Essays in Econometrics

by

Jack Ray Porter II

A.B., Applied Mathematics, Harvard University, 1991

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1996

Author......... ....
Department of Economics
June 17, 1996

Certified by.....
Whitney K. Newey
Professor of Economics
Thesis Supervisor

Certified by..... ....
Jerry A. Hausman
John and Jennie S. MacDonald Professor of Economics
Thesis Supervisor

Accepted by..
Richard S. Eckaus
Ford International Professor of Economics
Chairman, Departmental Committee on Graduate Students

Essays in Econometrics

by

Jack Ray Porter II

Submitted to the Department of Economics
on June 17, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Chapter 1. The model considered in this paper is $y_{it} = m(x_{it}) + \delta_i + \eta_{it}$, where δ_i is the individual effect, η_{it} the random disturbance, and m an unknown function to be estimated. The model is a natural nonparametric extension of the standard linear panel data model, maintaining an additive fixed effect. This paper presents two nonparametric estimators of the function, m. The two estimators correspond to the standard first difference and "within" linear panel estimators.

Chapter 2. Nonparametric panel estimation techniques from Chapter 1 are used with COMPUSTAT data on firms to investigate nonlinearities in the investment-q relationship as well as more recent theories of investment. We find that the nonparametric relationship of investment to tax-adjusted q suggests significant convex adjustment costs for large q and a small role for fixed costs. In addition, the relationship of investment to cash flow does not flatten out at very high levels of cash as implied by liquidity constraint models. Instead, the relationship remains stable, as in the free cash flow model or in models where cash flow signals investment opportunities.

Chapter 3. When econometric models are estimated by maximum likelihood, the conditional information matrix variance estimator is usually avoided in choosing a method for estimating the variance of the parameter estimate. However, the conditional information matrix estimator attains the semiparametric efficiency bound for the variance estimation problem. Unfortunately, for even moderately complex models, the integral involved in computation of the conditional information matrix estimator is prohibitively difficult to solve. Simulation is suggested to approximate the integral, and two simulation variance estimators are proposed. Monte Carlo results from a probit model reveal the attractiveness of these estimators in providing accurate confidence interval coverage rates compared to the standard maximum likelihood variance estimators. Monte Carlo results from a sample selection model show the possible gains in confidence interval accuracy in more complicated models where the choice of variance estimators is restricted.

Thesis Supervisor: Whitney Newey
Title: Professor of Economics

Thesis Supervisor: Jerry Hausman
Title: John and Jennie S. MacDonald Professor of Economics

Acknowledgments

I've had the very good fortune of being taught and advised by the absolute best teachers throughout my education. At MIT, my two official advisors' combined knowledge of economics, and econometrics in particular, is undoubtedly unmatched. Whitney Newey has followed my development from my first year in the program, employing me through the summers and advising me through the school years. Jerry Hausman has been the ideal secondary advisor. He has motivated me when necessary and provided valuable, honest advice at all phases of my research. Thomas Stoker should also be singled out for his advisory role and insightful suggestions. Many other professors provided comments and direction during my years in graduate school, Josh Angrist, Jushan Bai, Richard Blundell, Gary Chamberlain, Glenn Ellison, Sara Ellison, Dale Jorgenson, and Steve Pischke.

My classmates contributed enormously to this thesis both as intellectual colleagues and simply as friends. Nick Souleles and Dave Gross were my study group partners throughout the first years of classes, and I learned a great deal from our interactions. Austan Goolsbee deserves special mention as a co-author of Chapter 2 and a valued friend. Many others provided a much-needed laugh as well as a fruitful source of economic insights and questions, among them were Christine Jolls, Michael Lee, Steve Levitt, and Kim Rueben.

My friends and family provided outlets for thinking about things others than economics (what else is there?). My closest friends, Tom and Monica, were always there to laugh with (and often at) me. Of course, my family has always been the foundation of my support. I am ever grateful to Grandma, Kathy, Bob, Amy, Sean, Sue, Rich, Barb, Jon, Rachel, Marj, and yes, you too, Meg. My mother deserves a Ph.D. far more than I. Her strength and perseverence through the most difficult of times has humbled me on my best days and inspired me on my worst. Finally, one person has dealth with my struggles on a daily (and often hourly) basis while at the same time completing her own graduate education. Beth, you are the sun rising every morning. I dedicate this thesis to you.

To Beth

Contents

1 Nonparametric Estimation of a Flexible Panel Data Model 8

1.1 Introduction 8

1.2 Model 11

1.3 First Differencing Type Estimation 14

1.4 Difference From Means Type Estimation 26

1.4.1 Estimation in the Linear Model 27

1.4.2 Estimation in the General Model 29

1.5 Extensions 34

1.5.1 Time-varying Unknown Function 34

1.5.2 Lagged Regressors 36

1.5.3 Partially Linear Model 36

1.6 Conclusion 37

2 A Nonparametric Empirical Analysis of the q Theory of Investment 50

2.1 Introduction 50

2.2 A Simple q Model of Investment 51

2.3 Data 53

2.4 Empirical Results 53

2.4.1 Investment and q 53

2.4.2 Testing for Measurement Error 59

2.4.3 Cash Flow 62

2.5 Conclusion 64

3 Feasible Simulation Variance Estimation for Maximum Likelihood 68

3.1 Introduction 68

3.2 Standard MLE Variance Estimators 70

3.3 Simulation Estimators 75

3.4 Asymptotic Behavior Comparisons 77

3.5 Monte Carlo Results 80

3.5.1 Probit Model 80

3.5.2 Sample Selection Model 83

3.6 Conclusion 85

Chapter 1

Nonparametric Estimation of a Flexible Panel Data Model

1.1 Introduction

Panel data methods and nonparametric regression estimation have been two of the most significant developments in applied econometrics. Panel data is particularly valuable in empirical work since multiple observations on each cross-sectional unit across time provide the possibility of eliminating individual heterogeneity that can cause bias and inconsistency using cross-sectional data alone. As a result, the prevalence of rich panel data sets in the empirical literature has steadily increased. Nonparametric methods have also become increasingly popular in the applied econometrics literature. In particular, nonparametric regression allows researchers to avoid the restrictions of parametric modelling assumptions. In this paper, we hope to combine the most favorable aspects of panel data methods and nonparametric estimation in developing more general nonparametric regression estimators for the following panel data model: $y_{it} = m(x_{it})+\delta_i+\eta_{it}$, where $m(\cdot)$ is the unknown function to be estimated nonparametrically.

We will be considering fixed effects estimators that allow for correlation between x and δ. The assumption of an additive individual effect is critical to our estimation strategy. The additivity means that, as in the linear model, differencing transformations eliminate the fixed effect. First differencing and differencing from means are two of the primary transformations used in the linear panel model. We find that the first difference and difference from means (also known as "within") linear estimators have analogous nonparametric estimators for the nonparametric generalization of the standard linear model, given above.

To illustrate some of the issues involved in estimation of the nonparametric panel model given above, we compare first differences in the linear panel model to first differences in the nonparametric model. In the linear model we have $y_{it} - y_{i,t-1} = x_{it}\beta - x_{i,t-1}\beta + \eta_{it} - \eta_{i,t-1} = (x_{it} - x_{i,t-1})\beta + \eta_{it} - \eta_{i,t-1}$, while first differencing the nonparametric equation gives $y_{it} - y_{i,t-1} = m(x_{it}) - m(x_{i,t-1}) + \eta_{it} - \eta_{i,t-1} \neq m(x_{it} - x_{i,t-1}) + \eta_{it} - \eta_{i,t-1}$. In the linear model, the *first* equality suggests regressing $y_{it} - y_{i,t-1}$ on x_{it} and $x_{i,t-1}$, leading to two estimates of β corresponding to the coefficients for x_{it} and $x_{i,t-1}$. If we define a function $l(z_1, z_2) = m(z_1) - m(z_2)$, then the first equality from the nonparametric model can be rewritten as $y_{it} - y_{i,t-1} = l(x_{it}, x_{i,t-1}) + \eta_{it} - \eta_{i,t-1}$, so that nonparametric regression of $y_{it} - y_{i,t-1}$ on $(x_{it}, x_{i,t-1})$ would lead to a consistent estimate of l. But, our objective is to estimate m. From the linear model, we see that the linearity allows us to easily impose the restriction that the coefficients on x_{it} and $x_{i,t-1}$ are the same, leading to an estimate of β from regressing $y_{it} - y_{i,t-1}$ on $x_{it} - x_{i,t-1}$. In the nonparametric model, the same approach does not in general hold as the inequality above shows. Thus, in the nonparametric case generally, it is somewhat more involved to impose the restrictions necessary to move from estimation of $l(\cdot, \cdot)$ to estimation of $m(\cdot)$. Similar comparisons with the difference from means transformation lead to the same conclusion. Our objective here is to present methods that work with any nonparametric estimation technique (e.g. kernel, Fourier series, splines, etc.).

Series estimation provides a natural method of overcoming the above inequality by imposing the additive structure in estimation as in the linear case. With series estimation, $m(\cdot)$ is approximated with a linear (in the coefficients) expansion, and thus the estimate of l translates easily to an estimate of m as in the linear panel model case. Hence, in the estimation sections, the series results will be presented as a starting point for nonparametric estimation. These results will be followed by more general methods of nonparametric estimation applicable to any nonparametric smoother, including kernel regression.

One method that could overcome the above inequality but will not be explored here is Alternating Conditional Expectations (ACE) estimation. ACE is a method developed in Breiman and Friedman (1985) to impose additivity in nonparametric regression. It does have the advantage of being generally applicable to any nonparametric regression method, but the drawback of ACE is that it is an iterative technique whose distributional properties have not been fully developed in the simple i.i.d. case. Hence we search for other nonparametric estimation schemes. The nonparametric estimators introduced below will be computationally simpler, and we will be able to present some distributional results for these estimators that are useful for inference.

In Section 1.2, the assumptions for the nonparametric panel model are presented and discussed. Our nonparametric estimators, which are analogous to linear first difference and within estimators, are presented in Sections 1.3 and 1.4. Distributional results, asymptotic normality in particular, are also given in each section. In section 1.5, extensions of the basic model are considered along with the corresponding generalizations of the estimators to handle these cases.

Chapter 2 presents an empirical application of the methods discussed here. A panel of firm investment data is used to estimate the investment-q relationship. The standard adjustment costs theory of investment provides the starting point. Rather than imposing commonly-assumed quadratic parametric assumptions on the adjustment cost function, we can estimate a more flexible specification with our nonparametric approach. This flexible approach allows us to assess the role of nonlinearities in the investment-q relationship while still accounting for potential heterogeneity bias. Additionally, this example provides an opportunity to compare the various estimators suggested in the present chapter.

