat{f}_b(t) \right) \right|$$

$$A_b \leq \Delta_b \left[h^a \sup_{\theta, t} |\tau^*(t) / \hat{g}(t)| + h^a \sup_{\theta, t} |[1 - \tau^*(t)] / \hat{\rho}(t)| \right]$$

On \mathcal{A}^* , $\inf \hat{g}(t) \xrightarrow{p} \underline{g} > 0$. On the complement, $\inf \hat{\tau} \xrightarrow{p} 0$. Therefore, $A_b \xrightarrow{p} 0$ if $\Delta_b \xrightarrow{p} 0$. From Lemma 4:

$$\Delta_b = O(h^{-a}) o_p \left(h^{-1-b} N^{-1/2+\delta} \right), \quad \delta > 0.$$

Since $h = O(N^{-r})$, $\Delta_b \xrightarrow{p} 0$ for $r < (1/2 - \delta) / (1 + a + b)$.

For A_u , $A_u \xrightarrow{p} 0$ if $\Delta_u \xrightarrow{p} 0$, where:

$$\Delta_u \equiv h^{-a} \sup_{\theta, t} \left| \left(\hat{f}_u(t) - E\hat{f}_u(t) \right) \right| \leq h^{-a} \sup_{\theta, t} \left| \hat{f}_u(t) \right| + h^{-a} \sup_{\theta, t} \left| E\hat{f}_u(t) \right|$$

With a similar argument holding for both terms, for the first term:

$$h^{-a} E \sup_{\theta, t} \left| \hat{f}_u(t) \right| \leq h^{-a-1} \frac{1}{N} \sum_j E [(1 - b_j) |Y_j|],$$

Employing the tail assumption on Y_j , it suffices to show convergence to zero for:

$$h^{-(1+a)} \int_{h^{-b}}^{\infty} y / \left([1 + y^2]^{(df+1)/2} \right) dy \leq h^{-(1+a)} \int_{h^{-b}}^{\infty} y / \left(y^{(df+1)} \right) dy.$$

With $df > 1$, the above bound is

$$O \left[h^{-(1+a)} h^{(df-1)b} \right],$$

which converges to zero for

$$b = \varepsilon + (a + 1) / (df - 1), \varepsilon > 0.$$

Combining this restriction with that on r above (for $\Delta_b \xrightarrow{p} 0$) and letting $\lambda \equiv df / (df - 1)$, the uniform convergence for term A follows with: $r < (1/2 - \delta) / [\lambda(1 + \alpha) + \varepsilon]$.

B: Relative Bias

Let X_k be a continuous variable supported on $[a_k, b_k]$. For $c : 2a < c < 1$, write $\mathbf{1}(\mathcal{A})$ as an indicator on \mathcal{A} , and with X_k , $k = 1, \dots, K_c$ as a continuous component of X , define:

$$S_N \equiv \{x : a_k + h^c < x_k < b_k - h^c, k = 1, \dots, K_c\}$$

where the product is taken of the $k = 1, \dots, K_c$ continuous X -variables. On S_N , $\sup_{\theta} B \xrightarrow{p} 0$ from Lemma 3. On the complement of S_N , it can be shown that B vanishes if the probability on this set vanishes sufficiently fast ($0 < 2\alpha < c$).

The proof for (2) can be based on a similar argument. Alternatively, we can exploit trimming to establish uniformity in an $o_p(1)$ neighborhood of θ_0 . To outline the argument for one of the terms in (2), write:

$$\Delta \equiv \mathbf{1} \left(\hat{a} - X \left(\hat{\theta} - \theta \right) < X\theta < \hat{b} + X \left(\theta - \hat{\theta} \right) \right) \left| \nabla_{\theta}^k \left[\hat{f}(x\theta; \theta) - f(x\theta; \theta) \right] \right| / \hat{g}^*(x\theta; \theta)^a$$

Denote $|c|$ as the vector with i^{th} element $|c_i|$. Then, with $\hat{\delta} \equiv |X| \left| \hat{\theta} - \theta \right|$ and with τ as the indicator on $X\theta$ s.t. $\hat{a} - \hat{\delta} < X\theta < \hat{b} + \hat{\delta}$, Δ is bounded above by:

$$\begin{aligned} & \tau \left| \nabla_{\theta}^k \left[\hat{f}(x\theta; \theta) - f(x\theta; \theta) \right] \right| / \hat{g}^*(x\theta; \theta)^a \\ & \leq \frac{\tau}{\hat{g}^*(x\theta; \theta)^a} \left[\left| \nabla_{\theta}^k \left[\hat{f}(x\theta; \theta) - E\hat{f}(x\theta; \theta) \right] \right| + \left| \nabla_{\theta}^k \left[E\hat{f}(x\theta; \theta) - f(x\theta; \theta) \right] \right| \right] \end{aligned}$$

The proof for the first term is similar but simpler to that in (1) because \hat{g}^* is uniformly close to g and in large samples τg is bounded away from 0. The argument for the second

term follows from Lemma 3b with minor modifications.

The following lemma provides a result that is useful for the recentered test statistic.

