NONPARAMETRIC REGRESSION WITH A PARAMETRIC START FOR BINARY RESPONSE MODELS

by

Joshua Makori Ariga

A Thesis Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

in the Graduate College

THE UNIVERSITY OF ARIZONA

2001

ACKNOWLEDGEMENTS

I would like to express my thanks to Dr. Alan P. Ker for his guidance and patience in the course of this work. The completion of this project has benefited from his continued support and encouragement to undertake the subject matter. I have richly benefited from this experience.

I extend my thanks also to my committee members, Dr. Satheesh Aradhyula and Dr. Gary Thompson, for their useful comments and suggestions that helped transform a bare-borne draft to a wholesome thesis. Of course the support and patience of my family - Dinah, Luther, and Yvonne - made the program and thesis a possibility.

Contents

List of Figures

List of Tables

ABSTRACT

In the last few decades the econometric literature has generated a number of parametric, nonparametric, and semiparametric methods for solving binary response models. These procedures differ in the relative strength of their assumptions, convergence rates, consistency, efficiency, and prediction. This paper considers an estimator that assumes a parametric start and then couples this with a nonparametric correction factor. This model is compared with parametric estimators of the link function on a Monte Carlo study and an application using data from Spector and Mazzeo(1980) on effects of new teaching methods in economics on student grades.

Chapter 1

INTRODUCTION

1.1 Problem Statement

Qualitative response models have received a lot of attention in the econometric literature because they continue to have considerable application in studies involving participation in the labor force, choice of occupation, choice of transportation mode, purchase of consumer durables, analyzing voter behavior, and other areas. The last two decades have witnessed the application of semiparametric and nonparametric methods developed in the 1960s and 1970s to qualitative response models. Great interest was generated by Horowitz's (1993) comparison of the performance of several alternative parametric and semiparametric specifications when applied to the problem of solving the binary decision of selecting the mode of transportation to work. Horowitz(1993) regressed an individual's choice of transportation method (automobile versus transit) against difference in travel costs and travel times in and out-of-vehicle and number of cars owned by the household. Horowitz applied a number of parametric and semiparametric procedures to estimate the link function and

slope parameters. Some of these models were fitted by using the Klein and Spady (1993) procedure, maximum score method by Manski(1975) and Horowitz's(1992) smoothed maximum score methods[1].

A number of methods have been developed by econometricians to get solutions to binary response models. These methods can be broadly classified as parametric, nonparametric, and semiparametric in their approach to tackling the regression estimation of binary models. These methods differ in the relative strength of the assumptions and restrictions they impose on structural form in order to estimate the relevant regression parameters and function.

Economic theory does not reveal the structural form taken by variables but instead may shed some light on the nature of the relationships between certain variables - homogeneity, additivity, etc. Parametric approaches impose certain restrictions and assumptions to aid in identifying the structural forms prior to solving regression problems. Now, if the true structure is not from the assumed parametric family, then the model is misspecified. This implies that the regression estimates may not be very useful in explaining the data and practical applications like analyzing consumer behavior might not be reliable. Therefore, issues involving specification and selection of functional form are important to the quality of the results from binary response model estimation.

---

[1]Klein and Spady(1993) replaced the link function F() (Equation 1.2) with a nonparametric kernel and estimated the parameters by a semiparametric maximum likelihood method. The score methods of Manski(1975) and Horowitz(1992) apply the concept of getting a method that maximizes the number of correct predictions

1.2 Functional Forms

The estimation of the conditional mean regression function, $y \mid x$, is important in econometrics, as it is the most common method used to explain the relationship between variables. A number of regression estimators exist for solving binary response models of the general conditional estimation form

$$y_i = I(x_i\beta + u_i \geq 0) \tag{1.1}$$

where y_i are the decision choices available, x_i are the regressors, $I(.)$ is an indicator function, $x_i^{'}\beta$ is an index, and $(x_i\beta, u_i)$ are independent and identically distributed (Manski 1975). The i.i.d condition implies that $x_i\beta$ and u_i are not correlated and are randomly selected variables. The nature and application of index functions to binary regression is explained in Chapter 2 and 3 of this thesis. The expression in brackets (.) is the latent variable which is generally unobservable and provides the values taken by the endogenous variable y (Section 2.1). Equation 1.1 can be transformed to $y_i = F(x_i\beta) + u_i$ where F() is the distribution function of the unobservable u i.e. the link function between the latent variable and the endogenous variable Y. F() is usually unknown but when it is assumed to be the uniform distribution the expression becomes $y_i = x_i\beta + u_i$.

The condition $E(u \mid x) = 0$ helps map this equation to its probability form where Y takes a value of 1 conditional on X taking particular values, x, as shown below.

$$Pr(Y_i = I \mid X = x) = F(x_i\beta) \tag{1.2}$$

where $F(.)$ is an unknown function, $x_i\beta$ and is an index (Horowitz 1992a). The transformation from Equation (1.1) to (1.2) includes location normalization $E(u_i \mid$

$x_i\beta) = 0$ analogous to the parametric assumption of linear least squares regression $E(u \mid x) = 0$. Identification and prior restrictions on binary response models are discussed further in Section 2.1. Applications of this model include Horowitz (1993), Das (1991), and Melenberg and van Soest (1991). Horowitz applies this model to the choice of mode of travel to work while Melenberg and Soest look at the decision to rent or own a house subject to certain household characteristics (number of children, income, etc). Thompson and Kidwell (1998) apply the indirect utility approach to model consumer preference between organic and conventional produce and the choice between two stores conditional on certain variables (incomes, demographics, etc) using parametric methods. Their model incorporates simultaneity issues involving the two endogenous variables - type of store and produce[2]3.

The problem is to estimate both β and $F(.)$ from observations of (Y, X). This model nests some parametric specific forms. If $F(.)$ is the identity function the expression becomes a linear model. If $F(.)$ is the cumulative distribution function (cdf) of the normal or logistic distribution then the expression is the binary probit or logit model respectively (Greene 2000). When $F(.)$ is assumed to be unknown, as is the case with nonparametric regression, this becomes flexible and less restrictive than a parametric model while keeping the desirable features of parametric models. The estimation method and results depend on the a priori assumptions on the population or data generating process.

By far the most commonly applied regression methods for binary response models are the probit and logit. These approaches rely on prior assumptions that

[2]Thompson and Kidwell use parametric procedures to estimate coefficients in a model that regresses choice of produce or store against covariates involving demographics and income levels.

the link function $F(.)$ belongs to a specific family of parametric functions. This then makes it convenient for the relevant parameters to be estimated by common parametric estimators.

Nonparametric regression methods do not assume the functional form of the distribution of u_i but replace $F(.)$ with a kernel estimator which has appropriate or desirable properties. When the parametric structural assumption is correct, the estimates for β converge at a comparatively quicker rate of $n^{1/2}$ [3] than those of the nonparametric procedures. However, when the assumption is incorrect, parametric regression estimates may not converge to their true values. On the other hand, semiparametric methods combine parametric and nonparametric procedures with potential efficiency gains from both methodologies while minimizing their disadvantages (Ker and Coble 2001).

1.3 Research Objectives

Given the abundance of qualitative data in most surveys - decision to use certain technologies or not, insure crops or make major investment decisions, etc. - it is important to devise ways of estimating relevant parameters efficiently. The decision on whether to join the labor force, make a major investment, vote for a particular party or particular legislation, consume organic produce, purchase crop insurance, etc against the respective negative alternatives are all examples of binary response variables. For instance, a typical binary response model regresses the probability of joining the labor force against covariates like the number of children one has and income level of spouse or other benefactor.

[3] Section 3.1.1 and Ker and Coble (2001) expound on the concept of convergence.

A number of estimators are available to solve such models - probit, logit, linear probability model, and a host of nonparametric and semiparametric approaches. Ker and Coble(2001) have developed an estimator of the link function that incorporates some properties of the Hjort and Glad(1995) density estimator and Glad(1998) regression estimator that has desirable qualities for the solution of binary response models (Section 4.3). This estimator is an extension of the Hjort and Glad (1995) density estimator with a parametric start to the problem of qualitative response data in a nonparametric framework. The objective of this paper is to investigate this estimator's performance when compared to some of the existing alternatives. This functional form can be tested on its own merits and compared directly to the kernel regression estimator of Nadaraya (1964) and Watson (1964) and parametric models. The estimation process allows for consistency and flexibility particularly due to the weak restrictions and assumptions that accompany the model.

1.4 Organization of the Study

This paper will proceed as follows. Section 2.1 and 2.2 explain how binary response models are set up and how they are identified and estimated. Identification here refers to the conditions that should exist before $F(.)$ and β are estimated. Then the other Sections of Chapter 2 will embark on parametric estimation methods, their assumptions, performance, and problems emanating from their application to binary models. Chapter 3 will dwell on the relevant nonparametric methods applicable to binary data and the techniques for estimating the smoothing or bandwidth parameter that is crucial in estimation. Chapter 4 follows with a similar exposition on semiparametric or combined methods. In particular Section 4.3 and 4.3.1 explain how

Ker and Coble(2001) estimator is derived and some of its properties. Chapter 5 will deal with simulations, results and their interpretation, exposition on the application of the model, conclusions, summary, and recommendations for future research.

Chapter 2

PARAMETRIC METHODS

2.1 Binary Response Models

A typical binary choice model postulates that for each randomly selected individual i an observed choice variable y_i is related to $K \times 1$ vector of observable regressors or exogenous variables x_i by the expression $\boldsymbol{y_i = I(x_i\beta + u_i \geq 0)}$ as in Equation (1.1) where β ($K \times 1$ vector) is an unknown parameter, and u_i is an unobserved variable. The indicator function $I(.)$ is assigned 1 if the expression in brackets is satisfied and 0 otherwise (Ichimura and Scott 1998). For this model the threshold value k is 0, below and above which y takes certain values depending on the values taken by the latent variable within the brackets. If the latent variable $x_i\beta + u_i \geq 0$ then $y_i = 1$ and if $x_i\beta + u_i \geq 0$ then $y_i = 0$.