1.2 Model

The linear regression model $y_i = x_i'\beta + \varepsilon_i$ is usually the first one presented in an introductory econometrics class. This basic model has been expanded to $y_{it} = x_{it}'\beta + \delta_i + \eta_{it}$ for panel data and $y_i = m(x_i) + \varepsilon_i$ for nonparametric regression purposes. Both of these extensions have been widely examined and used in the literature. In this paper we study a specification that combines these two extensions into a model naturally made for considering nonparametric panel methods:

$$y_{it} = m(x_{it}) + \delta_i + \eta_{it} \tag{1.1}$$

$$E[\eta_{it}|x_{i1}, \ldots, x_{iT}] = 0 \tag{1.2}$$

where y_{it} is the dependent variable and x_{it} an $r \times 1$ vector of regressors observed across individuals $(i = 1, \ldots, N)$ and time $(t = 1, \ldots, T)$. The individual effects are denoted here by δ_i and the random disturbance by η_{it}. The object of interest is the unknown function $m(\cdot)$ to be estimated.

As currently posed, the model above is not identified. By subtracting a constant from the "true" function $m(\cdot)$ and adding the same constant to each individual effect, the specification in (1.1) and (1.2) still holds. In other words, letting $\underline{m}(x) = m(x) + c$ and $\underline{\delta}_i = \delta_i - c$, we still have $y_{it} = \underline{m}(x_{it}) + \underline{\delta}_i + \eta_{it}$ and $E[\eta_{it}|x_{i1}, \ldots, x_{iT}] = 0$. Thus we need an additional condition to pin down the level of the function $m(\cdot)$. The following is a natural choice of such a condition,

$$E(\delta_i) = 0. \tag{1.3}$$

This condition is used to establish only the level of m without affecting its shape. It is natural to require since intuitively it assumes that $m(\cdot)$ is defined as the function that goes through the middle of the data in an average sense. Thus, the individual effects represent a mean zero type disturbance from the true regression function. Clearly, the identifying condition (1.3) is different from the condition that $E[\delta_i|x_{it}] = 0$, which assumes away any potential correlation between the fixed effects and the regressors.

Thus, under $E[\delta_i|x_{it}] = 0$, heterogeneity bias is no longer an issue and any nonparametric regression estimator of m would be consistent. The condition (1.3) places no such restrictions on the correlation between the fixed effects and the regressors.

Equation (1.2) is the conditional mean zero assumption for the random disturbance. It imposes the restriction that any noncontemporaneous effects on the dependent variable must be derived completely through the individual effect. Obviously, the more restrictive assumption of strict exogeneity, $E[\eta_{it}|x_{i1}, \ldots, x_{iT}, \delta_i] = 0$, will also be sufficient, but not necessary, for the results below.

The model (1.1) departs from the standard linear panel model by replacing the parametric linear-in-the-regressors component with an unknown function of the regressors, $m(\cdot)$. One could consider even more general models such as $y_{it} = \tilde{m}(x_{it}, \delta_i) + \eta_{it}$, but without restricting $\tilde{m}(\cdot, \cdot)$ in some way the δ_i's and $\tilde{m}$ are not separately identified. In (1.1) we make the assumption that the individual effects, δ_i, enter additively. This assumption allows us to identify the model (using (1.3)) without making explicit assumptions about the function of the regressors x_{it}.[1] Under this additive assumption, individual heterogeneity can be simply thought of as a shift in the regression function. This particular additive identifying assumption was made for two reasons. First, an additive individual effect is the most commonly found assumption in parametric panel models. In particular, it follows naturally from the linear panel model. Second, it is a reasonable assumption to make in many applications, including the investment application given in Chapter 2. Also, in models where the individual effect does not enter additively, often some transformation of the dependent variable will lead to an additive fixed effect.

If one estimates using the standard linear panel model when, in fact, the data is generated by the model in (1.1) with some nonlinear $m(\cdot)$, then in general, the resulting estimates are biased and inconsistent in the sense that $x'\hat{\beta} \overset{p}{\not\rightarrow} m(x)$. Commonly one might estimate the coefficients of the linear model using first differences or within estimation. Given that the true model is not linear, these two estimators will

[1] We will refer to $m(\cdot)$ as the regression function, although strictly speaking $\tilde{m}(\cdot, \cdot)$ may be better named as such.

have different probability limits.[2] Thus, nontrivial differences in these two estimators could indicate the need for a more general model than the linear one that is used. Griliches and Hausman (1986) point out that differences in first difference and within estimators could indicate an errors-in-variables problem within the linear framework. If the linear first difference and within estimators are significantly different, one may still be able to discern the underlying cause of misspecification by considering the pattern of first and longer difference estimators. In the no serial correlation model from Griliches and Hausman, the first and longer differences estimators have a distinct pattern of declining bias. Such a distinctive pattern would not necessarily be present in our case of misspecifying a nonlinear model with a linear one.[3] Also, they suggest a specific test for serial correlation in the measurement errors. More generally, a large difference between within and first difference estimators is an indication of some misspecification.

Chamberlain (1982) suggests a multivariate regression test of heterogeneity bias for the linear panel model, $y_{it} = x_{it}'\beta + d_i + \varepsilon_{it}$. Let $y_i = (y_{i1}, \ldots, y_{iT})$ and $x_i = (x_{i1}', \ldots, x_{iT}')'$. Then if E^* is the multivariate linear predictor, we estimate Π from $E^*(y_i|x_i) = \Pi x_i$. To test for heterogeneity bias, Chamberlain tests if the off-diagonal elements of Π are all zero. Moreover, if heterogeneity bias is found, the linear model $E(y_{it}|x_i, d_i) = x_{it}'\beta + d_i$ still implies that Π will have equality of the off-diagonal elements within each column and equality of the diagonal elements. Thus a test of these restrictions on Π is a test of the assumptions of the linear model. But if (1.1) is the true model then $E^*(y_i|x_i) = E^*(m(x_{it}) + \delta_i + \eta_{it}|x_i) = E^*(m(x_{it})|x_i)$ $+E^*(\delta_i|x_i)$ $+E^*(\eta_{it}|x_i) = \gamma_t'x_i + \theta'x_i + 0 = (\gamma_t + \theta)'x_i$, for some γ_t and θ. Here since $E^*(m(x_{it})|x_i) = \gamma_t'x_i$ and γ_t is likely to vary by time period t (as indicated by its subscript), even for strictly stationary $\{x_{it}\}$, we would no longer find equality of the off-diagonal elements within a column of Π. Thus, rejection of equality in the above test of the linear specification with individual effects could indicate the need for use

[2] One may come up with special cases of nonlinear $m(\cdot)$'s where these two estimators do in fact have the same plims, but for actual applications such a result would be very surprising.

[3] This lack of declining bias also results in the Griliches and Hausman (1986) serial correlation case.

of a more general nonlinear model such as (1.1).

Generally speaking, tests for various kinds of misspecification are commonly carried out under the maintained hypothesis of linearity. Dropping the maintained assumption of linearity in favor of a more flexible model often provides another underlying misspecification to account for test rejections. In the next two sections, we consider nonparametric estimation of our nonlinear panel specification.

1.3 First Differencing Type Estimation

With the additive individual effect specified in (1.1), first differencing is a natural way of transforming the model back to a standard nonparametric model without heterogeneity bias. We have $dy_{it} = l(x_{it}, x_{i,t-1}) + d\eta_{it}$, and any nonparametric regression method allows us to estimate $l(\cdot,\cdot)$ consistently. The challenge is in trying to recover an estimate of $m(\cdot)$ from this first differenced equation.

Before introducing our more general first difference nonparametric estimator, we give an asymptotic distribution result for first differencing series estimators. In the series estimation setting, we can impose restrictions during estimation of l that allow us to straightforwardly recover an estimate of m. To carry out series estimation, we need a sequence of approximating functions $\{q_{kK}(\cdot)\}$. Let $q^K(x) = (q_{1K}(x), \ldots, q_{KK}(x))'$ be the $K \times 1$ vector of approximating functions. The aim of series estimation is to approximate the function m by some linear combination of the functions $\{q_{kK}\}$, and the choice of sequence must be such that the linear combination approximation to m improves (in a sense to be defined below) as K gets larger. ssing y_{it} on $q^K(x_{it})$ leads to an inconsistent estimate of m due to the correlation between x_{it} and δ_i. But in first differences we obtain a coefficient estimate $\hat{\beta}$ from least squares regression of dy_{it} on $p^K(x_{it}, x_{i,t-1})$, where $p^K(x_1, x_2) = q^K(x_1) - q^K(x_2)$. Then $q^K(\cdot)'\hat{\beta}$ consistently estimates $m(\cdot)$ to within an additive constant. In Theorem 1, we prove asymptotic normality of this estimator using Lemma 2, a distributional result for multivariate series regression given in the Appendix. Lemma 2 is stated with quite general assumptions (Assumptions A-1', A-2', A-3', A-4') applicable to any series approximating functions. To provide simple and verifiable primitive conditions in Theorem 1, we will narrow our focus to power series and regression splines, as in both Andrews (1991) and Newey (1994a). Distributional results for other approximating functions follow similarly from Lemma 2, the Appendix result.

Next we give the assumptions for Theorem 1.[4] Let $y_i = (y_{i1}, \ldots, y_{iT})'$, $x_i = (x_{i1}, \ldots, x_{iT})'$, and $dy_i = (dy_{i2}, \ldots, dy_{iT})'$. The first assumption contains a fourth moment condition on the residuals and the assumption that the data is i.i.d.

Assumption S-1 *$(y_1, \delta_1, x_1), \ldots, (y_N, \delta_N, x_N)$ i.i.d., $E[\| y_{it} - \delta_i - m(x_{it}) \|^4 |x_i]$ is bounded, and the smallest eigenvalue of $Var(y_i|x_i)$ is bounded away from zero.*

In Assumption S-2, we introduce a trimming function to bound the density of the regressors away from zero on a bounded support. Throughout this paper we will use λ to denote a fixed trimming function. This assumption along with our use of power series or splines ensures that, for some nonsingular transformation (denoted by B_K) of the approximating functions, the second moment matrix of approximating functions is uniformly bounded away from singularity. Thus, the least squares estimator is well-defined. Also, this assumption bounds the growth rate of the supremum norm of the vector of approximating functions, see Assumption A-2 in the Appendix. Let $f_{t1}(x_1, x_2)$ denote $f_{x_{it}, x_{i,t-1}}(x_1, x_2)$, the density of $(x_{it}, x_{i,t-1})$ evaluated at (x_1, x_2). Also, given the nonsingular transformation B_K above, define $\zeta_d(K) = \max_{|\rho| \leq d} \sup_x \| \lambda(x) B_K \partial^\rho p^K(x) \|$.