Lemma 6. Let $\tau_k M_{ik} \equiv E(Y_i | X_{ik})$ and $\hat{M}_{ik} \equiv \hat{E}_I(Y_i | X_{ik})$. Writing the window $h = O(N^{-r})$, refer to r as a window parameter. Then, assume that this estimated expectation is regular (D3) with window parameter $r_I : 1/6 < r_I < 1/4$. Consider the estimated "outer" expectation $\hat{E}_o(\hat{M}|V)$ and assume that it is regular with outer window parameter $r_o < r_I$. Below, we subscript h according to the window parameter upon which it is based (e.g., $h_o = O(N^{-r_o})$) Then:

$$\Delta_E \equiv \frac{1}{N} \sum_i \left[\hat{E}_o(\hat{M}_{ik} | V_{ik}) - \hat{E}_o(M_{ik} | V_{ik}) \right]^2 = o_p(N^{-1/2})$$

Proof. By definition, with $\delta_j \equiv \hat{M}_{jk} - M_{jk}$, $\Delta_E \leq \Delta_{E1} + \Delta_{E2}$, where:

$$\begin{aligned} \Delta_{E1} &\equiv \frac{1}{N} \sum_i \left[\sum_{j \neq i} \frac{1}{h_o^2 (N-1)^2} \delta_j^2 K_{ij}^2 \right] \\ \Delta_{E2} &\equiv \frac{1}{N} \sum_i \left[\sum_s \sum_{r \neq s} \frac{1}{(N-1)^2} \frac{|\delta_r| K_{ir}}{h_o} \frac{|\delta_s| K_{is}}{h_o} \right] \end{aligned}$$

Note that

$$|a| |b| \leq \max(a^2, b^2) \leq a^2 + b^2$$

Therefore, for Δ_{E2} , which converges in probability to 0 slower than Δ_{E1} :

$$0 < \Delta_{E2} < \frac{1}{N} \sum_i \left[\sum_s \sum_{r \neq s} \frac{1}{h_o^2 (N-1)^2} \left[\frac{\delta_r^2 K_{ir}^2}{h_o^2} + \frac{\delta_s^2 K_{is}^2}{h_o^2} \right] \right]$$

It suffices to show that $E(\Delta_{E2}) = o(N^{-1/2})$. From above:

$$E(\Delta_{E2}) = O(1) \left[E\left(\frac{\delta_r^2 K_{ir}^2}{h_o^2}\right) + E\left(\frac{\delta_s^2 K_{is}^2}{h_o^2}\right) \right]$$

Proceeding with the first term (the analysis for the second is identical), write: $\delta_r = \delta_r[i] + \delta_r^*[i]$, where $\delta_r^*[i] = O(1/hN)$ is the component of δ_r that depends on i and $\delta_r[i]$ is the remaining component after the i^{th} term has been removed. It can be shown that

$$\begin{aligned} E\left(\frac{\delta_r^2 K_{ir}^2}{h_o^2}\right) &= E\left(\frac{\delta_r^2[i] K_{ir}^2}{h_o^2}\right) + o(N^{-1/2}) \\ &= \frac{1}{h_o} E \left[E(\delta_r^2[i] | X_r) E\left(\frac{1}{h_o} K_{ir}^2 | X_r\right) \right] + o(N^{-1/2}) \end{aligned}$$

The first inner expectation is uniformly $O[\max(h_I^4, 1/(Nh_I))]$ while the second inner ex-

pectation is uniformly $O(1)$. Therefore, with $h_e = O(N^{-r_e})$, $h = O(N^r)$, and $r_e < r$:

$$E \left(\frac{\delta_r^2 K_{ir}^2}{h_o^2} \right) = O \left[\max(h_I^3, 1/(Nh_I^2)) \right] + o \left(N^{-1/2} \right) = o \left(N^{-1/2} \right), \quad 1/6 < r < 1/4.$$

Both the gradient and the moment conditions for the test statistic can be written as the sum of two components, each of which depends on estimated weights. The next two subsections show that in each of these components the weights may be taken as known.

2.2 Estimated Weights: $[Y - E(Y|v)] \hat{w}$

One of the components of the test statistic and of the gradient for the estimator depends on a weighted distance between the dependent variable and its expectation conditioned on an index. The following lemmas simplify this component.

Lemma 7. Define:

$$\hat{w}_i \equiv \left\{ \nabla_{\theta} \hat{E}_a(Y_i|V_i) \quad \text{or} \quad \hat{E}(Y_i|X_{ki}) \right\},$$

where \hat{E}_a is an adjusted expectation (D3) with window parameter $r : 1/8 < r < 1/4$. The expectation \hat{E} is regular with window parameter $r_k = r$. With $S_i \equiv X_{ki}$ or the index, V_i , define $\tau(S_i)$ as the indicator on $a < S_i < b$. Assume $E|Y_j^2|X_j| \leq \bar{\sigma}^2 = O(1)$. Then, with $u_i \equiv (Y_i - M_i)$:

$$D \equiv \sqrt{N} \langle \tau(V_i) u \tau(S_i) (\hat{w} - w) \rangle = o_p(1)$$

Proof. We provide the proof for $\hat{w}_i \equiv \nabla_{\theta} \hat{E}_a(Y_i|V_i)$, as the proof for the other weight is similar. Consider $\hat{w}_i^* \equiv \nabla_{\theta} \hat{E}(Y_i|V_i)$, where \hat{E} is regular with window parameter r . Since:

$$\sqrt{N} \langle \tau(V_i) u (\hat{w} - \hat{w}^*) \rangle = o_p(1),$$

we need to establish convergence in probability to 0 for

$$D^* \equiv \sqrt{N} \langle \tau(V_i) u (\hat{w}^* - w) \rangle$$

Recalling from (D3) that for regular expectations: $\delta \equiv \hat{w}_i^* - w_i = \nabla_{\theta} \left(\hat{f}_i / \hat{g}_i \right) - \nabla_{\theta} (f_i / g_i)$, this differential can be written as a sum of similar terms, one of which is given as

$$\nabla_{\theta} \hat{f}_i / \hat{g} - \nabla_{\theta} f_i / g_i = \left[g_i \left(\nabla_{\theta} \hat{f}_i - \nabla_{\theta} f \right) - \nabla_{\theta} f (\hat{g}_i - g_i) \right] / \hat{g}_i g_i$$