This is applicable to discrete choice models where y_i denotes whether a consumer i buys a certain good or service or in dealing with experimental willingness-to-pay models. For instance, economic theory postulates that the decision to purchase or not to purchase a car depends on its price. The assumption is that the consumer has a reservation price (which is unobservable, but probably determinable by conducting an experimental design). For instance if the market price falls below or is equal to

the reservation price then the decision to buy is made. However, if the market price is higher than the reservation price then the consumer decides to forego the investment. For this case the latent variable is the difference between the reservation and market price, $d = Pr - Pm$. If $d \geq 0$ then $y_i = 1$ and the consumer invests in a car. However, if $d < 0$ then $y_i = 0$ and no purchase is made. Thus y indicates the decision choices available to the consumer ($y_i = 1$ means he buys because $d \geq 0$ and $y_i = 0$ means no purchase because $d < 0$). The error u_i can be interpreted as unexplained unobservable effects (aside from price). The indicator function $I(.)$ thus assigns the new observable variable $y = 1$ if the latent variable is positive and $y = 0$ otherwise.

The model is also applicable in indirect utility studies to indicate that individual i gets an indirect utility mu_{ij} when alternative j is chosen, and therefore chooses $y = 1$ iff $mu_{i1} \geq mu_{i0}$. Then β is a vector of the marginal (indirect) utilities associated with the observable variables x_i (e.g. prices) and u_i denotes all other unobserved effects of indirect utility.

Heterogeneity (randomness in slope effects from observation to observation) can be incorporated into the model by treating β as a random variable (for more details readers are directed to Quandt (1956), Horowitz (1993), and McFadden (1976)). For this study β in Equation (1.1) is assumed to be constant across all observations. In addition, $F(.)$ denotes the cumulative distribution function of u_i. A number of estimators have been considered for this model including parametric, nonparametric, and semiparametric specifications discussed in the following sections. It is assumed that (x_i, y_i) are i.i.d and y_i is related to x_i through the index $x_i\beta$. The model satisfies the index restriction; $E(y \mid x) = E(y \mid x\beta)$ where E is the conditional expectation and $x\beta$ is the index or aggregator (Klein and Spady 1993) leading to the term single

index model for Equation (1.1), which is a useful variation of the linear regression formulation.

The index restriction implies that x is related to y only through the index $x_i\beta$. This, as will be elaborated in Section 2.1 and Chapter 3, is essential in alleviating the problem associated with large dimensions of X by aggregating the components of X into a single index. Horowitz (1993) and Powell, Stock, and Stoker (1989) provide a number of ways to estimate index coefficients. Powell, Stock, and Stoker (1989) applies density-weighted average derivative technique to estimate coefficients of index models. Their estimator is of the form $\boldsymbol{\delta} \equiv \boldsymbol{E}\left(\boldsymbol{f}(\boldsymbol{x})\frac{\boldsymbol{dg}}{\boldsymbol{dx}}\right) \equiv \boldsymbol{E}\left(\boldsymbol{f}(\boldsymbol{x})\frac{\boldsymbol{dG}}{\boldsymbol{d}(\boldsymbol{x_i\beta})}\right)\boldsymbol{\beta} = \boldsymbol{\gamma\beta}$ where $E(y \mid x) = G(x_i\beta)$ is the index regression function and $f(x)$ is the density function of x. Thus the weighted average of the derivative dg/dx is proportional to β . The derivatives of the density are then estimated by using kernels. Horowitz (1992) uses both nonparametric and semiparametric methods to estimate index model coefficients by utilizing quasi-maximum likelihood and maximum score methods.

The problem is to solve the model for $F(.)$ and β i.e. estimate both $F(.)$ and β. There are basically two approaches to estimating such regression models with the third procedure incorporating techniques borrowed from these two. If β (and therefore $x_i\beta$) is known then the conditional expectation of Y on X can be estimated by nonparametric methods by replacing $F(.)$ with an appropriate kernel estimator (Horowitz 1993). For the case where both F and β are unknown a similar nonparametric approach can be applied and β is replaced by its estimator by a process that minimizes the least squares cross validation function of Section 3.2.2 or other suitable technique. Alternatively, $F(.)$ can be assumed known a priori then the parameters to Equation (1.1) solved by applying parametric nonlinear techniques. Semipara-

metric methods combine these two approaches in a process that seeks synergism by gaining from the advantages of either of its parts while avoiding or mitigating their disadvantages.

2.2 Restrictions and Identification of models

Manski (1988) and Horowitz (1998) provide some criteria for the identification of binary response models. An observable binary variable y is assumed to be determined by an observable variable x and an unobservable random variable u. The pair (x, u) act on y through a linear function $x_i\beta + u_i$(the latent variable), where β is a k-dimensional parameter in Eq (1.1), $I(.)$ is an indicator function that takes the value 1 if the logical condition inside the brackets is satisfied and 0 otherwise as motivated in Section 2.1.

In economics, y may indicate the utility maximizing decision maker's observable choice between two alternatives (Manski 1988). If the realizations (y, x) are observable and drawn randomly from the population, then the set of conditional response probabilities can be expressed as $P_{1|x} = [Pr(y = 1 \mid x), x \in X]$. For each $x \in X$, the probability of the event $y = 1$ is $P_{1|x} = \int I[x\beta + u \geq 0]dF_{u|x}$ where $F_{u|x}$ is the distribution of u conditional on x. The response probabilities $P_{1|x}$ may be estimated by nonparametric regression of y on x if $F_{u|x}$ is replaced by a suitable kernel estimator of $F(.)$ (Section 3.2). Alternatively, the regression function and parameter estimates can be estimated by parametric nonlinear optimization methods if the distribution of u, $F_{u|x}$, is assumed. Manski (1988) suggests that the best predictor of y given x is always some function of the response probability $P_{1|x}$ like $I[P_{1|x} \geq .5]$ i.e. if the indicator function $I(.)$ is satisfied then $y = 1$ or 0 otherwise.

The problem of identification depends on whether $(\beta, F_{u|x})$ are identifiable or estimable. The probit and logit methods of Section 2.2.2 rely on two basic conditions for identification. First, for each $x \in X$, $F_{u|x}$ is a known, continuous, strictly increasing (monotonic) distribution function with median 0, and secondly, there is no proper linear subspace of $\Re^{\kappa}$ having probability 1 under F_x (where F_x is the probability distribution of x). This second condition simply means that there is no multicollinearity between X components i.e. there is no linear relationships between the Xs or equivalently X is a full rank matrix. Monotonicity is important in transforming the probabilities to their estimates back and forth thus ensuring a one-to-one mapping of probabilities and their estimates (Section 2.3.1 on LPM [4]).

Manski(1988) provides further identifying conditions which require technical proofs and readers are referred to Manski for these details. However, two other assumptions that are prominent in the literature are useful for identification of $(\beta, F_{u|x})$ in this study. The assumption of mean independence or what is also called location normalization,$E(u \mid x) = 0 \ \forall x \in X$, and the assumption by Horowitz (1998) on scale normalization that sets one of the coefficients of the β vector associated with continuous x to 1 i.e. $\beta_i = 1$ are applied in this study.

Mean independence, a common application by econometricians, implies that the residuals are independent and uncorrelated with expectation of zero given X - this does not imply imposition of normal distribution unless the model explicitly assumes this (like for the probit). Horowitz's restriction above is akin to setting the variance of u at a constant like in the classical linear regression, a form of scale normalization that maintains homoscedastic conditions.

[4]Linear Probability Model

2.3 Parametric Approaches

Parametric regression estimation imposes or assumes a distributional structure for the unobservable terms in Equation (1.1). Some of the models that make prior assumptions about the link function are discussed below. The parameter estimates in this case converge to their true values at the rate[5] of $n^{1/2}$. These methods for solving qualitative response models can be broadly categorized into linear-probabilistic and nonlinear models. The former is epitomized by the linear probability model (LPM) and the latter by the probit and logit models.

Some of the techniques applied in solving nonlinear models include Generalized Method of Moments (GMM) and nonlinear least squares or maximum likelihood estimation (MLE). The LPM, probit, and logit models are examples of parametric regression methods specifically suited to qualitative response models. Ker and Coble(2001) give a lucid discussion of the efficiency of the parametric and nonparametric techniques applied in crop yield density estimation in their article "Modeling yields"[6] . Scott(1992) also provides a comprehensive exposition of nonparametric kernel density estimators.

Parametric methods, aside from making the assumption that $F(.)$ is known a priori, also make certain necessary assumptions and or restrictions for solving the model. It is a common practice in econometrics to apply some scale and location normalization in order to identify β. For parametric models, such as the probit and logit models, scale normalization is executed by setting the variance of u to one and

[5]Section 3.1.1 and Ker and Coble (2001) expound on the concept of convergence.

[6]In estimating crop yield densities Ker and Coble(2001) contrast the results from Olkin and Spiegelman(1987), Hjort and Glad(1995), and Hjort and Jones (19)estimators

$(\frac{\pi}{3})^2$ for the probit and logit models respectively. Location normalization is usually set by assuming that the conditional expectation of u given x is zero,$E(u \mid x) = 0$. These are familiar concepts from ordinary least squares regression estimation.

2.3.1 The Linear Probability Model

The linear-probability model (LPM) assumes a linear relation between the probability of y taking on a value of 1 given the index $x_i\beta$. If vector X represents the regressors, a simple linear regression model in this case is given by $y_i = x_i\beta + u_i$ where $u \sim N(0, \sigma_u^2)$. If u_i and u_j are independent for $i \neq j$ (no heteroscedasticity) and X is a vector of random variables such that X and u are independent (othorgonality condition equivalent to mean independence) then we know that $E(u_i \mid x) = 0$ and $E(Y_i) = x_i\beta$.