Assumption S-2 *$\lambda(x_{it}, x_{i,t-1})$ is bounded and zero except on a compact set $\mathcal{X}$, where $f_{t1}(x_{it}, x_{i,t-1})$ is bounded away from zero for $t = 2, \ldots, T$.*

Smooth functions are well approximated by both polynomials and polynomial splines. Thus, Assumption S-3 requires sufficient smoothness of $m(\cdot)$, the function we are approximating, see Assumption A-3 in the Appendix.

[4] For further discussion of these assumptions see Newey (1994a). Andrews (1991) has a similar set of assumptions with a detailed discussion.

Assumption S-3 *There is a nonnegative integer s such that $m(x)$ is continuously differentiable to order s on $\Re^r$.*

To be explicit about the simple first difference series estimator of m, let

$$\hat{m}_F(x) = q^K(x)'\hat{\beta}, \text{ where}$$
$$\hat{\beta} = [\sum_{i=1}^{N}\sum_{t=2}^{T}\lambda(x_{it}, x_{i,t-1})p^K(x_{it}, x_{i,t-1})p^K(x_{it}, x_{i,t-1})']^{-1}$$
$$\cdot[\sum_{i=1}^{N}\sum_{t=2}^{T}\lambda(x_{it}, x_{i,t-1})p^K(x_{it}, x_{i,t-1})dy_{it}].$$

Notice that $\hat{\beta}$ is the least squares estimate from the projection of dy_{it} on $p^K(x_{it}, x_{i,t-1})$, the first differenced approximating function, with the addition of a trimming function, λ. Since series estimation of series term coefficients, β, involves first differenced approximating functions, we can assume q^K does not include a constant term. The Assumption S-3 and our use of polynomials or splines implies that there exists $\gamma_K, \beta_K, \alpha$ such that $|m - \gamma_K - q^{K\prime}\beta_K| = O(K^{-\alpha})$, as given by Assumption A-3' in the Appendix. Our estimator, $\hat{m}_F$, estimates m up to an additive constant, and γ_K is that additive constant.

Andrews and Newey show that variance estimation for linear functionals of series estimators is essentially the same as it is in least squares estimation for fixed K. Define the $K \times (T-1)$ matrix $p^K_{T-1}(x_i) = (p^K(x_{i2}, x_{i1}), \ldots, p^K(x_{iT}, x_{i,T-1}))$, $\lambda p^K_{T-1}(x_i) = (\lambda(x_{i2}, x_{i1})p^K(x_{i2}, x_{i1}), \ldots, \lambda(x_{iT}, x_{i,T-1})p^K(x_{iT}, x_{i,T-1}))$ and $\Sigma_F = E[\lambda p^K_{T-1}(x_i)\Omega(x_i)\lambda p^K_{T-1}(x_i)']$, where $\Omega(x_i) = \text{Var}(dy_i|x_i)$. Then,

$$V^K_F = q^K(x)'\left[E(\lambda p^K_{T-1}(x_i)p^K_{T-1}(x_i)')\right]^{-1}\Sigma_F\left[E(\lambda p^K_{T-1}(x_i)p^K_{T-1}(x_i)')\right]^{-1}q^K(x).$$

To estimate the variance, we define $\hat{\Sigma}_F = \frac{1}{N}\sum_{i=1}^{N}\lambda p^K_{T-1}(x_i)d\hat{\varepsilon}_i d\hat{\varepsilon}_i'\lambda p^K_{T-1}(x_i)'$, where $d\hat{\varepsilon}_{it} = dy_{it} - (\hat{m}_F(x_{it}) - \hat{m}_F(x_{i,t-1}))$ and $d\hat{\varepsilon}_i = (d\hat{\varepsilon}_{i2}, \ldots, d\hat{\varepsilon}_{iT})'$. Then,

$$\hat{V}^K_F = q^K(x)'\left[\sum_{i=1}^{N}\sum_{t=2}^{T}\lambda(x_{it}, x_{i,t-1})p^K(x_{it}, x_{i,t-1})p^K(x_{it}, x_{i,t-1})'/N\right]^{-1}\hat{\Sigma}_F$$

$$\cdot \left[\sum_{i=1}^{N}\sum_{t=2}^{T}\lambda(x_{it},x_{i,t-1})p^K(x_{it},x_{i,t-1})p^K(x_{it},x_{i,t-1})'/N\right]^{-1} q^K(x).$$

Asymptotic normality of $\hat{m}_F$ is shown in the following theorem.[5] The parameter α is determined by the choice of power series or splines via Assumption A-3 in the Appendix. The existence of this parameter is assured by Assumption S-3.

Theorem 1 *If Assumptions (S-1) - (S-3) are satisfied and the approximating functions $\{q_{kK}\}$ are either power series or regression splines, $\sqrt{n}K^{-\alpha} \longrightarrow 0$, $\zeta_0(K)^2K/n \longrightarrow 0$, then*

$$\sqrt{N}(V_F^K)^{-1/2}[(\hat{m}_F(x) - m(x)) + \gamma_K)] \xrightarrow{d} N(0,I)$$

and

$$\sqrt{N}(\hat{V}_F^K)^{-1/2}[(\hat{m}_F(x) - m(x)) + \gamma_K)] \xrightarrow{d} N(0,I).$$

This result for simple first difference series estimation is dependent on the fact that a restriction can be imposed during series estimation that leaves us with a natural estimate of m. The next problem is to find a general method that moves from estimation of l to estimation of m without depending on the particular nonparametric smoother used.

The solution comes from the partial means idea presented in Newey (1994b). Newey is interested in the estimation of additive models using kernel methods, although the partial means idea can be used for other purposes as well. To emphasize the general applicability of partial means estimators to any nonparametric estimator, we will use a generic nonparametric smoother notation. Given dependent variable data y and data on regressors X, let $S^x(y|X)$ be the nonparametric estimate of the regression function evaluated at x using data y and X. If we are interested in our nonparametric smoother evaluated at a data point $x = x_{it}$, then we can simplify our notation to $S^{it}(y|X)$. The general idea of the partial means estimator is as fol-

[5] All proofs are contained in the Appendix.

lows. Our first differenced model from above is $dy_{it} = l(x_{it}, x_{i,t-1}) + d\eta_{it}$. So if x_{it} is r-dimensional, then we treat the difference $m(x_{it}) - m(x_{i,t-1})$ as a function $l(x_{it}, x_{i,t-1})$ on a $2r$-dimensional domain. As suggested above, with the equation in this form, we can use *any* nonparametric regression estimator to obtain an estimate $\bar{l}(x_1, x_2) = S^{(x_1,x_2)}(dy|X, X_{-1})$. Then, we partial out with respect to one of the arguments of l to obtain an estimate $\bar{m}(\cdot)$ of $m(\cdot)$ to within an additive constant. The partialling out is achieved by averaging over the data with respect to one of the arguments of l; specifically, we let $\bar{m}(x) = \frac{1}{NT}\sum_{i,t}\bar{l}(x, x_{it})$. Now we have our partial means estimator $\bar{m}(x)$, which will consistently estimate $m(x) - E[m(x_{it})]$.

If we had stopped after obtaining our estimate $\bar{l}$, we could still obtain an estimate of the function $m(\cdot)$ to within an additive constant without the additional step of partialling out. For any fixed $\bar{x}$, $\bar{l}(x, \bar{x}) \xrightarrow{p} m(x) - m(\bar{x})$ which estimates $m(x)$ to within an additive constant just as partial means. The advantage of the partial means approach is that the second argument of l is "integrated out" through averaging, resulting in a significantly improved convergence rate (by an amount dependent on the dimension of the second argument) and hence significantly decreased standard errors. One might think of the averaging in partial means as an ameliorating influence on the curse of dimensionality.

While we cannot directly impose additivity in the partial means framework, we can take advantage of some of the special structure of our particular problem during estimation. Because the function $l(\cdot, \cdot)$ is just the difference of the same function $m(\cdot)$ evaluated at the two arguments of l, l has the following two properties:

- $l(x_1, x_2) = -l(x_2, x_1) \quad (= m(x_1) - m(x_2))$
- $l(x, x) = 0 \quad (= m(x) - m(x))$

Let $\tilde{l}(x_1, x_2) = \frac{1}{2}\bar{l}(x_1, x_2) - \frac{1}{2}\bar{l}(x_2, x_1)$. Then our estimate $\tilde{l}$ satisfies the first property, in particular $\tilde{l}(x_1, x_2) = -\tilde{l}(x_2, x_1)$, and under simple conditions for series and kernel partial means the second property will also hold for $\tilde{l}$.

We have not yet specified a particular choice of nonparametric smoother in the definition of $\bar{l}$. Next we discuss kernel partial means, where S is defined as a kernel weighted average of the dependent variable and the kernel weights depend on both the regressor data and the point of evaluation. Following an asymptotic normality result for the kernel regression case, we present the series result for partial means.

In our presentation of kernel regression partial means, we will use the Nadaraya-Watson kernel smoother,

$$\hat{l}(x_1, x_2) = \frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=2}^{T} K_\sigma \begin{pmatrix} x_1 - x_{it} \\ x_2 - x_{i,t-1} \end{pmatrix} dy_{it}}{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=2}^{T} K_\sigma \begin{pmatrix} x_1 - x_{it} \\ x_2 - x_{i,t-1} \end{pmatrix}}.$$

Here if $\mathcal{K}$ denotes a kernel, then we define $K_\sigma(u) = \frac{1}{\sigma^{2r}} \mathcal{K}(\frac{u}{\sigma})$. As before, let $f_{t1}(x_1, x_2)$ denote $f_{x_{it}, x_{i,t-1}}(x_1, x_2)$, the density of the random variables $(x_{it}, x_{i,t-1})$ evaluated at (x_1, x_2), which is independent of i since observations are assumed identically distributed across individuals, and let $\bar{f}_1(x_1, x_2) = \frac{1}{T-1} \sum_{t=2}^{T} f_{t1}(x_1, x_2)$. Nadaraya (1965) and Watson(1964) first suggested this flexible method of estimating regression functions. The denominator consistently estimates the density $\bar{f}_1(x_1, x_2)$, so we define $\hat{f}_{t1}(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x_1 - x_{it}, x_2 - x_{i,t-1})$ and $\hat{\bar{f}}_1(x_1, x_2) = \frac{1}{T-1} \sum_{t=2}^{T} \hat{f}_{t1}(x_1, x_2)$. The numerator consistently estimates $\frac{1}{T-1} \sum_{t=2}^{T} E[dy_{it} | x_{it} = x_1, x_{i,t-1} = x_2] f_{t1}(x_1, x_2) = l(x_1, x_2) \bar{f}(x_1, x_2)$. Thus the ratio $\hat{l}(x_1, x_2)$ provides a consistent estimate of $E[dy_{it} | x_1, x_2] = l(x_1, x_2)$ $(= m(x_1) - m(x_2))$. The possibility that the random denominator, which is a density estimator, is very small can cause technical difficulties. We follow the many applications (e.g. Manski (1984), Robinson (1988)) that have dealt with this problem previously and allow for a fixed trimming condition that bounds the denominator away from zero.