With similar arguments holding for the other terms, we analyze the first term. With $\Delta_i \equiv \nabla_{\theta} \hat{f}_i - \nabla_{\theta} f_i$, this term is given as:

$$\sqrt{N} \langle \tau(V_i) u \Delta / \hat{g} \rangle = D_1^* + o_p(1), \quad D_1^* \equiv \sqrt{N} \langle \tau(V_i) u \Delta / g \rangle.$$

Employing a mean-square convergence argument, $E \left[(D_1^*)^2 \right] = S + C$:

$$S \equiv E \langle u^2 \Delta^2 \rangle; \quad C \equiv \frac{1}{N} \sum_i \sum_{j \neq i} E(u_i u_j \Delta_i \Delta_j)$$

Taking an iterated expectation, S tends to zero. For C , write:

$$\Delta_i = \Delta_i [j] + \bar{\Delta}_i; \quad \Delta_j = \Delta_j [i] + \bar{\Delta}_j,$$

where $\bar{\Delta}_i$ and $\bar{\Delta}_j$ do not depend on Y_i or on Y_j . Then:

$$C = O(N)E(u_i \Delta_j [i]) E[u_j \Delta_i [j]] = O(N)O\left(\frac{1}{Nh^2}\right)^2 = O\left(\frac{1}{Nh^4}\right) = o(1).$$

With the exception of one weight component in the recentered moment conditions, the lemma above will be applied to simplify both the gradient for the estimator and the moment conditions. The complication, which is due to the recentering, is covered by Lemma 8 below.

Lemma 8. Referring to Lemma 6, let $M_{ik} \equiv E(Y_i | X_{ik})$ and $\hat{M}_{ik} \equiv \tau_k \hat{E}_I(Y_i | X_{ik})$. Let \hat{E}_I and \hat{E}_o be regular non-parametric expectations with respective windows r_I and r_o satisfying the restrictions in Lemma 6. Then:

$$\Delta \equiv \sqrt{N} \left\langle \hat{\tau} [Y - M] [\hat{E}_o(\hat{M}|V) - E(M_k|V)] \right\rangle \xrightarrow{p} 0$$

Proof. Let $\hat{w}_k \equiv \hat{E}_o(M_k|V)$, $w_k \equiv E(M_k|V)$, and $u \equiv [Y - M]$. Employing the same arguments as in the proof to Theorem 2 and Lemma 4.21 of Pakes and Pollard(1989), take the trimming function as known and write $\Delta = \Delta_1 + \Delta_2 + o_p(1)$,

$$\Delta_1 = \sqrt{N} \langle \tau u [\hat{w}_k - w_k] \rangle; \quad \Delta_2 = \sqrt{N} \left\langle \tau u [\hat{E}_o(\hat{M}|V) - \hat{E}_o(M_k|V)] \right\rangle$$

From Lemma 7, $\Delta_1 \xrightarrow{p} 0$. For the second term, the convergence rate in the expectation differential is not sufficient in itself to establish the desired result. Accordingly, in what follows we show that Δ_2 simplifies to a term whose expected square converges to zero.

To simplify Δ_2 , substitute from the definitions in the statement of the lemma to obtain:

$$\Delta_2 = N^{-1/2} \sum_{i=1}^n \frac{\tau_i u_i}{\hat{g}_i} \sum_{j \neq i}^n \frac{1}{h_o(N-1)} [\hat{M}_j - M_j] k_{ij}$$

With Δ'_2 defined by replacing \hat{g}_i with g_i in Δ_2 , it can be shown that $\Delta_2 = \Delta'_2 + o_p(1)$.

By definition:

$$\Delta'_2 = N^{-1/2} \sum_{i=1}^n \tau_i u_i \frac{1}{g_i} \sum_{j \neq i}^n \frac{1}{h_o(N-1)} \left[\frac{\hat{f}_{1j}}{\hat{g}_{1j}} - \frac{f_{1j}}{g_{1j}} \right] k_{ij}$$

It can also be shown that:

$$\Delta'_2 = N^{-1/2} \sum_{i=1}^n \tau_i u_i \frac{1}{g_i} \sum_{j \neq i}^n \frac{1}{h_o(N-1)} \left[\frac{\hat{f}_{1j}}{\hat{g}_{1j}} - \frac{f_{1j}}{g_{1j}} \right] [\hat{g}_{1j}/g_{1j}] k_{ij} + o_p(1)$$

Writing Δ''_2 for the expression above:

$$\begin{aligned}
\Delta_2'' &\equiv O\left(N^{-3/2}h_o^{-1}\right)\sum_{r=1}^n\frac{\tau_r u_r}{g_r}\sum_{j\neq r}^n[\hat{f}_{1j}-f_{1j}\hat{g}_{1j}M_j]k_{rj} \\
&= O\left(N^{-5/2}h_o^{-1}h_I^{-1}\right)\sum_{r=1}^n T_r, \quad T_r \equiv \frac{\tau_r u_r}{g_r}\sum_{j\neq r}^n\sum_{l\neq j}^n[Y_l k_{lj}^1 - k_{lj}^1 M_j]k_{rj}.
\end{aligned}$$

To complete the argument, we show that $E\left[(\Delta'')^2\right] = o(1)$. Squaring Δ_2'' and noting that $h_I^{-2} > h_o^{-2}$, the expectation of the cross-product terms is:

$$E(CP) = O(N^{-5})O(h_I^{-4})O(N^2)E(T_r T_s)$$

In T_r , for each j , there are $O(1)$ terms that depend on Y_r or on Y_s . Therefore, there are $O(N)$ such terms obtained by summing over j . Similarly, there are $O(N)$ such terms in T_s . Except for these $O(N^2)$ terms, all others vanish in expectation. Therefore:

$$E(CP) = O(N^{-5})O(h_I^{-4})O(N^2)O(N^2) = O\left(\frac{1}{Nh^4}\right)$$

For $h_I = O(N^p)$, $p < 1/4$, the above expectation vanishes. The argument for the squared terms in Δ_2'' is similar.