Let the probability that the Y equal one given the explanatory variables be given as $Pr(Y_i = 1 \mid X = x) = \pi_i$. Thus $Pr(Y_i = 0 \mid X = x) = 1 - \pi_i$ and from probability theory the expectation of Y is denoted by $E(Y_i) = \sum_{i=0,1} Pr(Y_i)Y_i = \pi_i(1) + (1 - \pi_i)(0) = \pi_i$. Thus the model implies a linear probability model of the form $\pi_i = x_i\beta$.

The LPM combines linearity with probability theory to model qualitative response data $Pr(Y_i = 1 \mid X = x) = \pi_i = F(x_i\beta)$ (Greene 2000), where π_i is the probability of the dependent variable y taking on values of 0/1 given that the explanatory variables X take on specific values x. This expression transforms to $\pi_i = x_i\beta$ because the assumption of linearity implies $F(.)$ is a uniform distribution.

The marginal responses of the probabilities to changes in independent variables are depicted by parameters β. The "probabilities" from LPM regression are not

confined to (0,1). When Y_i is binary, the error term is dichotomous as well and lacks normality; If $Y_i = 1$ then $u_i = 1 - \pi_i$ with probability π_i. If $Y_i = 0$ then $u_i = -\pi_i$ with probability $1 - \pi_i$ and $V(u_i) = \pi_i(1 - \pi_i)$. It is clear that v is variable across observations and can take negative values because the fitted values from OLS regression are not constrained to the (0,1) interval.

Weighted least squares (WLS) or feasible generalized least squares (FGLS) can be used to correct for heteroscedasticity. Another solution to this problem is to have a constrained LPM where π_i is contained within the unit interval while keeping the linearity between π_i and X within the interval.

However, this constrained model can generate unstable estimates especially at the boundaries and it is hard to fit data due to abrupt slope changes at $\pi_i = 0$ and $\pi_i = 1$. A number of studies show that the LS estimates of the LPM frequently give results similar to other formally justifiable methods like the logistic model described below(Greene 2000).

It can be inferred from this discussion that a model that gives results consistent with the underlying theory will be appropriate in this case. In principle, any continuous probability distribution defined over the real line will suffice. This explains why most practical applications replace the LPM model with either the probit or logit model with link function F (.) that maps all $x_i\beta$ to [0,1]. Figure 5.2 shows a graphical comparison of the probability estimates from the LPM and probit regressions from the same dataset.

2.3.2 The Logit and Probit Model

To correct the problem of probabilities, π_i , lying outside [0,1] we need a positively monotonic function that transforms $x_i\beta$ to the unit interval. It is useful, a priori, to have a *cdf* that is smooth and which approaches $\pi_i = 0$ and $\pi_i = 1$ asymptotically. It is advantageous if the *cdf* is strictly increasing so that $F^{-1}(.)$ transforms the probabilities to their corresponding regression estimates of Y_i, in a one-to-one relationship.

Any cumulative distribution function (in this case the logistic or normal distribution) meets this requirement. If Y is a dichotomous dependent variable taking values of one and zero with probability π_i and $1 - \pi_i$ respectively, then for the i^{th} observation, the probability can be depicted as $Pr(Y_i = 1 \mid x) = \pi_i$ where $\pi_i = F(x_i\beta)$ and $X_{n\times k}$ is a full column rank matrix and $\beta_{k\times 1}$ is a vector of parameters relating π to $x\beta$ which could be nonlinear. $F(.)$ is the normal or logistic distribution that transforms the estimates to the (0,1) interval. When the normal distribution is assumed the cdf is expressed as $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp^{-\frac{1}{2}z^2}\, dz$ and thus Equation (1.1) becomes

$$\pi_i = \Phi(x_i\beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp^{-\frac{1}{2}z^2}\, dz \tag{2.1}$$

where $\Phi(z)$ is the probit model and $z = x\beta$. $F(.)$ can also be the standardized logistic distribution $\Lambda(z) = \frac{1}{1+\exp^{-}z}$ if the $u_i's$ are distributed as logistic density function (pdf), with

$$\pi_i = \Lambda(x_i\beta) = \frac{1}{1 + \exp^{-}z} \tag{2.2}$$

giving the probabilities where $\Lambda(z)$ is the logistic or logit model. Both models can be conveniently expressed as in Equation(1.1).

For the standard normal cdf the standard deviation is set at $\sigma = 1$. When σ is replaced by the square root of the covariance-variance matrix of x i.e.$V(x) = x'\Sigma x$ then Equations (2.1) and (2.2) can accommodate heterogeneity or random-coefficient variation as dealt in Fischer and Nagin (1981) and Hausman and Wise (1978) in travel demand models.

Estimation of these models is straightforward using maximum likelihood methods to estimate the parameters; see Greene (2000), Quandt (1956), and McFadden (1976). The parameter estimates maximize the log of the product of the joint *pdf's* of multivariate functions from a known *dgp*. For a random sample of size n, ML estimator is defined by maximizing the log-likelihood function that corresponds to Equation (1.1)

$$\ln L_n(\beta) = \Sigma_{i=1}^{n}[y_i \ \ln F(x_i\beta) + (1 - y_i) \ \ln(1 - F(x_i\beta))] \tag{2.3}$$

where the error term u is assumed to have distribution $F(.)$ given x. The maximum likelihood estimates are obtained by maximizing Eq (2.3) over any unknown parameters (regression and distributional coefficients) corresponding to the *cdf* with appropriate normalization restrictions.

In some cases, the probit and logit models fit the data better by avoiding the assumption of linearity (LPM) thus revealing unique data behavior across the support (Greene 2000) as demonstrated by the graph in Figure 5.2. The logit's interpretation of the inverse transformation $F^{-1}(\pi_i)$ as log odds as a useful connotation related to prediction. Rearranging, and taking logs of the above logit model gives an interesting result; $Log\left(\frac{\pi_i}{1-\pi_i}\right) = F^{-1}(\pi_i) = x_i\beta$. This is the logit of π or the log of the odds that Y is one rather than zero. If the odds are even ($\pi_i = .5$) then the logit is zero i.e the

$E(Y \mid X = x) = 0$. If $\pi_i < .5$ the logit is negative and if $\pi_i > .5$ the logit is positive.

For the binary choice model, the probit and logit estimation methods are the most commonly applied and these restrict the error distribution to parametric families. If the parametric specification is incorrect then inconsistent estimates may result from the likelihood-based approaches. As Chen (2000) clearly enunciates

> .....specific functional forms for the error distribution cannot usually be justified by economic theory[7]

. However, when the assumed functional form is correct, convergence is attained at the parametric rate.

As a result of a priori assumptions about the dgp, regressions based on parametric methods are not necessarily always correct. If the assumptions made are not correct then the estimates may not converge to their true values. Policy decisions based on such results may not have the desired impact. However, if the initial assumptions are correct then parametric regression estimation of β may converge at a quicker rate to the true parameter values.

---

[7]Chen(2000) proposes a semiparametric MLE for both intercept and slope parameters under symmetry and index restrictions.

Chapter 3

NONPARAMETRIC METHODS

3.1 Introduction

Economic theory is often abstract without revealing any clear functional relationships and never implying or pointing to a specific data generating process. Parametric regression approaches rely on assumptions to help in the selection of the link function from the family of known parametric functions. Once this is identified, maximum likelihood estimators are consistent in estimating the unknown parameters of the likelihood function in Equation (2.3). The estimators are asymptotically normal with the inverse of the covariance matrix satisfying some efficiency lower bound criterion (Hardle and Linton 1994). This means that the inverse of the matrix of second derivatives (Hessian) of the nonlinear function that is to be optimized meets some Cramer-Rao-like measure of efficiency for the estimates. If the parametric model is not true, MLE estimates may not be efficient. Therefore, the restrictions imposed by a parametric approach can sometimes bear a cost.

It is not surprising that a lot of literature is now delving into different procedures that do not rely on strong prior restrictions. Nonparametric methods do not impose parametric restrictions on functional forms. Nonparametric smoothing tech-

niques have been used substantially in demand, food expenditures, and total income analysis for household surveys (Deaton 1991). These procedures are useful in displaying intricate features of a dataset (Greene 2000) and also as a tool of comparison against estimated parametric models. Some of the nonparametric methods include kernel estimators and related local smoothing methods, or series estimators such as truncated polynomials or spline methods. The discussion on nonparametric methods in this paper will concentrate on kernel estimators.

Nonparametric smoothing estimators offer robustness under a wider class of structural functional forms but are characterized by a slower convergence process dependent on the dimension of X. The smoothing parameter(Section 3.2.1) decreases so as to reduce bias to zero as sample size increases. A measure of bias is the squared difference between the expected value of the estimate (mean estimate) and its actual value.

Nonparametric regression usually makes modest assumptions about the shape of the regression function - some degree of functional differentiability, symmetry around zero $[k(u) = k(-u)]$, and integrability to one of the kernel $[\int k(u)du = 1]$ (Yatchew 1998). With nonparametric regression, inferences and estimators are less dependent on functional form or dgp assumptions. It is also important that nonparametric techniques can be useful in bolstering parametric procedures. If doubts exist about the parametric functional form, specification checks against nonparametric alternatives can provide comparison criteria and as will be discussed later, a joint parametric-nonparametric approach may offer additional efficiency.

3.1.1 Nonparametric Convergence Rates

Econometric literature on nonparametric regression estimation discusses various methods that can improve convergence rates. It is generally agreed that as the dimension of X increases, larger datasets are needed for efficient estimation. As the dimension of the X matrix becomes larger there may be need for an impracticably large sample in order to improve estimation precision (Stone 1980). But a number of techniques have been applied to mitigate this phenomenon, aptly called "the curse of dimensionality". These methods are discussed in the following sections.