The second step in estimation is to partial out with respect to one of the arguments of $\hat{l}$. As noted above, bounding the denominator away from zero via a trimming condition is necessary to apply the asymptotic theory to follow. To allow for this possibility, we introduce the weight function $\lambda(\cdot)$ which may be associated with fixed trimming. We suppose that $E[\lambda(x_{is})] = 1$. The kernel partial means estimators are

$$\begin{aligned} \hat{m}_{P,1}(x) &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\lambda(x_{js})\hat{l}(x, x_{js}) \\ \hat{m}_{P,2}(x) &= -\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\lambda(x_{js})\hat{l}(x_{js}, x). \end{aligned}$$

Note here that $\hat{m}_{P,1}$ denotes the estimator that partials out the second argument by averaging across the data from all T time periods using the kernel regression estimator $\hat{l}$. The second estimator $\hat{m}_{P,2}$ is the corresponding estimator partialling out with repect to the first argument. Let $\underline{m}(x) = m(x) - \frac{1}{T}\sum_{s=1}^{T} E[\lambda(x_{is})m(x_{is})]$ $(= \frac{1}{T}\sum_{s=1}^{T} E[\lambda(x_s)l(x, x_s)])$, then both of these partial means estimators estimate $\underline{m}(x)$.

As noted above, it is simple to define an estimator that satisfies two properties of the function l, $l(x_1, x_2) = -l(x_2, x_1)$ and $l(x, x) = 0$. Let $\tilde{l}(x_1, x_2) = \frac{1}{2}\hat{l}(x_1, x_2) - \frac{1}{2}\hat{l}(x_2, x_1)$. Then our estimate $\tilde{l}$ satisfies the first property, and if $\mathcal{K}$ is a kernel such that $\mathcal{K}(u_1, u_2) = \mathcal{K}(u_2, u_1)$ then $\tilde{l}(x, x) = 0$.[6] Now define the partial means estimator that imposes these conditions.

$$\begin{aligned} \hat{m}_P(x) &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\lambda(x_{js})\tilde{l}(x, x_{js}) \quad \left(= -\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\lambda(x_{js})\tilde{l}(x_{js}, x)\right) \\ &= \frac{1}{2}\hat{m}_{P,1}(x) + \frac{1}{2}\hat{m}_{P,2}(x). \end{aligned}$$

Thus, $\hat{m}_P(x)$ is simply the average of the estimators $\hat{m}_{P,1}(x)$ and $\hat{m}_{P,2}(x)$. In Theorem 2 below, we show asymptotic normality of these kernel partial means estimators. First the assumptions for Theorem 2 are discussed.

Our first assumption concerns the kind of kernel to be used. High-order (bias-reducing) kernels, introduced by Bartlett (1963), are used here to achieve the given convergence rate. The main idea is that we reduce the order of the bias for each moment assumed to be zero. Though the choice of kernel does not enter into the asymptotic variance to be given below, still the higher the order of the kernel the

[6] This condition holds when we use a product kernel, i.e. $\mathcal{K}(u_1, u_2) = \mathcal{K}(u_1) \cdot \mathcal{K}(u_2)$.

higher is its own variance, and practical experience suggests some carry over to our estimator. So we recommend a conservative choice of bias-reduction in the selection of a kernel. Many authors (e.g. Prakasa Rao (1983)) have presented straightforward simple methods of constructing kernels that satisfy the following assumption.

Assumption P-1 *There are positive integers Δ and ζ such that $\mathcal{K}(u_1, u_2)$ is differentiable of order Δ, the derivatives of order Δ are Lipschitz, $\mathcal{K}(u_1, u_2)$ is zero outside a bounded set, $\int \mathcal{K}(u)du = 1$, and for all $n < \zeta$, $\int \mathcal{K}(u)[\otimes_{l=1}^{n} u]du = 0$.*

The higher-order kernel assumption is effective in bias-reduction when used together with the next assumption. It imposes smoothness on the regression function and the joint density of the regressors from consecutive time periods.

Assumption P-2 *There is a nonnegative integer d and an extension of $l(x_1, x_2) \cdot f_{t1}(x_1, x_2)$ for $t = 2, \ldots, T$ to all of $\Re^{2r}$ that is continuously differentiable to order d on $\Re^{2r}$.*

Next we present the notation for the asymptotic covariance of the estimators.

$$V_{jt}^{P} = K_j \int \lambda(x_j(a))^2 \bar{f}_1(x_j(a))^{-2} \bar{f}_0(a)^2 \mathrm{Var}(dy_{it}|x_j(a)) f_{t1}(x_j(a))da$$

where $K_1 = \left[\int \left[\int K(u_1, u_2)du_2\right]^2 du_1\right]$ and $K_2 = \left[\int \left[\int K(u_1, u_2)du_1\right]^2 du_2\right]$. Now we are ready to present our asymptotic normality result for our first difference type estimators.

Theorem 2 *Suppose that (i) $E[\|dy_{it}|^4] < \infty$, $E[\|dy_{it}|^4|x_1, x_2]f_{t1}(x_1, x_2)$, and $f_{t1}(x_1, x_2)$ are bounded for $t = 2, \ldots, T$; (ii) Assumptions (P-1) and (P-2) are satisfied for $d \geq \zeta$; (iii) $\lambda(x, x_s) = \lambda(x_s, x) = \lambda(x_s)$ is bounded and zero except on a compact set where $f_{t1}(x, x_s)$ is bounded away from zero for $j = 1, 2$, $s = 1, \ldots, T$, and $t = 2, \ldots, T$; (iv) $\lambda(x_s)$ and $f_{x_s}(x_s)$ are continuous a.e. and $f_{x_s}(x_s)$ is bounded for all $s = 1, \ldots, T$; $E(dy_{it}^2|x_1, x_2)$ are continuous and for some $\epsilon > 0$ and for all $t = 1, \ldots, T$, $\int \sup_{\|\eta\| \leq \epsilon} E[dy_{it}^4|x = (x_1 + \eta, x_2)]f_{t1}(x_1 + \eta, x_2)dx_2 < \infty$; (v) $\frac{N\sigma^{2k-k_1}}{(\ln N)^2} \longrightarrow \infty$ and*

$N\sigma^{k_1+2\zeta} \longrightarrow 0$. Then,

$$\sqrt{N}\sigma^{\frac{k_1}{2}}(\hat{m}_{P,j}(x) - \underline{m}(x)) \xrightarrow{d} N(0, (\frac{1}{(T-1)^2}\sum_{t=2}^{T} V_{jt}^P))$$

$$\sqrt{N}\sigma^{\frac{k_1}{2}}(\hat{m}_P(x) - \underline{m}(x)) \xrightarrow{d} N(0, (\frac{1}{4(T-1)^2}\sum_{t=2}^{T} V_{1t}^P + V_{2t}^P))$$

If, in addition, $N\sigma^{3k-k_1} \longrightarrow \infty$, then $\sigma^{k_1}\hat{V}_{jt}^P \xrightarrow{p} V_{jt}^P$, where $\hat{V}_{jt}^P = \sum_{i=1}^N \hat{\psi}_{it,j}\hat{\psi}_{it,j}/N$, where $\hat{\psi}_{it,j} = \lambda(\tilde{x}_{it,j})\hat{l}(\tilde{x}_{it,j}) - \hat{m}_{P,j}(x) + \hat{\alpha}_{it,j} - \sum_{l=1}^N \hat{\alpha}_{lt,j}/N$, $\hat{\alpha}_{it,j} = (NT)^{-1}\sum_{l=1}^N \sum_{s=1}^T$ $\lambda(\tilde{x}_{ls,j})\hat{\hat{f}}_1(\tilde{x}_{ls,j})^{-1}$ $[dy_{it} - \hat{l}(\tilde{x}_{ls,j})]K_\sigma(\tilde{x}_{ls,j} - (x_{it}, x_{it-1}))$ and $\tilde{x}_{is,1} = (x, x_{is})$, $\tilde{x}_{is,2} = (x_{is}, x)$.

The regularity conditions of Theorem 2 are fairly standard and should be satisfied for a wide variety of applications. Condition (iii) gives λ the fixed trimming interpretation and helps to avoid the denominator problem. Conditions (i) and (iv) are useful dominance conditions. Condition (v) specifies the rate conditions on the bandwidth and incorporates undersmoothing.

The variance estimator given is the one suggested by Newey (1994b). It is a delta method type variance estimator obtained by considering our estimator as a functional of the kernel estimator. Actually the terms $\lambda(\tilde{x}_{it,j})\hat{l}(\tilde{x}_{it,j}) - \hat{m}_{P,j}(x)$ and $\sum_{l=1}^N \hat{\alpha}_{lt,j}/N$ are both asymptotically negligible so that $\sigma^{k_1}\sum_{i=1}^N \hat{\alpha}_{it,j}\hat{\alpha}_{it,j}/N$ will also be consistent for V_{jt}. We see that there are no terms corresponding to the asymptotic covariance of $\hat{\alpha}_{it,j}$ and $\hat{\alpha}_{is,k}$ (for any $s \neq t$ or $j \neq k$). In the proof of Lemma 3, it is shown that such covariance terms are asymptotically negligible. Thus, the only terms that appear in the asymptotic variance correspond to the variance of $\hat{\alpha}_{it,j}$, which indicates a significant correlation only between terms where the data from the same time period (t) is used in the kernel estimation and the same argument (j) is being partialled out. Estimates of the asymptotically negligible cross-terms are left out of the variance estimator, but from the proof of Lemma 3 it is clear how to include the estimates of the asymptotically negligible covariance terms, if desired. Such an inclusion could serve to improve the small sample performance in specific applications.

We do not give the asymptotic distribution corresponding to evaluation of the estimator at a finite set of distinct points, but it is straightforward to show that it is multivariate normal with diagonal asymptotic variance matrix. The intuition for the asymptotic covariances being zero is that as the bandwidth approaches zero, the neighborhood of positive kernel weight around each point shrinks so that the overlap is negligible.