2.3 Estimated Weights: $\left[\hat{E}(Y|V) - E(Y|V)\right]\hat{w}$,

This weighted component appears in both the test statistic and the gradient for the estimator. The lemmas below show that it is close in probability to a simplified term.

Lemma 9. With $h = O(N^{-r})$, $\frac{1}{8} < r < \frac{1}{4}$, then, with \hat{w} as :

$$(a) : \frac{\partial \hat{E}(Y|V)}{\partial \theta}; \quad (b) : \hat{E}(Y|X_k); \quad \text{or} \quad (c) : \hat{E}[\hat{\tau}_k \hat{E}(Y|X_k)|V],$$

$$\Delta \equiv \sqrt{N} \left[\left\langle \hat{\tau}_v (\hat{M} - M) \hat{w} \right\rangle - \left\langle \tau_v (\hat{M} - M) w \right\rangle \right] = o_p(1)$$

Proof. The arguments for (a), (b), and (c) are similar. For \hat{w} in (c), write:

$$\Delta \equiv \sqrt{N} \left\langle \tau_v (\hat{M} - M) (\hat{w} - w) \right\rangle + \sqrt{N} \left\langle \tau_v (\hat{M} - M) (\hat{\tau}_v - \tau_v) \hat{w} \right\rangle$$

For the first term, the result follows from Lemmas 2 and 3. The argument for the second term is similar (see the section of the proof of Theorem 2 relating to indicators).

Lemma 10. (A Linear Characterization) Under the same window condition as in Lemma 9 and with \hat{M} as the vector with i^{th} element $\hat{M}_i \equiv \hat{f}_i/\hat{g}_i$:

$$\sqrt{N} \left[\left\langle (\hat{M} - M) w (\hat{g}_v/g_v) \right\rangle - \left\langle (\hat{M} - M) w \right\rangle \right] = o_p(1)$$

Proof. The proof follows from Lemmas 2 and 3.

Appendix B

1 Notation, Assumptions, and Intermediate Results

Here, we provide the large sample theory for the estimated parameters in the differenced expenditure equation. Klein, Shen, and Vella (2009), hereafter KSV, provide the large sample theory for the estimates of a class of joint binary double index models that includes the model considered here. Since these results are required for the differenced expenditure equation, we begin by briefly summarizing them. Referring to Section 2, let

$$\alpha_o = \begin{pmatrix} \beta_I \\ \beta_A \end{pmatrix}, V(\alpha_o) = \begin{pmatrix} V_I(\beta_I) \\ V_A(\beta_A) \end{pmatrix} \equiv V_o$$

and their corresponding estimators:

$$\hat{\alpha} = \begin{pmatrix} \hat{\beta}_I \\ \hat{\beta}_A \end{pmatrix}, V(\hat{\alpha}) \equiv \hat{V}$$

With H_o as the hessian w.r.t. the binary quasi-likelihood in KSV, it is shown:

$$\sqrt{N}(\hat{\alpha} - \alpha_o) \xrightarrow{d} W \sim N(0, -H_o^{-1})$$

As the estimator will depend on nonparametric expectations, we next provide their definitions. For Z as any variable in the model (X or Y), $K_{Aij} \equiv K[(v_{Ai} - v_{Aj})/h]$, $K_{Iij} \equiv K[(v_{Ii} - v_{Ij})/h]$, the estimated conditional expectation is denoted as $\hat{E}_{zi} \equiv \hat{E}(Z|V_{Ai} = v_{Ai}, V_{Ii} = v_{Ii})$ and is given by:

$$\begin{aligned} \hat{E}_{zi} &\equiv \hat{f}_i/\hat{g}_i \equiv \hat{E}_{zi}(V(\alpha_o)) \\ \hat{f}_i &\equiv \frac{1}{(N-1)h} \sum_{j \neq i} Z_j \hat{\tau}_j K_{Aij} K_{Iij} \\ \hat{g}_i &\equiv \frac{1}{(N-1)h} \sum_{j \neq i} \hat{\tau}_j K_{Aij} K_{Iij} \end{aligned}$$

where $\hat{\tau}_j$ is a trimming function that provides protection from small denominators. We set $h = O(N^{-r})$ with $r = 1/5$ as the point-wise optimal window parameter.

Recall that we have the expenditure equation:

$$Y_E = X_c \beta_c + c + I\theta_E + \varepsilon_E$$

Define control functions:

$$G_d(V_o) \equiv E(\varepsilon_E | A = 1, I = d, Z) = E(\varepsilon_E | A = 1, I = d, V_o) \text{ where } d \in \{1, 0\}.$$

We can rewrite the expenditure equation as

$$\begin{aligned} Y_E &= X_c \beta_c + c + I\theta_E + G_d(V_o) + u_d^* \\ \text{where } u_d^* &= \varepsilon_E - G_d(V_o) \end{aligned}$$

$$E[u_d^*|A=1, I=d, Z] = E[u_d^*|A=1, I=d, V_o] = 0$$

Partitioning X_c into X_{c1} and X_{c0} according to whether $I = 1$ or 0 :

$$Y_E = \begin{cases} X_{c1}\beta_c + c + \theta_E + G_1 + u_1^* & : A=1, I=1 \\ X_{c0}\beta_c + c + G_0 + u_0^* & : A=1, I=0 \end{cases}$$

Since the control function is unknown, we employ an extension of Peter Robinson's differencing method (Robinson, 1988):

$$Y_E - E(Y_E|A=1, I=d, V_o) = \begin{cases} [X_{c1} - E(X_{c1}|A=1, I=d, V_o)]\beta_c + u_1^* & : d=1 \\ [X_{c0} - E(X_{c0}|A=1, I=d, V_o)]\beta_c + u_0^* & : d=0 \end{cases}$$