3.1.2 Dimensionality and Single Index Models

Local nonparametric regression averagers, because of their flexible functional forms, estimate the parameters by a process that averages across neighboring data points. This leads to larger and larger data sets being required for the estimates to converge to their true values, what is termed the curse of dimensionality. The rate of convergence for density estimators is provided by Stone (1980, 1982) as $0_p\left(\frac{1}{N^{2m/(2m+d)}}\right)$ where d is the dimension of X, m is the smoothness or differentiability of the function, and N is the number of realizations. Hardle and Linton (1994) also indicate that for nonparametric regression estimation this depends on the dimension of X. This is the rate at which a consistent estimator of β i.e.$\hat{\beta}$ converges to the true value.

The single index model (SIM) structure avoids the curse of dimensionality that is associated with SIM-less nonparametric methods. What the SIM model structure does is aggregate the X components to a single index $x\beta$ and then the regression of Y on X (multivariate) becomes a simpler univariate regression of Y on $x\beta$.

The simple SIM structure can be extended to a multiple-index format $E(Y \mid$

$X) = x_0\beta_0 + G(x_1\beta_1.....x_m\beta_m)$ that splits the model into a number of indexes. This then becomes an additive index model composed of linear and nonlinear parts. The rate of convergence for parametric regression is not dependent on the number of explanatory variables and the variance of the estimator goes to zero at a rate of $n^{1/2}$ irrespective of the number of explanatory variables.

3.1.3 Partially Linear and Additive Models

To avoid the slow convergence associated with the curse of dimensionality, restricted models can be used - nonparametric additive models of Hastie and Tibshirani (1990), or semiparametric models like the partially linear and index models. Partially linear models proposed by Ai and McFadden (1997) where x_1 and x_2 are non-overlapping divide the model into linear and unknown functional forms $E(Y \mid x_1, x_2) = x_1\beta_1 + g(x_2)$. For a partial linear regression model on the unit square of the form $z\beta + f(x) + u$ where f is unknown, f can be estimated at a rate equivalent to that of nonparametric function of one variable while $z\beta$ is estimated by parametric techniques.

The curse of dimensionality is thus circumvented by use of partially linear or additive models of the form $f(z, x_1, x_2, x_3) = z\beta + \Sigma_1^3 \ f_i(x_i)$, where $z\beta$ is linear and $f_i(x_i)$are additive nonparametric sub-models. The overall rate of convergence depends on the nonparametric part of the model. For such hybrid regression function, the parametric part of the model converges to the true model at a rate of $n^{1/2}$. But the overall rate of convergence for the full model will depend on the nonparametric part of the model, which can be improved by applying the SIM model as discussed above.

Additive models, where several nonparametric sub-models are summed up but run separately, are useful in reducing the dimensionality problem. Additive models of the form $E(Y \mid X) = g_1(x_1) + g_2(x_2) + ... + g_k(x_k)$ where $k = dim(x)$ and g_i are unknown functions are useful in this regard. For example, on the unit square, f_1 and f_2 in $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$ can be approximated at a rate commensurate to that of a nonparametric function of one variable (Yatchew 1998). The rate of convergence for the additive model depends on the pure nonparametric component of the model.

3.1.4 Smoothness or Differentiability Conditions

Smoothness or differentiability of the function can also reduce approximation error. In nonparametric regression, smoothness conditions (especially bounds on derivatives) are crucial to the consistency of estimation and determination of convergence rates. Sufficient smoothness allows consistent derivatives of the estimator itself. If g is twice differentiable on the unit interval with $g^{'}$ and $g^{''}$ bounded by L and g is evaluated at equidistant values of x then it can be approximated by a Taylor series expansion $g(x_0) = g(x_t) + g^{'}(x_t)(x_0 - x_t) + \frac{1}{2}g^{''}(x^*)(x_0 - x_t)^2 \; where x^* \in [x_0 - x_t]$ If $g(x_0)$ is approximated by $g(x_t) + g^{'}(x_t)(x_0 - x_t)$ the error is $0(x_0 - x_t)^2 = 0(\frac{1}{N^2})$. $g^{'}(x_t)$ is not observable but the bound on the second derivative implies that $g^{'}(x_t) - \frac{[g(x_{t+1} - g(x_t)]}{[x_{t+1} - x_t]}$ is $0(\frac{1}{N})$ so that $g(x_0) = g(x_t) + \frac{[g(x_{t+1} - g(x_t)]}{[x_{t+1} - x_t]}(x_0 - x_t) + 0(\frac{1}{N^2})$. This is a linear estimation joining observations with straight lines (Yatchew 1998).

The curse of dimensionality is mitigated by application of semiparametric estimation (the parametric part converges at a faster rate but overall convergence rate depends on the nonparametric part of the model), using higher-order differentiable functions, and applying the principles of the single index structure.

3.2 The Nadaraya-Watson Estimator

A number of nonparametric regression estimators have been developed over the years. The kernel estimator that applies the principle of local averaging[8] (this study uses this) and nonparametric least square estimation related to spline estimation are some examples.

These estimators use the concept of local averaging that gives weights $w_i(x)$ depending on each x and is of the general form $\hat{m}(x) = \sum_1^n w_i(x)y_i$, where y is the dependent variable. Higher weights are given to observations near x and lower weights to observations further away. The simplest kernel estimator takes the form $y = f(x) + u$. To get weights,$w_i(x)$, we use a unimodal function centered at zero that declines in either direction at a rate controlled by a scale or smoothing parameter, h, whose selection criteria is the subject of discussion in the next section. The symmetrical nature of this function ensues that equal weights are assigned to equidistant neighbors on either side of a particular x.

Kernels can take various forms but should conform to certain desired properties. These functions $k(.)$ should be smooth, symmetric around zero and integrate to one i.e. $k(u) = k(-u)$ and $\int k(u)du = 1$. Smoothness conditions imply that the kernel should be differentiable to a certain degree. Thus kernels can be classified by the extent of their differentiability, with higher order kernels giving rise to higher moments. Define $u_j(k) = \int u^j k(u)du$ and $V_j(k) = \int k(u)^j du$. The order p of a kernel is the first nonzero moment; $u_j = 0,\ j = 1.....p-1\ u_p \neq 0$, implying that

[8]Others include nearest neighbor estimation as studied by Stone (1977) and others. Prakasa-Rao (1983) surveys several of these methods. Econometric application of nonparametric estimators can be found in McFadden(1976)

orders between 1 and p vanish. For the case of density estimation, smoothness helps to remove the noise that causes sudden spikes or jumps along the support.

Borrowing from statistics, we can estimate $m(.)$ in $y_i = m(x_i) + u_i$ by estimating the joint density, $f(x, y)$, and then integrating as $m(x) = \int y\ f(x, y)dy\ / f(x, y)dy$ to estimate y. We know that $\int k(u)du = 1$ so from $\hat{f}(x, y) = \frac{1}{n}\Sigma k_h(x - x_i)k_h(y - y_i)$ it can be further inferred that $\int \hat{f}(x, y)dy = \frac{1}{n}\Sigma k_h(x - x_i)$ which leads to the function depicted as $\int y\hat{f}(x, y)dy = \frac{1}{n}\Sigma k_h(x - x_i)y_i$. Define the weights as

$$w(x_i) = (nh)^{-1}k_h(.)\ /(nh)^{-1}\Sigma k_h(.) \tag{3.1}$$

where $k_h = k[(x - x_i)/h]$. The shape of the weights is determined by k and their magnitude by h, the smoothing parameter. A nonparametric regression function estimator (Nadaraya 1964; Watson 1964) then becomes

$$\hat{m}(x_i) = (nh)^{-1}\Sigma_{i=1}^{n} y_i\ k_h(.)\ /(nh)^{-1}\Sigma_{i=1}^{n} k_h(.) \tag{3.2}$$

The choice of k is less important than the selection of h over which observations are averaged and whose selection criteria is the subject of the next section. This paper will compare the Nadaraya-Watson regression estimator with the parametric and the modified Hjort and Glad (1995) estimator of Section 4.3 and 4.3.1.

A number of nonparametric kernel-type estimators exist and most of them rely on local averaging with weights at each point. Other nonparametric estimators include k-nearest neighbors (Macki 1981), local polynomial estimators (Fan and Gijbels 1992; Tibshirani 1984), and spline estimators (Ansley, Kohn, and Wong. 1993). For the bivariate case, the kernel regression estimator becomes a product of the individual kernels (Yatchew 1998). Scott (1992) explains the estimation of multivariate densities.

3.2.1 Selecting the Smoothing Parameter

The objectives of smoothing include finding the underlying structure or density and also to construct estimators from probability distributions (regression models). A decision has to be made on how many observations are included in the estimation of each point. A large window includes more observations, thereby reducing variances, but obscures subtle nonlinearity, or increases bias due to inclusion of more dissimilar observations. On the other hand, a small bandwidth or window improves detection of nonlinearity, or reduces bias, but involves fewer observations, thus increasing variance.

The optimal window usually gives the smallest value of squared bias and variance or mean integrated squared error (Powell and Stoker 1995). Powell and Stoker (1995) look at optimal bandwidth choice for density-weighted averages, a technical exposition that this paper will not dwell on. The literature on bandwidth selection is quite extensive and includes several automatic (data-driven) methods for choosing bandwidths in applications.

Most of the literature covered generally agrees that the selection of $k(.)$ is not as important as that of h, the smoothing parameter. It is very important to have a method of selecting h, the smoothing parameter. Each nonparametric regression method requires a decision on how much to smooth the data. A large h oversmooths the data and may obscure intricate data relationships (Yatchew 1998). A large h will oversmooth and increase bias and obscure important features of the regression function. A small h will track the data too closely and impair out-of-sample prediction accuracy.