Choice of bandwidth is always an important consideration in kernel methods. Theorem 2 gives the rate conditions that govern that choice asymptotically, but, as often happens, provides little practical guidance for a particular data set. It is important to notice that some "undersmoothing" is employed to establish asymptotic normality. As a practical matter, undersmoothing indicates that the selected bandwidth σ should be chosen to be less than the bandwidth which minimizes the mean square error (as might be suggested by a cross-validation criterion). To avoid a further level of complexity, the theorem given here does not address the issue of data-driven choices for σ. Still, cross-validation might provide a starting point from which the bandwidth can be decreased. For multidimensional x, one may want to scale the various dimensions by, for example, the inverse of the standard deviations of each dimension. Again, we avoid that additional complication in the above theorem, but it is straightforward to modify the theorem to take into account rescaling for multidimensional x (see Robinson (1983)).

For completeness and as a basis for comparison with the simple first difference series estimators, we turn now to series partial means estimation. Again r is the dimension of x_{it}. In the first step of partial means estimation, dy_{it} is regressed on the $2r$-dimensional approximating functions, $q^K(x_{it}, x_{i,t-1})$. Because we are interested in the function m, which is defined on an r-dimensional domain, we must have a dimension reduction in the second step to obtain an estimate of m. The dimension reduction is accomplished through averaging out with respect to one of the arguments. This method contrasts with the simple first difference series case, where the approximating functions used were the differences of approximating functions on an r-dimensional space. In that case, the dimension reduction is built into the first step of series regression by our choice of approximating functions. That choice imposes the restriction that the function to be estimated (l) on the $2r$-dimensional domain is defined as the difference of the same function (m) on two different r-dimensional domains.

In the first step, regressing dy_{it} on $q^K(x_{it}, x_{i,t-1})$ yields the $2r$-dimensional vector of estimated series-term coefficients, $\bar{\beta} = [\sum_{i=1}^N \sum_{t=2}^T \lambda_t(x_{it}, x_{i,t-1}) q^K(x_{it}, x_{i,t-1}) \cdot q^K(x_{it}, x_{i,t-1})']^{-1} [\sum_{i=1}^N \sum_{t=2}^T \lambda_t(x_{it}, x_{i,t-1}) q^K(x_{it}, x_{i,t-1}) dy_{it}]$. Again, λ is a trimming function to bound the density away from zero. In the second step, we partial out by averaging, leading to the series partial means estimators

$$\hat{m}_{M,1}(x) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \lambda(x, x_{it}) q^K(x, x_{it})' \bar{\beta}$$

and

$$\hat{m}_{M,2}(x) = -\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \lambda(x_{it}, x) q^K(x_{it}, x)' \bar{\beta}.$$

Imposing the restrictions as before, $\hat{m}_M(x) = \frac{1}{2}\hat{m}_{M,1}(x) + \frac{1}{2}\hat{m}_{M,2}(x)$.

Again variance estimation is just as in least squares estimation for fixed K. Let $\bar{q}_1^K(x) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \lambda(x, x_{it}) q^K(x, x_{it})$, $\bar{q}_2^K(x) = -\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \lambda(x_{it}, x) q^K(x_{it}, x)$, and $\bar{q}^K(x) = \frac{1}{2}\bar{q}_1^K(x) + \frac{1}{2}\bar{q}_2^K(x)$. Define the $K \times (T-1)$ matrix $q_{T-1}^K(x_i) = (q^K(x_{i2}, x_{i1}), \ldots, q^K(x_{iT}, x_{i,T-1}))$, $\lambda q_{T-1}^K(x_i) = (\lambda(x_{i2}, x_{i1}) q^K(x_{i2}, x_{i1}), \ldots, \lambda(x_{iT}, x_{i,T-1})\ q^K(x_{iT}, x_{i,T-1}))$ and $\Sigma_M = E[\lambda q_{T-1}^K(x_i)\ \bar{\Omega}(x_i)\ \lambda q_{T-1}^K(x_i)']$, where $\bar{\Omega}(x_i) = \mathrm{Var}(dy_i | x_i)$, as before. Then, for $j = 0, 1, 2,$[7]

$$V_{M,j}^K = \bar{q}_j^K(x)' \left[E(\lambda q_{T-1}^K(x_i) q_{T-1}^K(x_i)')\right]^{-1} \Sigma_M \left[E(\lambda q_{T-1}^K(x_i) q_{T-1}^K(x_i)')\right]^{-1} \bar{q}_j^K(x).$$

To estimate the variance, define $\hat{\Sigma}_M = \frac{1}{N} \sum_{i=1}^N \lambda q_{T-1}^K(x_i) d\hat{\varepsilon}_i d\hat{\varepsilon}_i' \lambda q_{T-1}^K(x_i)'$, where $d\hat{\varepsilon}_{it} = dy_{it} - (\hat{m}_M(x_{it}) - \hat{m}_M(x_{i,t-1}))$ and $d\hat{\varepsilon}_i = (d\hat{\varepsilon}_{i2}, \ldots, d\hat{\varepsilon}_{iT})'$. Then, for $j = 0, 1, 2,$

$$\hat{V}_{M,j}^K = \bar{q}_{M,j}^K(x)' \left[\sum_{i=1}^N \sum_{t=2}^T \lambda(x_{it}, x_{i,t-1}) q^K(x_{it}, x_{i,t-1}) q^K(x_{it}, x_{i,t-1})'\right]^{-1} \hat{\Sigma}_M$$

[7] A subscript '$M, 0$' is used interchangeably with the subscript 'M'.

$$\left[\sum_{i=1}^{N}\sum_{t=2}^{T}\lambda(x_{it},x_{i,t-1})q^K(x_{it},x_{i,t-1})q^K(x_{it},x_{i,t-1})'\right]^{-1}\bar{q}_{M,j}^K(x).$$

The next theorem shows asymptotic normality of these series partial means estimators.

Interestingly, because we have focused on polynomial and spline series approximations, the conditions of Theorem 1 did not even include assumptions on the how well the series terms approximate m (or l). Polynomials and splines have known approximating properties, so no further assumptions are needed on the approximating functions. Instead, the conditions are on the specification of the first differenced model. Partial means estimation is performed on the same first differenced model, and these conditions from Theorem 1 are also sufficient to give the partial means result. Again, the parameter $\bar{\alpha}$ is determined by the choice of approximating functions via Assumption A-3.

Theorem 3 *If Assumptions (S-1) - (S-3) are satisfied and the approximating functions $\{q_{kK}\}$ are either power series or regression splines, $\sqrt{n}K^{-\bar{\alpha}} \longrightarrow 0$, $\zeta_0(K)^2K/n \longrightarrow 0$, then for $j = 0, 1, 2$,*

$$\sqrt{N}(V_{M,j}^K)^{-1/2}[\hat{m}_{M,j}(x) - (m(x) - E[\lambda_j(x)m(x)])] \xrightarrow{d} N(0, I)$$

and

$$\sqrt{N}(\hat{V}_{M,j}^K)^{-1/2}[\hat{m}_{M,j}(x) - (m(x) - E[\lambda_j(x)m(x)])] \xrightarrow{d} N(0, I).$$

This result is a direct application of Lemma 1 from the Appendix. We use $2r$-dimensional polynomial or spline approximating functions with a functional defined as the expectation of l with respect to one of the arguments. In Chapter 2, we compare a partial means estimator and a simple first difference series estimator for our investment application and find quite similar performance.

1.4 Difference From Means Type Estimation

The estimation procedure described in this section is the nonparametric analog of the within estimator for a linear panel data model. The analogy will become clear in Section 1.4.1 as the estimator is applied in the linear model setting. This procedure has two steps. First, the individual effects are treated as fixed effects and estimated. Second, the estimated fixed effects are in turn used in estimation of the unknown function $m(\cdot)$ from (1.1).

Suppose we could *observe* the fixed effects, then we could define a new variable $y^*_{it} = y_{it} - \delta_i$. If we treat y^* as the dependent variable of interest, then we could obtain an estimate of m using standard nonparametric regression techniques on (y^*, X). In other words, we could consider the slightly modified model $y^*_{it} = m(x_{it}) + \eta_{it}$, which is in standard form for nonparametric regression estimation of the function m. This basic idea motivates the second step in our estimation procedure. Specifically, if estimates of the fixed effects, $\hat{\delta}_i$, are obtained in the first step, then we can use those estimates to form $\hat{y}^*_{it} = y_{it} - \hat{\delta}_i$. Our estimate for $m(\cdot)$ will be the nonparametric regression estimate using the data $(\hat{y}^*, X)$. We underline the generality of this procedure by presenting it with the same generic nonparametric smoother notation used to describe general partial means estimation. Ideally we would like to estimate m by $\hat{m}(x) = S^x(y^*|X)$, but we do not have the true fixed effects values so our estimate is $\hat{m}(x) = S^x(\hat{y}^*|X)$, using the fixed effects estimates from the first step. For fixed T, we cannot consistently estimate δ_i, but the idea is that $\hat{m}$ will be consistent as the nonparametric smoother "averages" across the errors $\hat{\delta}_i - \delta_i$.

The first step produces the fixed effect estimates used, as just described, in the second step. The estimated fixed effects will be those that minimize the sum of squared distances between $\hat{y}^*$ and its regression estimate at each data point. We will use the same smoother notation introduced above (note that a superscript *it* denotes x_{it}), although a different nonparametric regression estimator could conceivably be used in the first and second steps. The fixed effects are estimated as follows,

$$\begin{aligned}\{\hat{\delta}_1,\ldots,\hat{\delta}_N\} \in \underset{\delta_1,\ldots,\delta_N}{\operatorname{argmin}} \; & \textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it}-\delta_i)-S^{it}(y^*(\delta)|X)]^2 \\ \text{s.t.} \; & \textstyle\sum_{i=1}^{N}\delta_i = 0.\end{aligned} \tag{1.4}$$

Here we use the notation $y^*(\delta)$ to emphasize that y^* is a function of the fixed effects being optimized over. The constraint $\sum_i \delta_i = 0$ is the estimation embodiment of the identifying condition given in (1.3). It is interesting to note that this objective function produces the nonparametric regression estimates for the partially linear model as given in Robinson (1988). If δ_i is replaced by the linear term to be partialled out and S^{it} is a kernel regression smoother, then the solution to the optimization problem in (1.4) is exactly Robinson's estimator. So, as we will find in section 1.5, including an additive linear-in-the-regressor term to the model (1.1) will result in a natural extension of the estimation procedure given in (1.4).[8]

The discussion of the conceptual framework for estimation has purposely been presented as only a general outline. In the next two subsections we will examine this estimation procedure in more specific settings. First, this approach to estimation will be applied in the familiar linear model. Then, the remainder of the section will focus on nonparametric series and kernel regression smoothers.