Defining

$$Y^* = \begin{pmatrix} Y_E - E(Y_E|A=1, I=1, V_o) \\ Y_E - E(Y_E|A=1, I=0, V_o) \end{pmatrix},$$

$$X^* = \begin{pmatrix} X_{c1} - E(X_{c1}|A=1, I=1, V_o) \\ X_{c0} - E(X_{c0}|A=1, I=0, V_o) \end{pmatrix},$$

$$u^* = \begin{pmatrix} u_1^* \\ u_0^* \end{pmatrix},$$

we can rewrite the differenced Y_E equation as

$$Y^* = X^*\beta_c + u^*.$$

Since OLS is not feasible, we need to replace all true values with the corresponding estimates. Define

$$\hat{Y}^* = \begin{pmatrix} Y_E - \hat{E}(Y_E|A=1, I=1, \hat{V}) \\ Y_E - \hat{E}(Y_E|A=1, I=0, \hat{V}) \end{pmatrix}$$

$$\hat{X}^* = \begin{pmatrix} X_{c1} - \hat{E}(X_{c1}|A=1, I=1, \hat{V}) \\ X_{c0} - \hat{E}(X_{c0}|A=1, I=0, \hat{V}) \end{pmatrix}$$

$$\Delta_Y(\hat{\alpha}) = Y^* - \hat{Y}^*; \quad \Delta_X(\hat{\alpha}) = \beta_c(X^* - \hat{X}^*)$$

The feasible OLS estimating equation can now be written as:¹

$$\hat{Y}^* = \hat{X}^*\beta_c + \varepsilon, \quad \varepsilon = u^* - \Delta_Y(\hat{\alpha}) + \Delta_X(\hat{\alpha})$$

With the OLS estimator having the following form:

$$\sqrt{N}(\hat{\beta}_c - \beta_{co}) = \left(\hat{X}^{*'}\hat{X}^*/N\right)^{-1} \sqrt{N}\hat{X}^{*'}(u^* - \Delta_Y(\hat{\alpha}) + \Delta_X(\hat{\alpha}))/N,$$

we now proceed to show that this estimator is consistent and asymptotically distributed as normal.

Lemma 1. Double convergence. With $\langle xy \rangle \equiv \sum x_i y_i / N$, assume:

¹We note that all available observations are employed in estimating the expectations used. However, in the final OLS step, for technical reasons we need to exclude observations where estimated indices are too close to their support boundaries.

$$\langle \hat{a}\hat{a} \rangle = O_p(N^{-t}), \langle \hat{b}\hat{b} \rangle = O_p(N^{-s}), \text{ where } s + t > 1.$$

Then, $\sqrt{N} \langle \hat{a}\hat{b} \rangle = o_p(1)$.

Proof. The proof follows directly from Cauchy's inequality:

$$\left[\sqrt{N} \langle \hat{a}\hat{b} \rangle \right]^2 \leq N \langle \hat{a}\hat{a} \rangle \langle \hat{b}\hat{b} \rangle$$

Lemma 2. Employing notation and results above, with Ω as a positive definite matrix:

$$\hat{X}^{*'} \hat{X}^* / N \xrightarrow{P} \Omega$$

Proof. We begin by showing that

$$(a) : \hat{X}^{*'} \hat{X}^* / N - X^{*'} X^* / N \xrightarrow{P} 0$$

To see that this result must hold, write:

$$\begin{aligned} \hat{X}^* - X^* &= E(X|V) - \hat{E}(X|\hat{V}) \\ &= E(X|V) - \hat{E}(X|V) \\ &\quad + \hat{E}(X|V) - \hat{E}(X|\hat{V}) \end{aligned}$$

The first term $E(X|V) - \hat{E}(X|V)$ goes to zero from KSV. For the second term, which arises because the index has been estimated, from a Taylor series expansion:

$$\hat{E}(X|V) - \hat{E}(X|\hat{V}) = \nabla_{\alpha} \hat{E}(X|V^+)(\hat{\alpha} - \alpha_o) = \nabla_{\alpha} E(X|V)(\hat{\alpha} - \alpha_o) + o_p(1)$$

The result in (a) now follows. To complete the argument, from standard convergence arguments

$$(b) : X^{*'} X^* / N \xrightarrow{P} \Omega.$$

Lemma 3. With $\hat{E}_i(\alpha_0) = \hat{f}_i / \hat{g}_i$, assume that

$$\sum \left[\hat{f}_i - \hat{g}_i E_i \right]^2 / N = O_p(N^s)$$

and define:

$$\hat{B} \equiv \sum \left[\hat{E}_i(\alpha_0) - E_i(\alpha_0) \right] \hat{\tau}_i \hat{w}_i / N,$$

where \hat{w}_i is to be viewed as an estimated weight that converges to a fixed weight:

$$\sum [\hat{w}_i - w_i]^2 / N = O_p(N^t), s + t > 1$$

Then: $\sqrt{N} \hat{B} \xrightarrow{P} 0$.