There has been a lot of research covering this area by a number of authors,

some of which is the subject of the next section. Some of these techniques include least squares cross-validation, double cross-validation (Chiu 1991), plug-in methods, and double smoothing. Lee and Solo (1999) came up with what they call a PURE technique that combines non-asymptotic plug-in and Unbiased Risk Estimation techniques based on mean and weighted mean integrated squared errors (MISE and WMISE). For these and other approaches, readers are directed to the respective articles.

3.2.2 Cross-Validation

For a nonparametric regression expression of the form $y_i = m(x_i) + u_i$, the Nadaraya-Watson estimator of $m(x)$ is given by Eq (3.2) where $k(.)$ is the kernel function. The precision of this estimator depends on the bandwidth. A frequent measure of accuracy is the mean averaged squared error (MASE) i.e. how close the estimate $\hat{m}_h$ is to the true m. Cross-validation and Akaike's Information Criterion (AIC) are some of the automatic or data-driven regression smoothing processes. Hardle, Hall, and Marron (1992) show that such methods may exhibit a significant amount of across-sample variability.

The model used in this thesis applies a technique commonly referred to as Least Squares Cross-Validation (LSCV or CV). This is a global bandwidth selection method rooted on the principle of minimizing the CV function shown in Eq (3.4) below. This approach estimates $m(.)$ as

$$\hat{m}_h, (x_j) = (nh)^{-1}\Sigma_{j\neq i}^{n} y_i \; k_h(.) \; /(nh)^{-1}\Sigma_{j\neq i}^{n} k_h(.) \tag{3.3}$$

with the j^{th} observation being left out in the estimation process, as shown on the summation signs. As the sums in the numerator and denominator indicate, $\hat{m}_h$ is

estimated without the j^{th} observation, the so called leave-one-out technique with the estimate of m at each x_i using all other observations, then predicting the values of m at the omitted observation (n separate kernels estimates are needed for each h). The CV function can then be expressed as

$$CV(h) = n^{-1}\Sigma[\hat{m}_h, (x_j) - y_j]^2 \tag{3.4}$$

(Clark 1995) and is minimized with respect to h.

Then the optimal h minimizes the expression above and $cv(\hat{h})\ /cv(h_{opt}) \rightarrow 1$. The question in this selection process is how to arrive at the range of values over which the search for the optimal h will take place. Hardle and Linton (1994) suggests that the minimization of CV be taken over restricted sets of bandwidths $H_n = [n^{-(\frac{1}{5}-\varsigma)}\ n^{-(\frac{1}{5}+\varsigma)}]$ for $\varsigma > 0$. Therefore, the set of values H_n will depend on the number of observations used with the asymptotic property that $\lim h \rightarrow 0\ as\ n \rightarrow \infty$. The intuition is that as $\lim h \rightarrow 0$, $k_h \rightarrow 1$ if $x\beta - x_i\beta \geq 0$ and $k_h \rightarrow 0$ if $x\beta - x_i\beta < 0$. Thus k_h acts as an indicator function in this case. It is instructive to note that $k_h = k[(x\beta - x_i\beta)/h]$, where k is the usual kernel.

3.2.3 Plug-in Techniques

Ruppert, Sheather, and Wand (1995) developed a "direct plug-in" selection process relying on an asymptotically optimal expression akin to Silverman's(1986) shown below. The idea is to plug in estimates of σ^2 and n to obtain h. Gasser and Muller (1984) came up with "iterative plug-in" selectors while Fan and Gijbels (1995) developed selectors involving the plug-in notion. One of these plug-in selectors is the

Silverman (1986) rule expressed as

$$h = 1/364\Big[\frac{v_2(k)}{u_2^2(k)}\Big]^{\frac{1}{5}}\ \hat{\sigma} n^{-\frac{1}{5}} \tag{3.5}$$

where σ^2 is the sample variance and $v(.)$ and $u(.)$ are as described under Section 3.2 on orders of kernels . For details of plug-in techniques that apply asymptotic approximation procedures to determine optimal local bandwidths see Hardle and Marron(1985). With all these procedures the idea is to get an h that minimizes the sum of variance and bias squared i.e.$E[var + bias^2]$. For those keen on getting more details on these and other methods Abramson (1982), Jones, Linton, and Nielsen (1995), and Jones and Foster (1993) give an exposition of the bias reduction methods.

A lot of potential exists for use of nonparametric regression in parameter estimation and as a supplement to parametric regression. However, some aspects of nonparametric regression may slow this expansion. Nonparametric regression techniques are theoretically more complex than conventional parametric models but not in their practical application (Ker and Coble 2001). Secondly, nonparametric regression techniques are computationally intensive. However, nonparametric regression relies on some form of local averaging or least squares estimation, a familiar concept from parametric modeling.

In nonparametric estimation, $E(Y \mid X)$ is assumed to satisfy smoothness or differentiability conditions but no assumptions are made about its shape or the shape of its dependence on X. This does not only maximize flexibility but also minimizes the risk of misspecification error.

Chapter 4

SEMIPARAMETRIC METHODS

4.1 Introduction

As noted earlier, economic theory rarely points at the functional form of interest of the unobservable terms in Equation (1.1). Incorrect imposition of structure to Equation (1.1) may lead to inconsistent estimates of parameters of interest. Numerous econometric regression estimation problems involve unknown functions and unknown finite-dimensional parameters. Sometimes a parametric form may do well for some part of the regression function while the shape of the other portion is not quite clear or discernible a priori. In such cases, where prior knowledge exists, a semiparametric approach can be more appropriate where parametric and nonparametric techniques are combined. For a random variable Y taking 0/1 values and with X matrix of covariates, the probability of $Y = 1$ conditional on X is expressed as $P(Y = 1 \mid X = x) = F(x_i\beta)$ where F is the distribution function ,Equation (1.2).

If $F(.)$ is assumed to be known a priori, as in the case of the logit and probit, the only problem is to estimate β and this can be done by MLE. However, $F(.)$ is not known in all cases. If $F(.)$ is misspecified, MLE estimates of β may be inconsistent and predictions based on them unreliable. Many estimation problems have

unknown distribution functions that influence the relationship between observed variables (Horowitz 1991). Semiparametric methods estimate models with both unknown $F(.)$ and β. Semiparametric methods may reveal insights about the data that may not be available when using parametric estimation (Figure 5.2). Semiparametric methods for binary response data use the single index model to aggregate X. SIMs are useful in dimension reduction of X and are of the form $E(Y \mid x) = G(x_i\beta)$ where β is unknown $k \times 1$ constant and G is unknown function. The quantity $x_i\beta$ is the index. G and β are to be estimated from observations of (Y, X).

Semiparametric methods relax some of the assumptions of parametric estimation models of conditional mean functions like linear and binary probit models while maintaining the desirable features of these models. SIMs avoid the curse of dimensionality by aggregating the dimensions of X. G can then be estimated at a rate comparable to that of a single dimension observable quantity $x\beta$. β can also be estimated at the parametric rate thus equaling, in terms of rate of convergence in probability, the respective nonparametric mean regression estimation of G and parametric model for estimating β. The dimension-reduction feature of SIMs is a powerful tool in regression.

4.2 Some Semiparametric Estimators

A number of "semiparametric" methods have emerged that do not impose parametric functional forms for the error distribution in the estimation of β in the model of Equation (1.1); see Cosslett (1983), Manski(1975), Ichimura (1993), Horowitz (1998), and Klein and Spady (1993). Klein and Spady (1993) introduced estimators that attain the efficiency bounds of Cosslett(1987) and Chamberlain(1986,1992) for

the slope parameters under the independence and index restrictions.

The Klein-Spady estimator takes the form shown in Equation (1.2) where $F(.)$ is an unknown function characterized by a parametrically specified index, not necessarily a distribution function, with range [0,1]. For the Klein-Spady estimator the estimate of $F(.)$ is obtained by replacing $F(.)$ with a nonparametric kernel, without imposing distributional assumptions on the cdf of u; however it does not take into account heteroscedasticity (random effects).

Klein and Spady (1993), under the index restriction that the unobservable u depends on x only through the index $x\beta$, proposed a semiparametric ML estimator for β by maximizing

$$\ln L_{ks}(\beta) = \Sigma_{i=1}^{n} \tau_{ni}[y_i \ \ln F_n(x_i\beta) + (1 - y_i) \ \ln(1 - F_n(x_i\beta))] \tag{4.1}$$

where F_n is a nonparametric estimator for the unknown distribution of $u - \alpha_0$ and τ_{ni} are some trimming functions adopted for technical convenience[9] , and α_0 is the intercept.

By further assuming that the error term is independent of explanatory variables, they showed that the estimator achieves Cosslett's (1987) and Chamberlain's(1986, 1992) semiparametric efficiency bound. Cosslett (1987) looked at semiparametric efficiency bounds for the parameters in a binary model with the assumption that the error term is independent of the regressors (a typical econometric location normalization restriction) and a number of other location restrictions. From the semiparametric literature scale normalization takes two forms,$\mid \beta_0 \mid = 0$ or the coefficient of one component of β is set to one. Location normalization is set by assuming that the

[9]to avoid outliers and zero denominators

distribution of u conditional on x is zero.

Gozalo and Linton (2000) introduced a nonparametric regression estimator that uses prior information about the regression shape in the form of a parametric model and then nonparametrically encompasses the model. The authors applied the model to binary data on Horowitz's (1993) transport choice model dataset. Fan and Ullah (1996) developed an estimator that adds a parametric estimator and NW estimator to obtain asymptotic properties like those of the Glad (1998) estimator discussed below. Jones, linton, and Nielsen(1995) proposed a purely nonparametric estimator in a multiplicative manner. An initial kernel estimator is multiplied with a nonparametric correction factor reducing bias but increasing variance in the process.