1.4.1 Estimation in the Linear Model

In the linear model, $m(x_{it}) = x_{it}\beta$ and the object is to estimate β. Here, the smoother is simply the projection on X, so $S^{it}(y|X) = x_{it}(X'X)^{-1}X'y$. Suppose β includes a constant term, then identification requires that the estimation of the fixed effects in equation (1.4) be performed under the restriction that the sum of fixed effects is zero.

[8] It is also essentially the objective function used in Ichimura(1993). But Ichimura obtains estimates of a single index model by minimizing over parameters that are a part of the kernel component.

Hence the minimization problem is:

$$\{\hat{\delta}_1,\dots,\hat{\delta}_N\} \in \underset{\delta_1,\dots,\delta_N}{\text{argmin}} \; \textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it}-\delta_i) - x_{it}(X'X)^{-1}X'(y-\delta\otimes e_T)]^2$$
$$\text{s.t. } \textstyle\sum_{i=1}^{N}\delta_i = 0 \tag{1.5}$$

where e_T is a $T \times 1$ vector of ones and δ is the $N \times 1$ vector of fixed effects so that the Kronecker product $\delta \otimes e_T$ is an $NT \times 1$ vector that corresponds to the $NT \times 1$ vector y of dependent variable observations. Substituting $\delta_1 = -\delta_2 - \dots - \delta_N$, we then have an unconstrained optimization problem. In vector form,

$$\min_{\delta_{-1}} \; (y - D^1\delta_{-1} - P_X(y - D^1\delta_{-1}))'(y - D^1\delta_{-1} - P_X(y - D^1\delta_{-1}))$$

where $P_X = X(X'X)^{-1}X'$ is the projection on X, $\delta_{-1} = (\delta_2,\dots,\delta_N)'$, and D^1 is the $NT \times N-1$ matrix such that $D^1\delta_{-1}$ is the $NT \times 1$ vector version of the fixed effects with $-\delta_2 - \dots - \delta_N$ in place of δ_1, the fixed effect for the first individual. The minimization above is familiar from Generalized Least Squares and the closed form solution is well-known, $\hat{\delta}_{-1} = [D^{1\prime}(I-P_X)D^1]^{-1}D^{1\prime}(I-P_X)y$.

In the second step of the estimation procedure we obtain an estimate of $x\beta$ at any given x, or more simply an estimate of β, $\hat{\beta} = (X'X)^{-1}X'(y - D^1\hat{\delta}_{-1}) = (X'X)^{-1}X'(I - D^{1\prime}[D^{1\prime}(I-P_X)D^1]^{-1}D^1(I-P_X))y$. Since a constant term is included in the regression, let $\beta = (\alpha, \gamma')'$, where α is the constant term, and $X = [e_{NT}\ \tilde{X}]$. It is a well-known result that $\hat{\gamma} = [\tilde{X}'(I_{NT} - (\frac{1}{T}I_N \otimes e_Te_T'))\tilde{X}]^{-1}\tilde{X}'(I - (\frac{1}{T}I_N \otimes e_Te_T'))y$, which is exactly the within estimator so familiar in the linear case. So we have shown that using our method with an OLS smoother on the linear model leads to the standard within estimator. This example then provides the intuition for why the nonparametric estimator introduced above is the analog of the difference from means estimator.

1.4.2 Estimation in the General Model

Within the flexible nonlinear framework, we consider kernel and series choices for the nonparametric smoother. An asymptotic normality theorem is given for the series regression case. Conceivably, one could choose different nonparametric smoothers for the first and second steps, though we do not explore that possibility here. More likely, one could easily generalize the following to include different bandwidth choices or different numbers of series terms for the first and second steps.

We begin by rewriting the minimization (1.4) used to obtain fixed effects estimates with a kernel regression smoother,

$$\{\hat{\delta}_1,\ldots,\hat{\delta}_N\} \in \underset{\delta_1,\ldots,\delta_N}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\lambda_{it}[(y_{it}-\delta_i)-\sum_{j=1}^{N}\sum_{s=1}^{T}w_{js}^{it}(y_{js}-\delta_j)]^2 \tag{1.6}$$

$$\text{s.t.} \sum_{i=1}^{N}\delta_i = 0$$

where $w_{js}^{it} = \frac{\frac{1}{NT}K_\sigma(x_{it}-x_{js})}{\frac{1}{NT}\sum_{k=1}^{N}\sum_{r=1}^{T}K_\sigma(x_{it}-x_{kr})}$ and $\lambda_{it} = \lambda(x_{it})$.

The minimization in (1.6) has a familiar closed form solution. First, we vectorize and substitute in $\delta_1 = -\delta_2 - \ldots - \delta_N$ to produce an unconstrained optimization. Let δ_{-1} and D^1 be defined as before, and $W = [w_{js}^{it}]$ such that $Wz =$ kernel regression estimate of z. Let $\Lambda = \text{diag}(\lambda_{it})$, the matrix with elements λ_{it} on the diagonal and zeros on the off-diagonal.

$$\begin{aligned} &\min_{\delta_{-1}} [(y-D^1\delta_{-1})-W(y-D^1\delta_{-1})]'\Lambda[(y-D^1\delta_{-1})-W(y-D^1\delta_{-1})] \\ &\iff \min_{\delta_{-1}} [\Lambda^{\frac{1}{2}}(I-W)(y-D^1\delta_{-1})]'[\Lambda^{\frac{1}{2}}(I-W)(y-D^1\delta_{-1})] \end{aligned}$$

This minimization is in exactly the same form as the one seen in generalized least squares, where $(I-W)'\Lambda(I-W)$ plays the role served by the inverse of the residual variance-covariance matrix in GLS. So the solution has a familiar form.

$$\begin{aligned} \hat{\delta}_{-1} &= [D^{1\prime}(I-W)'\Lambda(I-W)D^1]^{-1}D^{1\prime}(I-W)'\Lambda(I-W)y \\ &= [\sum_{i,t}\lambda_{it}(D_{it}^1-\hat{D}_{it}^1)(D_{it}^1-\hat{D}_{it}^1)']^{-1}[\sum_{i,t}\lambda_{it}(D_{it}^1-\hat{D}_{it}^1)(y_{it}-\hat{y}_{it})] \end{aligned}$$

Here, $D^1_{it} = ((i,t)^{th}$ row of $D^1)'$ and $\hat{D}^1_{it} = ((i,t)^{th}$ row of $WD^1)'$. With $\hat{\delta}_{-1}$ expressed in this latter form, the similarity to Robinson's (1988) estimator is unmistakeable. If we replace $D^1\delta_{-1}$ by $X\beta$, then the solution is $\hat{\beta} = [\sum_{i,t} \lambda_{it}(x_{it} - \hat{x}_{it})(x_{it} - \hat{x}_{it})']^{-1}[\sum_{i,t} \lambda_{it}(x_{it} - \hat{x}_{it})(y_{it} - \hat{y}_{it})]$, which is exactly Robinson's estimator for the partially linear model.

Now that we have our fixed effects estimates ($\hat{\delta}_{-1}$ as above and $\hat{\delta}_1 = -\hat{\delta}_2 - \ldots - \hat{\delta}_N$), the kernel regression estimate of the function $m(\cdot)$ follows simply,

$$\hat{m}(x) = \textstyle\sum_{i=1}^N \sum_{t=1}^T w^x_{it}(y_{it} - \hat{\delta}_i)$$

$$\text{where } w^x_{it} = \frac{\frac{1}{NT} K_h(x - x_{it})}{\frac{1}{NT}\sum_{k=1}^N \sum_{r=1}^T K_h(x - x_{kr})} \text{ are kernel weights.}$$

Just as first differencing eliminates the fixed effects, this procedure can also be thought of as a transformation to eliminate the fixed effects. By vectorizing once again, we can observe how this elimination works. Let $P_D = D^1[D^{1\prime}(I - W)'\Lambda(I - W)D^1]^{-1}D^{1\prime}(I - W)'\Lambda(I - W)$, $Q_D = I - P_D$, and w^x is the $NT \times 1$ vector with elements w^x_{it}.

$$\begin{aligned}
\hat{m}(x) &= w^{x\prime}(y - D^1\hat{\delta}_{-1}) \\
&= w^{x\prime}\{I - D^1[D^{1\prime}(I - W)'\Lambda(I - W)D^1]^{-1}D^{1\prime}(I - W)'\Lambda(I - W)\}y \\
&= w^{x\prime}Q_D y \\
&= w^{x\prime}Q_D(m(X) + D^1\delta_{-1} + \Delta_1 + \eta) \\
&= w^{x\prime}Q_D(m(X) + \Delta_1 + \eta),
\end{aligned}$$

where Δ_1 is an $NT \times 1$ vector with the first T entries equal to $\delta_1 + \delta_2 + \ldots + \delta_N$ and the remaining entries equal to zero. The last equality is the key one that actually shows the elimination of the fixed effects term, $D^1\delta_{-1}$; it occurs because $D^1 \in$ nullspace(Q_D). But this orthogonality is exactly what eliminates the fixed effects, just as differencing does in linear panel estimation.[9]

[9] Strictly speaking, the fixed effects are not completely eliminated since the Δ_1 term contains the δ's. This fact does not cause a problem as they enter only as a sum which we know converges to zero (by the identification condition) and all but the first T rows of Δ_1 are identically zero.

An interesting efficiency consideration comes from applying our difference from means estimator to the model $y_{it} = \beta + \delta_i + \eta_{it}$. From the minimization in (1.6), we have $\hat{\delta}_i = (\frac{1}{T}\sum_t y_{it}) - (\frac{1}{NT}\sum_i\sum_t y_{it})$, which leads to $\hat{\beta} = \frac{1}{NT}\sum_i\sum_t y_{it}$. But $\hat{\beta}$ is exactly the same estimator that would come from using linear within estimation on this model. So in this example there is no efficiency lost by our nonparametric difference from means estimator relative to the usual linear difference from means estimator. More generally, let $B = \{b_1, \ldots, b_p\}$, where $b_j \in \Re^k$ and $p < \infty$, be a discrete set. If for all i and t, $x_{it} \in B$, then estimation of $m(\cdot)$ reduces to a parametric problem. As above, for a small enough choice of bandwidth, $\hat{m}(a_j)$ from the *nonparametric* difference from means estimator is exactly equal to the estimate from the *linear* difference from means method. Hence as desired, no efficiency is lost in this dummy variables framework.