Proof. To simplify \hat{B} , we first deal with the estimated denominator in $\hat{E}_i(\beta_o)$, Write \hat{B} as:

$$\hat{B} \equiv \sum \left[\hat{f}_i / \hat{g}_i - E_i \right] \hat{\tau}_i \hat{w}_i / N,$$

From double convergence (Lemma 1):

$$\sqrt{N} [\hat{B} - \hat{B}^*] \xrightarrow{p} 0, \hat{B}^* = \sum [\hat{f}_i/\hat{g}_i - E_i] [\hat{g}_i/g_i] \hat{\tau}_i \hat{w}_i/N$$

Employing similar arguments, it can be shown that

$$\sqrt{N} [\hat{B}^* - U_N] \xrightarrow{p} 0, U_N = \sum [\hat{f}_i/\hat{g}_i - E_i] [\hat{g}_i/g_i] \tau_i w_i/N$$

The lemma will now follow if $\sqrt{N}U_N \xrightarrow{p} 0$. To proceed, in U_N substitute the expressions for \hat{f}_i and \hat{g}_i above to obtain:

$$U_N = \binom{N}{2}^{-1} \sum \sum (\rho_{ij} + \rho_{ji}) / 2$$

where $\rho_{ij} = \frac{1}{h} (Z_j K_{ij} - E_i K_{ij}) w_i / g_i$

There are two properties of ρ_{ij} that we will require to analyze U_N . First, its expectation is 0. To establish this result, note that:

$$E [(Z_j K_{ij} - E_i K_{ij}) | X] = H(V)$$

Therefore, from iterated expectations

$$E [\rho_{ij}] = \frac{1}{h} E [H(V) w_i / g_i] = E [H(V) E(w_i | V) / g_i] = 0$$

As a second property of ρ_{ij} , it can be shown that it does not explode too fast in that $E(\rho_{ij}^2) = O(N)$. Therefore, from Powell, Stock, and Stoker (1989) and Serfling (1980):

$$\sqrt{N}(U_N - \hat{U}_N) \xrightarrow{p} 0$$

where $\hat{U}_N = N^{-1/2} \sum (E(\rho_{ij} | X_i, Y_i) + E(\rho_{ji} | X_i, Y_i))$.

From the properties of the weight function, $E(\rho_{ji} | X_i, Y_i) = 0$. For the first term:

$$E [E(\rho_{ij} | X_i, Y_i)] = 0$$

Furthermore, it can be shown that the first term has the form:

$$E(\rho_{ij} | X_i, Y_i) = h^2 T_i,$$

where T_i is bounded. Therefore:

$$\sqrt{N} \hat{U}_N = \sqrt{N} h^2 \sum T_i / N$$

converges to zero in probability as it has expectation of 0 and variance that converges to zero.

Lemma 4. Referring to the form for the estimator shown above:

$$\hat{X}^{*'}(u^* - \Delta_Y(\hat{\alpha}) + \Delta_X(\hat{\alpha}))/N \xrightarrow{p} 0$$

Proof. For the first term, since $\hat{X}_i^* \rightarrow X^*$, it can be shown that:

$$\hat{X}^{*'}u^*/N - X^{*'}u^*/N \xrightarrow{p} 0,$$

with $X^{*'}u^*/N \xrightarrow{p} 0$. For the second term, from a Taylor series expansion:

$$\hat{X}^{*'}\Delta_Y(\hat{\alpha})/N = \hat{X}^{*'}\Delta_Y(\alpha_0)/N + \nabla_{\alpha} \left[\hat{X}^{*'}\Delta_Y(\alpha_0) \right] (\hat{\alpha} - \alpha_0) + o_p(1)$$

From Lemma 3, the first term vanishes, while the second term vanishes from the asymptotic form for the estimator for the index parameters shown above. The argument for the $\Delta_X(\hat{\alpha})$ term is identical.

2 Main Results

Theorem 1 Consistency. $\hat{\beta}_c - \beta_c \xrightarrow{p} 0$

Proof. Consistency is immediate from Lemma 1-2. ■

Theorem 2 Normality. Define Ω as above and let

$$\begin{aligned} d &= \nabla_{\alpha} [-\Delta_Y(\alpha_0) + \Delta_X(\alpha_0)] \text{ and } A \equiv p \lim \left(X^{*'}d/N \right) \\ C_1 &\equiv E(S_{1N}S'_{1N}), \quad S_{1N} \equiv \sqrt{N}X^{*'}u^*/N \\ C_2 &\equiv A(H_o^{-1})A' \end{aligned}$$

Then:

$$\sqrt{N} \left(\hat{\beta}_c - \beta_c \right) \xrightarrow{d} W,$$

where W is distributed as:

$$\begin{aligned} W &\sim N(0, \Sigma), \\ \Sigma &= \Omega^{-1} [C_1 + C_2] \Omega^{-1} \end{aligned}$$

Proof. Recall that

$$\sqrt{N} \left(\hat{\beta}_c - \beta_c \right) = \left(\hat{X}^{*'}\hat{X}^*/N \right)^{-1} \sqrt{N}\hat{X}^{*'}(u^* - \Delta_Y(\hat{\alpha}) + \Delta_X(\hat{\alpha}))/N$$

From Lemma 1, $\left(\hat{X}^{*'}\hat{X}^*/N \right) \xrightarrow{p} \Omega$. Therefore, it suffices to show that the other (gradient) term is normally distributed in large samples with covariance matrix $C_1 + C_2$.

To establish this result, we first simplify the gradient term by showing that we may replace \hat{X}^* with X^* . Consider:

$$\sqrt{N}(\hat{X}^* - X^*)'(u^* - \Delta_Y(\hat{\alpha}) + \Delta_X(\hat{\alpha}))/N$$

With $w_i \equiv u^*$, from Lemma 2, the first term vanishes. For the second and third terms, from a Taylor expansion in $\hat{\alpha}$, both terms vanish from double convergence arguments.