The estimator used in this thesis utilizes the above dimension reduction features. This estimator combines a nonparametric kernel estimator with a parametric start in the regression of binary response data. The form taken by this estimator and its theoretical assumptions are a culmination of the parametric and nonparametric literature and assumptions that have already been discussed in the preceding sections. The relevant assumption, restrictions and related literature will be summarized in the next section.

4.3 Ker and Coble (2001) Estimator

This section explains the workings of the Hjort and Glad (1995) density estimator and how it can be modified to conform to regression estimation of binary response models as in Glad (1998) below. Hjort and Glad (1995) introduced a density estimator that combines a kernel-type correction function with a parametric start. The idea behind this estimator is to begin with a parametric estimate $f(x;\hat{\beta})$ and then

multiply with a kernel-type estimate of the correction function, $r(x) = f(x)/f(x;\hat{\beta})$. The authors propose that

$$\hat{f}(x) = n^{-1}\Sigma k_h(X - x_i)\frac{f(x_i;\hat{\beta})}{f(X;\hat{\beta})} \tag{4.2}$$

be the estimator, where $k(.)$ is the kernel. The parametric start need not closely approximate the true density for the estimator to work well. The authors show that the estimator will work well even with a crude parametric start. Ker and Coble (2001) used the Hjort and Glad density estimator in the estimation of crop yield densities and found that it performed better than comparable density estimators[10].

Glad (1998) came up with the regression counterpart to the above Hjort and Glad (1995) density estimator. The conditional regression function $m(x) = E(Y \mid X = x)$ is estimated by multiplying a pilot parametric estimate $f(x;\hat{\beta})$ with a nonparametric estimator to give $m(x) = f(x;\hat{\beta})\ [f(x)/f(X;\hat{\beta})]$ (Glad 1998). The estimator for the nonparametric correction factor is given by $\hat{r}(x) = f(x)/f(X;\hat{\beta})$ and combining this with the initial parametric start estimate gives the estimate of $m(x)$ as $\hat{m}(x) = f(x;\hat{\beta})\ \hat{r}(x)$, which is the Glad (1998) regression estimator. For information on the asymptotic properties of this estimator readers are directed to the article for technical proofs.

The intuition for the Glad (1998) estimator is that when the initial parametric estimate,$f(x;\hat{\beta})$, captures some features of the shape of $m(x)$, the correction factor,$r(x)$, will be less variable than $m(x)$ itself and thus easier to estimate nonparametrically. The parametric start estimate $f(x;\hat{\beta})$ can be obtained by using any

[10]Ker and Coble(2001) have developed an estimator of the link function for binary response models that has a flavor of the Hjort and Glad's(1995) density estimator and Glad's(1998) regression estimator. This paper will investigate how this estimator performs alongside other existing methods.

parametric method - linear LS regression, nonlinear regression, etc.

The Glad (1998) estimator differs from that Ker and Coble (2001) in a number of respects. The latter is flavored with some properties of the former but is specifically designed to estimate the link function of binary response models. Again Ker and Coble (2001) estimator incorporates the SIM dimension-reduction properties which aid in enhancing convergence rates. The estimator has been modified by incorporating the principles of the single index model in estimating conditional regressions involving qualitative response data. How this is done is explained under the section on SIMs including an explanation of how the link function and parameters are identified and estimated (Sections 2.1, 2.2, and 4.3.1). The proposed semiparametric estimator will take the following form before further modifications (section 4.3.1) are implemented

$$\hat{m}(x) = (nh)^{-1} \sum y_i k_h(Z - z_i) \frac{f(z_i; \hat{\beta})}{f(Z; \hat{\beta})} \; / \; (nh)^{-1} \sum k_h(Z - z_i) \tag{4.3}$$

where $z_i = x_i\beta$ is the index and $\hat{\beta}$ are the set of parameters, and h is the smoothing parameter.

4.3.1 Modification, Identification and Estimation

Before implementing both the nonparametric and Ker and Coble (2001) estimator, Equations (3.2) and (4.3), there is need to make some technical adjustments to avoid problems associated with the denominator approaching zero (Horowitz 1991). To implement these modifications we let B be the compact set containing β. We need to average over certain values of $x\beta$ that exceed a certain small positive value to avoid values close to zero. For the purposes of this thesis the estimation of each point was conducted within two standard deviations of that point. So the data points used satisfy these two criteria. Define A_x and A_{nx} as sets $A_x = x : F(x\hat{\beta}, \hat{\beta}) \geq \eta$

and $A_{nx} = x :|| \ x\hat{\beta} - x\hat{\beta}^* \ ||\leq 2h_n$ for some $x^* \in A_x$ where $\eta > 0$ is a constant and $||\ .\ ||$ is the Euclidean norm. Let $I(.)$ be an indicator function $I(.) = 1$ if the event in the parenthesis occurs and $I(.) = 0$ otherwise. Define $j_j = I(x_j \in A_x)$ and $J_{nj} = I(x_j \in A_{nx})$. These changes will confine the divisors to values away from zero and at the same time avoid averaging over extremely large outliers.

When Equation (4.3) takes into account the above changes we get a semiparametric estimator of the form

$$\hat{m}(x) = \frac{(nh)^{-1} \sum \ y_i \ j_{nj} \ w(x_j) \ k_h(Z - z_i) \ \frac{f(z_i;\hat{\beta})}{f(Z;\hat{\beta})}}{(nh)^{-1} \sum \ j_{nj} \ w(x_j) \ k_h(Z - z_i)} \tag{4.4}$$

where the numerator is composed of $f(z_i.)$ (the parametric start at point i), and $k_h(.)/f(Z.)$ (the correction factor), and the denominator is a normalization term. Thus the minimization of Equation (3.4) (Section 3.2.2) gets the estimates of β. The weight function, w, was set to 1 for this study but other similar applications have used the variance as the weight.

β and $F(.)$ are identified as explained in the section under identification and estimation of binary response models (Section 2.1). β is estimated by optimization techniques applied to Equation (3.4). The sums in Equation (4.4) and (3.4) should be restricted to observations i for which the probability density of $x\beta$ at the point $x_i\beta$ exceeds a small positive number to avoid the denominators getting close to zero as explained above.

Chapter 5

ANALYSIS, RESULTS, AND CONCLUSIONS

5.1 Data and Sampling Technique

The Monte Carlo exercise utilized data recovered randomly from the uniform distribution and then the disturbances incorporated from three different data generating processes from the Marron and Wand(1992) group of test densities, as depicted graphically in Figure 5.1. The first dataset includes disturbances from the standard $N(0,1)$. The second dataset is recovered from a mixture of normal with weights i.e. $1/5N(0,1) + 1/5N(\frac{1}{2},(\frac{2}{3})^2) + 3/5N(\frac{13}{12},(\frac{5}{9})^2)$, a skewed unimodal distribution. The other data was similarly recovered from a kurtotic unimodal distribution, $2/3N(0,1) + 1/3N(0,(\frac{1}{10})^2)$. In all these cases three independent variables were randomly generated and used to generate the dependent variable using coefficients that are known a priori (Section below). Then these parameters were estimated using the different approaches and their mean squared errors noted.

For the application of the model, data from Spector and Mazzeo(1980)were used. This data were used to analyze the effectiveness of new teaching techniques in economics.The dependent variable is an indicator of whether students' grades(GRADES) improved after exposure to a new method of teaching economics(a binary variable in-

dicator, PSI). The other independent variables include grade point average(GPA), and the score on a pretest that shows entering knowledge of the material (TUCE). The probit model, NW and Ker and Coble(2001) estimators were applied to the data, the slopes estimated, and the out-of-sample prediction for these methods compared.

5.2 Simulations

The Monte Carlo study involved 500 simulations on each of the samples of size 50, 100, and 500 for the three data generating processes mentioned above. The probit model, the Nadaraya-Watson SIM estimator, and the Ker and Coble (2001) estimator were applied to the data to get these estimates. This thesis applies these estimators to data generated by a Monte Carlo operation in order to investigate their finite-sample properties. Given a true regression function $m(x)$ we generated n realizations for X_i randomly from the uniform distribution [0,1]. For the first simulation process, the binary endogenous response variable was generated by adding normally distributed noise. For the KC estimator, a probit estimate acts as the initial pilot start and is then multiplied by the nonparametric correction factor applying the LSCV for the smoothing parameter. For this experiment, the fact that we know the design density beforehand is not utilized in the regression process at all; the kernel assumes no such knowledge.

For the case of the standard normal data, the parametric assumption of normality is correct and theory postulates that the estimates from probit regression will converge to the true values. However, for the skewed and kurtotic datasets this assumption is not true and theory predicts less consistent or more biased probit estimates. The KC estimator is expected to have improved convergence when the

parametric start is correct. Parameters are then estimated and their mean squared errors computed.

Several experiments were carried out with different true regression functions and sample sizes of 50, 100, and 500 and different magnitudes of noise (Gaussian, skewed, and kurtotic). Based on these simulations the average of the parameter estimates is determined. For each estimation process the squared bias and variance are calculated for each parameter estimate. The sum of these two terms gives what we call the mean squared error (MSE).

5.3 The Results

5.3.1 Monte Carlo

The results are presented for the three regression estimates in Tables 5.1 to 5.6 at the end of this chapter. In the first simulation, the pilot estimate is correct (the probit estimate of the Gaussian) and belongs to the true parametric family. In the second and third simulation (skewed and kurtotic) the initial parametric estimate is not correct.

From the all the simulation results there is a reduction in MSE as sample size increases even with a rough parametric start. This declines for all sample sizes and indicates a clear downward spiral as n increases. Theory expounds that this should vanish to zero as n increases. However, this experiment was unable to reach such levels due to the time demanded to run such models on computer. It is noteworthy that even at such relatively small sample sizes the estimator exhibits significant bias reduction properties even with incorrect start.