Now we turn to difference from means series estimation. First, we show asymptotic normality of the estimator that uses the within transformation of the series terms. Second, we show that, under certain conditions on the trimming function, the use of a series smoother in our general difference from means procedure results in an estimator that is only an additive constant away from the simple within series estimator. This latter outcome provides further elucidation of the analogy between the general nonparametric estimation procedure outlined in this section and within estimation.

To describe simple within series estimation, let $\tilde{d}$ denote the difference from means transformation (e.g. $\tilde{d}y_{it} = y_{it} - \frac{1}{T}\sum_{s=1}^{T} y_{is}$). Let $q^K(x) = (q_{1K}(x), \ldots, q_{KK}(x))'$ be a $K \times 1$ vector of approximating functions for m. Define for each j, $\tilde{p}_{jK}(x_{it}) = q_{jK}(x_{it}) - \frac{1}{T}\sum_{s=1}^{T} q_{jK}(x_{is})$, then $\tilde{p}^K(x_{it}) = (\tilde{p}_{1K}(x_{it}), \ldots, \tilde{p}_{KK}(x_{it}))'$. Our within series estimator will be $\hat{m}_W(x) = q^K(x)'\tilde{\beta}$, where $\tilde{\beta}$ is obtained from least squares regression of $\tilde{d}y_{it}$ on $\tilde{p}^K(x_{it})$,

$$\tilde{\beta} = \left[\sum_{i=1}^{N}\sum_{t=2}^{T} \tilde{\lambda}(x_i)\tilde{p}^K(x_{it})\tilde{p}^K(x_{it})'\right]^{-1} \sum_{i=1}^{N}\sum_{t=2}^{T} \tilde{\lambda}(x_i)\tilde{p}^K(x_{it})\tilde{d}y_{it}.$$

Again $\tilde{\lambda}(\cdot)$ is a fixed trimming function.

Variance estimation is as in least squares with fixed K. Let $\tilde{dy}_{\cdot} = (\tilde{dy}_{i1}, \ldots, \tilde{dy}_{iT})'$. Now define $\tilde{p}_T^K(x_i) = (\tilde{p}^K(x_{i1}), \ldots, \tilde{p}^K(x_{iT}))$ and $\Sigma_W = E[\tilde{\lambda}(x_i)\tilde{p}_T^K(x_i)\tilde{\Omega}(x_i)\tilde{p}_T^K(x_i)']$, where $\tilde{\Omega}(x_i) = \mathrm{Var}(\tilde{dy}_i|x_i)$. Then,

$$\begin{aligned} V_W^K &= q^K(x)\left[E(\tilde{\lambda}(x_i)\tilde{p}_T^K(x_i)\tilde{p}_T^K(x_i)')\right]^{-1}\Sigma_W \\ &\quad \cdot \left[E(\tilde{\lambda}(x_i)\tilde{p}_T^K(x_i)\tilde{p}_T^K(x_i)')\right]^{-1} q^K(x)'. \end{aligned}$$

Before giving the variance estimator, define $\hat{\Sigma}_W = \frac{1}{N}\sum_{i=1}^N \tilde{\lambda}(x_i)\tilde{p}_T^K(x_i)\tilde{d}\hat{\varepsilon}_i\tilde{d}\hat{\varepsilon}_i'\tilde{p}_T^K(x_i)'$, where $\hat{\varepsilon}_{it} = y_{it} - \hat{m}_W(x_{it})$ and $\tilde{d}\hat{\varepsilon}_i = (\tilde{d}\hat{\varepsilon}_{i1}, \ldots, \tilde{d}\hat{\varepsilon}_{iT})'$. Then,

$$\begin{aligned} \hat{V}_W^K &= q^K(x)'\left[\sum_{i=1}^N\sum_{t=1}^T \tilde{\lambda}(x_i)\tilde{p}^K(x_{it})\tilde{p}^K(x_{it})'\right]^{-1}\hat{\Sigma}_W \\ &\quad \left[\sum_{i=1}^N\sum_{t=1}^T \tilde{\lambda}(x_i)\tilde{p}^K(x_{it})\tilde{p}^K(x_{it})'\right]^{-1} q^K(x). \end{aligned}$$

Again, Assumptions S-1 and S-3 can be used to prove asymptotic normality of our within series estimator. Assumption S-2 is modified only slightly to apply to the trimming function used here. Let $f(x_i)$ denote the density of x_i.

Assumption S-2' *$\tilde{\lambda}(x_i)$ is bounded and zero except on a compact set $\tilde{\mathcal{X}}$ where $f(x_i)$ is bounded away from zero for $t = 2, \ldots, T$.*

If Assumption S-3 is satisfied, then, when restricting attention to power series and regression splines, Assumption A-3' in the Appendix holds. Thus there exist $\tilde{\alpha}$, $\tilde{d}$, $(\tilde{\beta}_K)$, $(\tilde{\gamma}_K)$ such that $|m - \tilde{\gamma}_K - q^{K\prime}\tilde{\beta}_K|_{\tilde{d}} = O(K^{-\tilde{\alpha}})$.

Theorem 4 *If Assumptions S-1, S-2', and S-3 are satisfied, the approximating functions $\{q_{kK}\}$ are either power series or regression splines, $\sqrt{N}K^{-\tilde{\alpha}} \longrightarrow 0$, $\tilde{\zeta}_0(K)^2K/N \longrightarrow 0$, then*

$$\sqrt{N}(V_W^K)^{-1/2}[(\hat{m}_W - m(x)) + \tilde{\gamma}_K] \xrightarrow{d} N(0, I)$$

and

$$\sqrt{N}(\hat{V}_W^K)^{-1/2}[(\hat{m}_W - m(x)) + \tilde{\gamma}_K] \xrightarrow{d} N(0, I).$$

Now consider plugging a series smoother into our general estimator. Let $\breve{q}^K(x) = (1, q^K(x)')'$ be the approximating functions with a constant term added to the approximating functions used above. The first step is

$$\{\breve{\delta}_1, \ldots, \breve{\delta}_N\} \in \underset{\delta_1, \ldots, \delta_N}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} \lambda_{it}[(y_{it} - \delta_i) - \breve{q}^K(x_{it})'\{\sum_{j=1}^{N}\sum_{s=1}^{T} \lambda_{js}\breve{q}^K(x_{js})$$
$$\cdot \breve{q}^K(x_{js})'\}^{-1}\{\sum_{j=1}^{N}\sum_{s=1}^{T} \lambda_{js}\breve{q}^K(x_{js})(y_{js} - \delta_j)\}]^2$$

$$\text{s.t.} \sum_{i=1}^{N} \delta_i = 0.$$

Then letting $\breve{Q} = (\breve{q}^K(x_{11}), \ldots, \breve{q}^K(x_{NT}))'$, we can vectorize and solve for $\breve{\delta}_{-1}$. In the second step, we obtain our estimator of m, $\breve{m}_W(x) = \breve{q}^K(x)'\breve{\gamma}$, where $\breve{\gamma} = (\breve{Q}'\Lambda\breve{Q})^{-1}\breve{Q}'\Lambda(y - D^{1\prime}\breve{\delta}_{-1})$.

From least squares, we know that $\breve{\gamma}$ can be expressed in another way. Let U $= I - \Lambda^{1/2}D^1(D^{1\prime}\Lambda D^1)^{-1}D^{1\prime}\Lambda^{1/2}$; then $\breve{\gamma} = (\breve{Q}'\Lambda^{1/2}U\Lambda^{1/2}\breve{Q})^{-1}\breve{Q}'\Lambda^{1/2}U\Lambda^{1/2}y$. Let $\breve{\gamma} = \begin{pmatrix} \breve{\rho} \\ \breve{\beta} \end{pmatrix}$ with $\breve{\rho}$ the scalar constant term and $\breve{Q} = (e_{NT}, Q)$. Hence, $\breve{m}_W(x) = \breve{\rho} + q^K(x)'\breve{\beta}$. Then as in the linear model, we can give an expression for $\breve{\beta}$ from the partitioned matrix inverse formula, $\breve{\beta} = (Q'H'HQ)^{-1}Q'H'Hy$, where $H = G\Lambda^{1/2}$ and $G = (U - U\Lambda^{1/2}e_{NT}(e_{NT}'\Lambda^{1/2}U\Lambda^{1/2}e_{NT})^{-1}\ e_{NT}'\Lambda^{1/2}U)$. Fortunately, G can be expressed quite simply as the identity minus a block diagonal matrix with N blocks of size $T \times T$. If $\mu_i = (\sqrt{\lambda_{i1}}, \ldots, \sqrt{\lambda_{iT}})'$, then the i^{th} block is $(\sum_t \lambda_{it})^{-1}(\mu_i\mu_i')$. Thus we can express $\breve{\beta}$ as the coefficient estimate from a regression of Hy on HQ. From our expression for G, we can see that, as a transformation, H is very similar to a standard difference from means. Specifically, an element of the H-transformed y variable is $[Hy]_{it} = \sqrt{\lambda_{it}}(y_{it} - (\sum_{s=1}^{T} \lambda_{is})^{-1}\sum_{s=1}^{T} \lambda_{is}y_{is})$. Given this expression for the H-transformed variables, we have proven the following assertion:

Claim 1 *If the trimming function $\lambda(x_{it})$ is chosen to be the trimming function from the within series estimation above ($\lambda(x_{it}) = \tilde{\lambda}(x_{it})$) and $\lambda(x_{it}) = 0$ or 1 for all i and t, then $\breve{m}_W(x) = \breve{\rho} + \tilde{m}_W(x)$.*

From Claim 1, we learn that, under reasonable conditions on the trimming function, the general difference from means procedure using a series smoother is the same (up to an additive constant) as the within series estimation. Also, from the general expression for the H transformation, we gain further intuition on how the general estimation procedure given in this section is just an extension of within estimation.

1.5 Extensions

In many applications, generalizations of the model (1.1) may be of primary interest. Following we will describe without proof the modifications of our estimators necessary to handle some of the extensions most likely to be encountered.