Proceeding with the simplified gradient term, we need to establish asymptotic normality for:

$$S_N \equiv \sqrt{N} X^{*'} (u^* - \Delta_Y(\hat{\alpha}) + \Delta_X(\hat{\alpha})) / N$$

Recalling that $d = \nabla_{\alpha} [-\Delta_Y(\alpha_0) + \Delta_X(\alpha_0)]$, Taylor expanding in $\hat{\alpha}$:

$$\begin{aligned} S_N &= S_{1N} + S_{2N} + S_{3N}, \\ S_{1N} &\equiv \sqrt{N} X^{*'} u^* / N \\ S_{2N} &\equiv \sqrt{N} X^{*'} [-\Delta_Y(\alpha_0) + \Delta_X(\alpha_0)] / N \\ S_{3N} &= \left(X^{*'} d / N \right) \sqrt{N} (\hat{\alpha} - \alpha_0) \end{aligned}$$

With $w_i \equiv X^*$, from Lemma 4 the second term vanishes in probability. With $A \equiv p \lim \left(X^{*'} d / N \right)$ as above:

$$\begin{aligned} S_{3N}^* &\equiv A \sqrt{N} (\hat{\alpha} - \alpha_0) \\ S_N^* &= S_{1N} + S_{3N}^*, \end{aligned}$$

Then, as S_{2N} vanishes in probability,

$$S_N - S_N^* = \left[\left(X^{*'} d / N \right) - A \right] \sqrt{N} (\hat{\alpha} - \alpha_0)$$

Since the first component converges in probability to 0 and the second component converges in distribution, the product of these two components converges in probability to 0 (Slutsky). From above it suffices to analyze S_N^* :

$$S_N^* = \sqrt{N} X^{*'} u^* / N - \sqrt{N} A (\hat{\alpha} - \alpha_0) \equiv u_1 - u_2$$

To further simplify the analysis, as in typical selection models, we show that the covariance between these error components is zero. To this end, suppose for observation i we have $A_i = 1$ and $I_i = 1$ (the argument for the $A_i = 1$ and $I_i = 0$ case will be identical), then:

$$E \left[u_{1i} u_{2i}' \right] = [E (u_{1i} u_{2i} | A_i = 1, I_i = 1, V)] P_{11} \equiv C * P_{11}$$

To show that this covariance is zero, it suffices to show that $C = 0$. With $\bar{A}_i \equiv \{A_j, j \neq i\}$ and with $\bar{I}_i \equiv \{I_j, j \neq i\}$

$$\begin{aligned} C &= E \left[E (u_{1i} u_{2i} | A_i = 1, I_i = 1, \bar{A}_i, \bar{I}_i, V) \right] \\ &= E \left[E (u_{1i} | A_i = 1, I_i = 1, \bar{A}_i, \bar{I}_i, V) u_{2i} \right] \\ &= E \left[E (u_{1i} | A_i = 1, I_i = 1, V) u_{2i} \right] = 0, \end{aligned}$$

because of the recentered form for u_1 . From the form of S_N^* , a standard central limit theorem applies to yield normality, with expectation 0 and covariance matrix given by the sum of

the covariance matrices for the two elements of S_N^* , namely:

$$S_N^* \xrightarrow{d} S^{*\sim} N(0, C_1 + C_2),$$

$$C_1 \equiv E(S_{1N}S'_{1N}); C_2 \equiv E(S_{3N}S'_{3N}),$$

where the first component accounts for heteroscedasticity and the second accounts for parameter estimation uncertainty. The theorem now follows. ■

Remark 1 *With minor changes in notation, the above theorem also applies in a GLS step. With estimates given from above, define the residual and conditional variance function as:*

$$\hat{\varepsilon} = \hat{Y}^* - \hat{X}^* \hat{\beta}_c$$

$$\hat{S}^2(X) = \hat{E}(\hat{\varepsilon}^2 | \hat{v}_A, \hat{v}_I)$$

With all variables defined relative to $\hat{S}(X)$ and with C_1 above redefined as the identity matrix, Theorem 2 immediately extends to the GLS version of this estimator.

References

- Ahn, H. (1997), Semiparametric Estimation of a Single-Index Model with Nonparametrically Generated Regressors, *Econometric Theory* 13, 3-31.
- Bhattacharaya, P.K. (1967), Estimation of a Probability Density Function and its Derivatives. *Indian Journal of Statistics Series A*, 373-383
- Bierens, Herman J., (1990), A consistent conditional moment test of functional form. *Econometrica* 58, 1443-1458.
- Blundell, R. W. and Powell, J. L., 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71(3), 655-679.
- Cardon, J. H. and Hendel I., 2001. Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey. *RAND Journal of Economics* 32(3), 408-27.
- Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 1985-2007.
- Chiappori, P. and Salanie B., 2001. Testing for Asymmetric Information in Insurance Markets, *Journal of Political Economy* 108(1), 56-78.
- Climov, D. , M. Delecroix & L. Simar (2002), Semiparametric estimation in single index Poisson regression: a practical approach, *Journal of Applied Statistics* 29, 1047-1070.
- Delgado, M A. & J. Mora (1995), Nonparametric and semiparametric inference with discrete regressors. *Econometrica* 63, 1477-1484.
- Delgado, M A. & T. Stengos (1994), Semiparametric specification testing of non-nested econometric models. *Review of Economic Studies* 61, 291-303.
- Duan, N. et al., 1983. A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business & Economic Statistics*, 1, 2, 115-126.
- Duan, N. et al., 1984. Choosing Between the Sample-Selection Model and the Multi-Part Model. *Journal of Business & Economic Statistics* 2, 283-289.
- Duan, N. et al., 1985. Comments on Selectivity Bias. *Advances in Health Economics and Health Services Research* 6, 19-24.
- Fraga, M. & O. Martins (2001), Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal, *Journal of Applied Econometrics* 16, 23-39.
- Gerfin, M. (1996), Parametric and Semi-parametric Estimation of the Binary Response Model of Labor Market Participation, *Journal of Applied Econometric* 11, 321-39.
- Gorgens, T. (2000), Semiparametric Estimation of Single-Index Transition Intensities, *Econometric Society World Congress 2000 Contributed Papers* 0596, Econometric Society.