Table 5.1 shows the MSE values for the three estimators for each of the βs

when the parametric start is correct. The results confirm the theoretical expectations and results from other practical applications. For the normal standard data (correct parametric model), the probit outperforms the other estimators by converging quicker within the range of realizations used i.e. positing the lowest MSE values. Convergence in this case is measured by the magnitude of the mean squared error (MSE) across the realizations. The KC estimator does as well as or better than the NW regression estimator in most of these cases; this bolsters theoretical expectations that the combined estimator sort of 'averages" its component parts. The parameter estimates for the probit model are therefore closer to their true values (Table 5.4).

For the skewed and kurtotic regressions, theory predicts better performance from approaches that do not assume link functions a priori. For the skewed regression the KC estimator and NW methods outperform the probit model, with the KC estimator giving the minimum MSE. The MSE for the probit model is higher in this case than what it is for the Gaussian dgp. The MSE results from the kurtotic dgp mirror those of the skewed distribution. For the probit model , the MSE values from the Gaussian noise are lower than those of the kurtotic and skewed disturbances. The other estimators do better than the probit model in the cases involving non-Gaussian disturbance data.

with the wrong parametric estimates, the KC estimator utilizes information from its nonparametric part to correct the wrong start. This agrees with theoretical expectations from Hjort and Glad (1995) and Glad (1998).

5.3.2 Application

Table 5.7 gives the results of the application exercise. The parameter estimates (not the slopes)for the parametric estimation are shown against those of the NW and KC estimators including the correct predictions for each model. For this exercise it was not possible to calculate the MSE because the true values for the parameters are unknown. But a cursory look reveals that the parameter values for all the models are not very "dissimilar". However the KC estimator gives better prediction results.

5.4 Summary and Conclusions

A number of methods exist for solving endogenous response models and the choice of which method to use hinges on a number of considerations. If in some way the data generating process is known to be the normal distribution then the probit model would be the appropriate tool to use. On the other hand, this information may not be easily available and decisions have to be made on how to proceed with the regression process. The nature of the functional form to use is therefore very important in the solution of such problems. What is desirable is a flexible functional form that is able to estimate the model without recourse to assumptions about the link function. The KC estimator is such a form and this study has shown some of its competitive properties when compared to other forms. The KC method combines flexibility and an initial parametric estimate to give results that are more competitive than the other methods discussed in this thesis.

Table 5.1: Gaussian: Mean Squared Error (MSE)

Estimator	*Parameters*	$N = 50$	$N = 100$	$N = 500$
Nonparametric	β_1	0.078	0.013	0.002
	β_2	0.178	0.030	0.010
Parametric	β_1	0.004	0.002	0.000
	β_2	0.022	0.020	0.002
KC	β_1	0.028	0.014	0.002
	β_2	0.158	0.028	0.006

Table 5.2: Skewed: Mean Squared Error (MSE)

Estimator	*Parameters*	$N = 50$	$N = 100$	$N = 500$
Nonparametric	β_1	0.076	0.020	0.004
	β_2	0.405	0.067	0.027
Parametric	β_1	0.456	0.371	0.347
	β_2	2.380	1.770	1.660
KC	β_1	0.070	0.019	0.006
	β_2	0.140	0.070	0.025

Table 5.3: Kurtotic: Mean Squared Error (MSE)

Estimator	*Parameters*	$N = 50$	$N = 100$	$N = 500$
Nonparametric	β_1	0.109	0.042	0.004
	β_2	0.385	0.152	0.033
Parametric	β_1	0.466	0.402	0.354
	β_2	2.914	2.082	1.963
KC	β_1	0.090	0.037	0.004
	β_2	0.280	0.080	0.016

Table 5.4: Gaussian: Parameter Estimates
True Coefficients $\beta_1 = 3,\ \beta_2 = 7$

Estimator	*Parameters*	$N = 50$	$N = 100$	$N = 500$
Nonparametric	β_1	2.72	2.88	3.04
	β_2	6.58	6.83	7.10
Parametric	β_1	3.06	3.05	3.02
	β_2	7.15	7.14	7.05
KC	β_1	3.17	3.12	3.02
	β_2	7.40	7.17	7.07

Table 5.5: Skewed: Parameter Estimates
True Coefficients $\beta_1 = 3,\ \beta_2 = 7$

Estimator	*Parameters*	$N = 50$	$N = 100$	$N = 500$
Nonparametric	β_1	2.82	3.08	3.05
	β_2	6.47	6.86	7.14
Parametric	β_1	3.67	3.61	3.59
	β_2	8.54	8.33	8.29
KC	β_1	3.23	3.10	3.04
	β_2	7.25	7.14	7.04

Table 5.6: Kurtotic: Parameter Estimates
True Coefficients $\beta_1 = 3,\ \beta_2 = 7$

Estimator	*Parameters*	$N = 50$	$N = 100$	$N = 500$
Nonparametric	β_1	2.74	2.85	3.05
	β_2	6.48	6.69	7.16
Parametric	β_1	3.68	3.63	3.59
	β_2	8.70	8.44	8.40
KC	β_1	3.15	3.10	3.05
	β_2	7.29	7.11	7.10

Table 5.7: Application Results

Estimator	*Parameters*	*EstimatedValue*	*PercentPrediction*
Nonparametric	β_1	−7.750	87.20
	β_2	1.525	
	β_3	0.055	
	β_4	1.425	
Parametric	β_1	−7.452	81.70
	β_2	1.626	
	β_3	0.052	
	β_4	1.426	
KC	β_1	−7.650	89.50
	β_2	1.675	
	β_3	0.455	
	β_4	1.385	

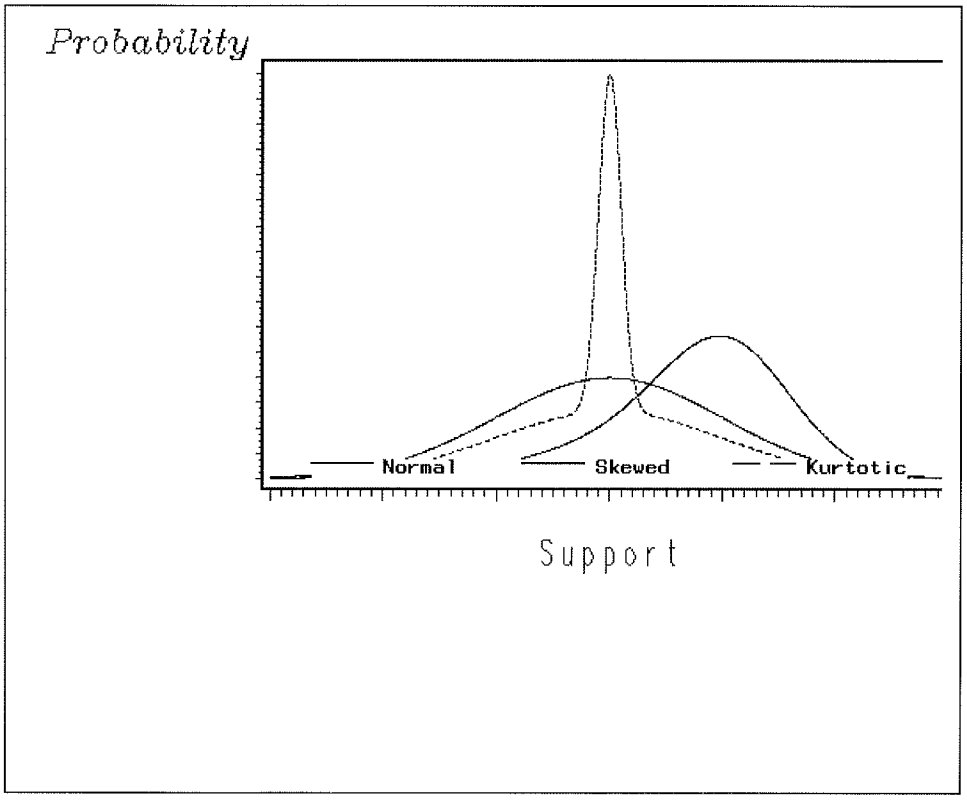

Fig. 5.1: Data Generating Processes: Gaussian, Skewed, and Kurtotic

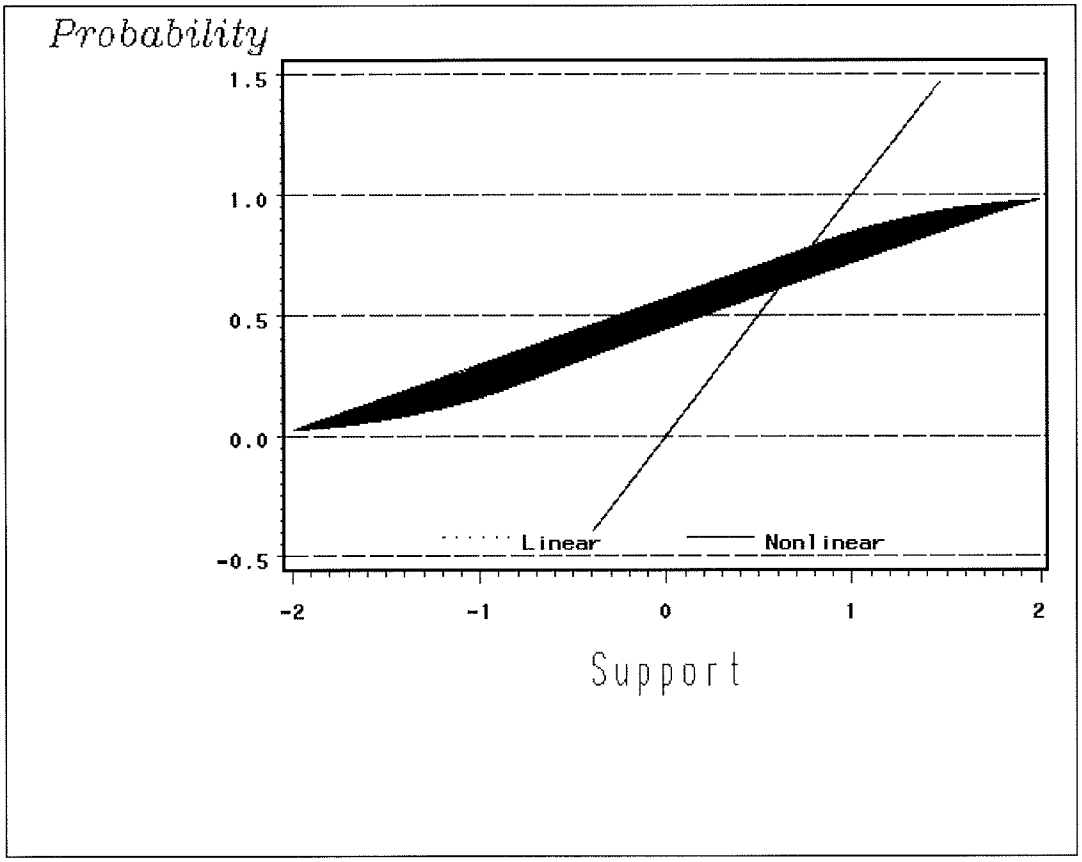

Fig. 5.2: A Comparison: The Probit and Linear Probability Model

Reference

Ai, C., and D. L., McFadden. "Estimation of Some Partially Specified Nonlinear Regression Models." *Journal of Econometrics* 76(January 1997):1-37.

Abramson, I."On Bandwidth Variation in Kernel Estimates - a Square Root Law." *Annals of Statistics* 10(1982):1217-1223.

Ansley, C.F.,R.Kohn and C.Wong."Nonparametric Spline Regression with Prior Information." *Biometrika* 80(1993):75-88.

Chamberlain, G."Asymptotic Efficiency in Semiparametric Models with Censoring." *Journal of Econometrics* 32(1986):189-218.

————-."Efficiency Bounds for Semiparametric Regression." *Econometrica* 60(1992):576-596.

Chen, S."Efficient Estimation of Binary Choice Models Under Symmetry." *Journal of Econometrics* 96(2000):183-199.

Chiu, S. T. "Some Stabilized Bandwidth Selectors for Nonparametric Regression." *The Annals of Statistics* 19(1991):1528-1546.

Cosslett, S. R. "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model." *Econometrica* 51(1983):765-782.

————. "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models." *Econometrica*55(1987): 559-587.

Das, S."A Semiparametric Structural Analysis of the Idling of Cement Kilns." *Journal of Econometrics* 50(1991):235-256.

Deaton, A.*The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore:John Hopkins University Press, (1991).

Fan, J., and I.Gijbels. "Variable Bandwidth and Local Regression Smoothers." *Annals of Statistics* 20(1992):2008-2036.

Fan, J., and I. Gijbels."Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation." *Journal of the Royal Statistical Society* B 57(1995):371-394.

Fan, Y., and A. Ullah."Asymptotic Normality of a Combined Regression Estimator." Unpublished, University of Windsor, 1996.

Fischer, G., and D. Nagin."Random Versus Fixed Coefficient Quantal Choice Models ."in C. Manski and D. McFadden(editors),*Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, Mass.:MIT Press, 1981.

Gasser, T., and H.G.,Muller. "Estimating Regression Functions and their Derivatives by the Kernel Method." *Scandinavian Journal of Statistics* 11(1984):171-185.

Glad, I.K. "Parametrically Guided Non-parametric Regression." *Scandinavian Journal of Statistics* 25(1998):649-668

Gozalo, P., and O. Linton. "Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression." *Journal of Econometrics* 99(November 2000):63-106.

Greene, W.H. *Econometric Analysis. Fourth Edition.* New York University, New Jersey: Prentice Hall, 2000.

Hausman, J.A., and D.A. Wise. A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heteregenous Preferences,*Econometrica* 46(1978): 403-426.

Hastie, T.J., and R.J. Tibshirani.*Generalized Additive Models.* Chapman and Hall, 1990.

Hardle, W., and O. Linton. Applied Nonparametric Methods. *Handbook of Econometrics* 4(1994): 2297-2314.

Hardle, W., and J.S. Marron. "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation." *Annals of Statistics* 13(1985):1465-1481.

Hardle, W., P. Hall, and J.S.,Marron."Regression Smoothing Parameters That Are Not Far From Their Optimum." *Journal of the American Statistical Association* 87(1992):227-233.

Hjort, N.L., and I.K. Glad. "Nonparametric Density Estimation with a Parametric Start." *The Annals of Statistics* 23(1995):882-904.

Horowitz, J.L. Semiparametric methods in econometrics. *Lecture Notes in Statistics* 131 New York: Springer, 1998.

————."Semiparametric Estimation of a Work-trip Mode Choice Model." *Journal of Econometrics* 58(1993):49-70.

————."A Smoothed Maximum Score Estimator for the Binary Response Model." *Econometrica* 60(1992):505-531.

Ichimura, H., and T. T. Scott. "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution." *Journal of Econometrics* 86(October 1998):269-95.

Ichimura, H., and T. S. Thompson. "Maximum Likelihood Estimation of a Binary Choice Model With Random Coefficients of Unknown Distribution." *Journal of Econometrics* 86(1998):269-295.

Jones, M.C., and P.J. Foster. "Generalized Jacknifing and Higher Order Kernels." Forthcoming in Journal of Nonparametric Statistics,1993

Jones, M.C., O.Linton and J.P. Nielsen. "A Simple Bias Reduction Method for Density Estimation." *Biometrika* 82(June 1995):327-338.

Ker, A.P. "Nonparametric Estimation of Possibly Similar Densities." Forthcoming(October 2000).

Ker, A.P., and B.K. Goodwin. "Modeling Yields." Forthcoming (November 2000).

————. "Nonparametric Estimation of Crop Insurance rates Revisited. " *American Journal of Agricultural Economics* 83(May 2000):463-478.

Klein, R.W., and R.S. Spady. "An Efficient Semiparametric Estimator of the Binary Response Model." *Econometrica* 61(1993):387-421.

Lee, T.C.M.,and V.Solo. "Bandwidth Selection for Local Linear Regression: A Simulation Study." *Computational Statistics* 14(1999):515-532.

Macki, Y.P. "Local Properties of K-NN Regression Estimates." SIAM J.Alg.Disc.Meth. 2. (1981):311-23.

Manski, C.F."Maximum Score Estimation of the Stochastic Utility Model of Choice." *Journal of Econometrics* 3(1975):205-228.

————-. "Identification of Binary Response Models." *Journal of the American Statistical Association* 83(1988):729-738.

Marron, J.S., M. P. Wand. "Exact Mean Integrated Squared Error." *Annals of Statistics* 20(June 1992):712-736.

McDonald, J.B. "An Application and Comparison of Some Flexible Parametric and Semi-Parametric Qualitative Response Models." *Economics Letters* 53(1996):145-152

McFadden, D.L. "Quantal Choice Analysis:A Survey." *Annals of Economic and Social Measurement* 5(1976):363-390.

———————."Specification of Econometric Models." *Econometrica*, forthcoming, 1998.

Melenberg, B., and A.van Soest. "Parametric and Semi-parametric Modeling of Vacation Expenditures." Discussion Paper no 9144, *centER for Economic Research*, Tilburg, Holland, 1991.

Nadaraya, E.A."On estimating Regression." *Theory of Probability and Its Application* 9(1964):186-190.

Olkin, I., and C.H., Spiegelman. " A semiparametric approach to density estimation." *Journal of American Statistical Association* 82(1987): 858-865.

Powell, J.L., and T.M. Stoker. "Optimal Bandwidth Choice for Density-Weighted Averages." *Journal of Econometrics* 75(1995):291-316.

Powell, J.L., J.H. Stock, and T.M. Stoker. "Semiparametric Estimation of Index Coefficients." *Econometrica* 57(1989):1403-1430.

Prakasa-Rao, B.L.S. Nonparametric Functional Estimation. New York: Academic Press, 1983.

Quandt, R. "Probabilistic Theory of Consumer Behavior." *Quarterly Journal of Economics* **volume** (70), (1956):507-536.

Ruppert,D., Sheather, S.J., and M.P., Wand. "An Effective Bandwidth Selector for Local Least Squares Regression." *Journal of the American Statistical Association* 90(1995):1257-1270.

Scott, D.W. Multivariate Density Estimation: Theory, Practice, and Visualization. New York:Wiley, 1992.

Spector, L., and M. Mazzeo ." Probit Analysis and Economic Education." *Journal of Economic Education* 11 (1980):37-44.

Stone, C.J. "Optimal Rates of Convergence for Nonparametric Estimators." *Annals of Statistics* 8(1980):1348-1360.

————————. "Optimal Global Rates of Convergence for Nonparametric Regression." *Annals of Statistics* 8(1982):1040-1053.

————————. "Consistent Nonparametric Regression." *Annals of Statistics* 5(1977):595-620.

Thompson, G.D., and J. Kidwell. "Explaining the Choice of Organic Produce: Cosmetic Defects, Prices, and Consumer Preference." *American Journal of Agricultural Economics* 80(1998):277-287.

Tibshirani, R."Local Likelihood Estimation." Phd Dissertation, Stanford University, California, 1984.

Watson,G.S. "Smooth Regression Analysis." *Sankhya Series* 26 (1964):359-372.

Yatchew, A. "Nonparametric Regression Techniques in Economics." *Journal of Economic Literature* 36(1998):669-721.