Gorgens, T. & J. L. Horowitz (1999), Semiparametric Estimation of a Censored Regression Model with an Unknown Transformation of the Dependent Variable, *Journal of Econometrics* **90**, 155-191.

Hadley, J., Holahan J., 2003. Covering the Uninsured: How Much Would It Cost? Health Affairs, 2003.

Hardle, W. & E. Mammen (1993), Comparing nonparametric versus parametric regression fits. *Annals of Statistics* **21(4)**, 1926-1947.

Hardle, W., E. Mammen & M. Muller (1998), Testing parametric versus semiparametric modelling in generalized linear models. *Journal of American Statistical Association* **93**, 1461-1474.

Hardle, W., V. Sponkoiny & S. Sperlich (1997), Semiparametric single index versus fixed link function modelling. *Annals of Statistics* **25**, 212-243.

Hausman J. A. et al., 1998. Misclassification in the dependent variable in a discrete response setting. *Journal of Econometrics* 87, 239-269.

Heckman, J. J., 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic Social Measurement* 5, 4, 475-492.

Heckman, J. J., 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47, 1, 53-161.

Hoeffding, H. (1963), Probability Inequalities for Sums of Bounded Random Variables, *Journal of the American Statistical Association* **48**, 13-30.

Holly, A. et al., 2002. Hospital services utilization in Switzerland: The role of supplementary insurance. Institute of Health Economics and Management, University of Lausanne, manuscript.

Honore, B. E. & J. L. Powell, Pairwise Difference Estimation of Nonlinear Models, in *D. W. K. Andrews and J. H. Stock, eds., Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg* (Cambridge: Cambridge University Press, 2005), 520-53.

Horowitz, J. L. & V. G. Sponkoiny (2001), An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative, *Econometrica* **69**, 599-631.

Horowitz, J. L. & W. Hardle (1994), Testing a parametric model against a semiparametric alternative. *Econometric Theory* **10**, 821-848.

Ichimura, H. (1993), Semiparametric least squares(SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics* **58**, 71-120.

The Kaiser Family Foundation, Kaiser Fast Facts. Health Insurance Coverage in America, 2006.

Klabunde C. N. et al., 2000. Development of a comorbidity index using physician claims data. *Journal of Clinical Epidemiology* 53, 1258-1267.

Klein, R. W. (1993), Specification tests for binary choice models based on index quantiles, *Journal of Econometrics* **59**, 343-375.

Klein R. W. and Shen C., 2008. Bias Corrections in Testing and Estimating Semiparametric, Single Index Models. unpublished manuscript.

Klein, R. W. and Spady R. H., 1993. An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica* **61**, 387-421.

Klein R. W., Shen C., and Vella F., 2009. Joint Binary Selection and Treatment Models. unpublished manuscript.

Klein, R. W. & F. Vella (2007), Estimating a class of triangular simultaneous equations models without exclusion restrictions, manuscript.

Lee, L., 1982. Some approaches to the correction of selectivity bias. *Review of Economic Studies* **49**(3), 355-372.

Lindau S.T. et al., 2006. Health literacy as a predictor of follow-up after an abnormal Pap smear: a prospective study. *Journal of General Internal Medicine* **21**(8), 829-834.

Maddala, G. S., 1985. A Survey of the Literature on Selectivity Bias as it Pertains to Health Care Markets. *Advances in Health Economics and Health Services Research* **6**,3-18.

Manning, W. G. et al., 1987. Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review* **77**, 251-277.

Miller E., Banthin, J. S., and Moeller, J. F., 2003. Covering the Uninsured: Estimates of the Impact on Total Health Expenditures for 2002. Agency for Healthcare Research and Quality.

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and two-part model in health econometrics. *Journal of Health Economics* **17**, 247-281.

Newey, W. K. (1985), Maximum likelihood specification testing and conditional moment tests, *Econometrica* **53**, 1047-1070. Newey, W. K., F. Hsieh & J. Robins (2004), Twicing Kernels and a Small Bias Property of Semiparametric Estimators, *Econometrica* **72**, 947-962.

Newhouse, J. and the Insurance Experiment Group, 1993. Free for all? Lessons from the RAND Health Insurance Experiment, Harvard University Press, Cambridge.

Pakes, A. & D. Pollard, (1989), Simulation and the asymptotics of optimization estimators, *Econometrica* **57**, 1027-1058.

Powell, J. L. , J. H. Stock, and T. M. Stoker, 1989. Semiparametric Estimation of Weighted Average Derivatives. *Econometrica* **57**, 1403-1430.

Puhani, P. A., 2000. The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys* **14**(1),53-68.

RAND Health Insurance Experiment [in Metropolitan and Non-Metropolitan Areas of the United States], 1974-1982.

Robinson, P.M., 1988. Root-n-consistent semiparametric regression. *Econometrica* 56, 931-954.

Rothschild, M. and Stiglitz, J. E., 1976. Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *The Quarterly Journal of Economics* 90(4), 630-49.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Steinvil A. et al., 2008. Relation of educational level to inflammation-sensitive biomarker level. *American Journal of Cardiology* 102(8),1034-9.

Vera-Hernandez, M., 1999. Duplicate coverage and demand for healthcare. The case of Catalonia. *Health Economics* 8, 579-598.

Wooldridge, J. M., 2002. *Econometrics of cross section and panel data*, MIT Press, Cambridge.

Vita

Chan Shen

- 2009 Ph.D in Economics, Rutgers University, New Brunswick, New Jersey
- 2007 M.S. in Statistics, Rutgers University, New Brunswick, New Jersey
- 2004 M.A in Economics, Rutgers University, New Brunswick, New Jersey
- 2002 B.A in Economics, Fudan University, Shanghai, China