1.5.1 Time-varying Unknown Function

We consider the generalization of (1.1) to a model with a time-varying unknown function.

$$y_{it} = m_t(x_{it}) + \delta_i + \eta_{it} \tag{1.7}$$

Modification of the partial means first difference estimator is straightforward. Define $\bar{l}_t(x_1, x_2) = S^{(x_1,x_2)}(dy_t|X_t, X_{t-1})$. To estimate $m_t(\cdot)$ in the second step, we must partial out the second argument of $\bar{l}_t$ from the $y_{it} - y_{i,t-1}$ equation ($\bar{m}_{1t}$) or partial out the first argument of the $y_{i,t+1} - y_{it}$ equation ($\bar{m}_{2,t+1}$). So we define the time-varying partial means estimators to be $\bar{m}_{1t} = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\bar{l}_t(x, x_{js})$ and $\bar{m}_{2t} = -\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\bar{l}_t(x_{js}, x)$. Note that the conditions imposed in defining $\tilde{l}$, the restricted estimator, no longer hold. Specifically, now $l_t(x_1, x_2) = m_t(x_1) - m_{t-1}(x_2) \neq -l_t(x_2, x_1)$ and it follows that $l_t(x, x) \neq 0$. As a result, we must use $\bar{m}_{1t}$ and $\bar{m}_{2,t+1}$ rather than the restricted estimator $.5\bar{m}_{1t} + .5\bar{m}_{2t}$. For kernel smoothers, asymptotic normality follows as in section 1.3 from the kernel partial means theory in the Appendix.[10] Similarly, the distributional results for series smoothers also generalize to this case.

[10] Lemma 3 is the crucial one for proving asymptotic normality. Its proof is sufficiently general to handle the case of averaging across time-varying estimators.

The difference from the means estimator is quite simple to modify for the model (1.7). Note that in the second step when estimating $m_t(x)$, using the estimated fixed effects from the first stage, we only want to use time period t observations, so $m(x) = S^x(\hat{y}_t^*|X_t)$. Similarly, when estimating the fixed effects in the first step, $S^{it}(y^*(\delta)|X)$ is replaced by $S^{it}(y_t^*(\delta)|X_t)$.

Thus, for a kernel smoother,

$$\begin{aligned} \hat{m}_t(x) &= \textstyle\sum_{i=1}^N w_{it}^x(y_{it} - \hat{\delta}_i) \\ \text{where } w_{it}^x &= \frac{\frac{1}{N}K_\sigma(x-x_{it})}{\frac{1}{N}\sum_{k=1}^N K_\sigma(x-x_{kt})}. \end{aligned} \tag{1.8}$$

In the first step, we now zero out all but own time period weights so that

$$w_{js}^{it} = \begin{cases} \frac{\frac{1}{N}K_\sigma(x_{it}-x_{jt})}{\frac{1}{N}\sum_{k=1}^N K_\sigma(x_{it}-x_{kt})} & \text{if } s = t \\ 0 & \text{if } s \neq t. \end{cases}$$

One particular form of a time-varying function is $m_t(\cdot) = m(\cdot) + \alpha_t$. With additive time effects, interest again centers on estimation of m. Similar to the discussion concerning condition (1.3), we again face an identification problem without restricting the time effect in some way. From the following condition, we can interpret the time effects as mean zero deviations from the true regression function

$$\sum_{t=1}^T \alpha_t = 0. \tag{1.9}$$

The first difference partial means estimators are consistent up to an addivitive constant, so the time effects merely become part of that constant. The estimators given above are still valid. Specifically, $\bar{m}_{1t}(x)$ estimates $\underline{m}(x) + \alpha_t - \alpha_{t-1}$.

For the difference from means type estimator, we could use the time-varying estimators above, e.g. $\hat{m}(x) = \frac{1}{T}\sum_{t=1}^T \hat{m}_t(x)$, with $\hat{m}_t(x)$ from (1.8). This estimator averages across the time effects which should eliminate them by (1.9). Alternatively, the time effects could be estimated jointly with the individual effects in the first stage. Define $y_{it}^{**} = y_{it} - \delta_i - \alpha_t$, which we sometimes write as an explicit function of δ and

α to emphasize the dependence.

$$\{\hat{\delta},\hat{\alpha}\} \in \underset{\delta_1,\ldots,\delta_N,\alpha_1,\ldots,\alpha_T}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it}-\delta_i-\alpha_t)-\sum_{j=1}^{N} S^{it}(y^{**}(\delta,\alpha)|X)]^2$$

$$\text{s.t.} \sum_{i=1}^{N}\delta_i = 0 \quad \text{and} \quad \sum_{t=1}^{T}\alpha_t = 0$$

where the T-vector of estimated time effects is $\hat{\alpha} = (\hat{\alpha}_1,\ldots,\hat{\alpha}_T)$ and $\hat{m}(x) = S^x(\hat{y}^{**}|X)$. This method is described in more detail in section 1.5.3.

1.5.2 Lagged Regressors

Now we add lagged regressors to (1.1),

$$y_{it} = m(x_{it}, x_{i,t-1}) + \delta_i + \eta_{it}.$$

The only change required in the partial means estimators is the use of second differences. Then partial out with respect to one of the arguments of $m(\cdot,\cdot)$. Clearly, if more lags were added, then higher differences could be used. The second (or higher) differences estimator theory is the same as the first difference, only requiring $T \geq 3$ instead of $T \geq 2$.

The difference from means estimator remains the same treating $(x_{it}, x_{i,t-1})$ as the argument to $m(\cdot)$. So if x_{it} is r-dimensional, then we would use a nonparametric smoother on $2r$-dimensional regressor space.

1.5.3 Partially Linear Model

Finally we add a linear component to (1.1),

$$y_{it} = m(x_{it}) + z_{it}'\beta + \delta_i + \eta_{it}.$$

Taking first differences,

$$dy_{it} = l(x_{it}, x_{i,t-1}) + dz'_{it}\beta + d\eta_{it}.$$

Then Robinson's (1988) estimator gives a $\sqrt{N}$-consistent estimator of β, $\hat{\beta}$. Next, set $\tilde{y}_{it} = y_{it} - z'_{it}\hat{\beta}$ and proceed as in section 1.3 with $\tilde{y}_{it}$ as the dependent variable. Since $\hat{\beta}$ is $\sqrt{N}$-consistent, there is no effect on the standard error estimates for any of the partial means estimators given there.

As mentioned in section 1.4, the objective function that yields our first step fixed effects estimates in the difference from means procedure is the same objective function that gives Robinson's estimator. Thus it is simple to alter the difference from means estimator to include the linear component. We estimate the fixed effects and the parameters of the linear component jointly in the first step. For example with the kernel regression smoother,

$$\{\hat{\beta}, \hat{\delta}_1, \ldots, \hat{\delta}_N\} \in \underset{\beta, \delta_1, \ldots, \delta_N}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it} - Z'_{it}\beta - \delta_i) - \sum_{j=1}^{N}\sum_{s=1}^{T} w^{it}_{js}(y_{js} - Z'_{it}\beta - \delta_j)]^2$$

$$\text{s.t.} \sum_{i=1}^{N} \delta_i = 0.$$

The second step estimation of m then becomes

$$\hat{m}(x) = \sum_{i=1}^{N}\sum_{t=1}^{T} w^{x}_{it}(y_{it} - Z'_{it}\hat{\beta} - \hat{\delta}_i).$$

1.6 Conclusion

We have found that the within and first difference transformations that eliminate the individual effect in a linear panel model have analogs in the nonparametric panel extension of the linear model that also eliminate the individual effect. These two transformations then lead to two corresponding series estimators, and we are able to present more general nonparametric estimators that are applicable to any nonparametric smoother, not just series. In the first difference case, we use the partial means idea to provide a general first difference estimation scheme that works with any nonparametric regression estimator. In the difference from means approach, we obtain our estimator by employing an objective function analogous to the one that produces Robinson's (1988) estimator in the partially linear model. We showed that the general difference from means procedure reduces to within estimation when applied to the linear model and reduces to simple series estimation on within-transformed approximating functions in the general nonparametric model.

An empirical example in Chapter 2 will show the potential usefulness of these estimators. In cases where the linear panel model is rejected, the more flexible model (1.1) might be useful in analysis and might also avoid arbitrary parametric assumptions often unsupported by theory. In our investment example, an examination of the linear panel estimates and multivariate regression results will lead us to reject the linear model. The application in Chapter 2 also presents the opportunity to compare different estimation techniques. For instance, we can confirm the similarity of series estimation by partial means and by simple series first differencing.

As larger panels of data become increasingly available to economists, the desire for nonparametric panel methods will likely increase. The model analyzed in this paper is a natural starting point for nonparametric panel empirical work. We have found that, as in the linear panel model, there is more than one consistent estimator of the regression function. These estimators are straightforward to use and allow the researcher to estimate nonlinear relationships in the panel data framework without parametric assumptions on the shape of that relationship. Thus, these estimators will be useful in a wide variety of applications, in addition to our investment example.

Appendix

Series

Andrews (1991) and Newey (1994a) give general results proving asymptotic normality of linear functionals of series estimators. We will present a multivariate version of Newey's result for scalar functionals of series estimators. We begin with some multivariate regression notation. Let Y be the dependent variable and X an R-dimensional vector of regressors.[11] We observe J values of (Y, X) for each individual, and the data, $(Y_{11}, X_{11}), \ldots, (Y_{NJ}, X_{NJ})$, generated by the model $G_0(X) = E[Y|X]$. We assume that $E[Y_{ij}|X_{i1}, \ldots, X_{iJ}] = E[Y_{ij}|X_{ij}] = G_0(X_j)$ and that J is fixed and finite. We suppose that interest centers on estimation of a scalar linear functional, $\theta_0 = a(G_0)$.

To carry out series estimation, we need a sequence of approximating functions $\{q_{kK}(\cdot)\}$. Let $q^K(X) = (q_{1K}(X), \ldots, q_{KK}(X))'$ be the $K \times 1$ vector of approximating functions. Series estimation of G_0 is accomplished by a simple least squares projection on the approximating functions,

$$\hat{G}(\cdot) = q^K(\cdot)'\hat{\beta}_K,$$

$$\text{where } \hat{\beta}_K = [\sum_{i=1}^{N}\sum_{j=1}^{J} q^K(X_{ij})q^K(X_{ij})']^{-1}[\sum_{i=1}^{N}\sum_{j=1}^{J} q^K(X_{ij})Y_{ij}].$$

Our estimate of θ_0 is $\hat{\theta} = a(\hat{G})$. If we define $A = (a(q_{1K}), \ldots, a(q_{KK}))$, then $\hat{\theta} = A\hat{\beta}_K$.

Variance estimation follows similarly as in multivariate least squares with K fixed. Define $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$ and $X_i = (X_{i1}, \ldots, X_{iJ})'$. Let $q_J^K(X_i) = (q^K(X_{i1}), \ldots, q^K(X_{iJ}))$. The variance of the functional estimator and the variance estimate are

$$\begin{aligned} V_K &= A\left[E(q_J^K(X_i)q_J^K(X_i)')\right]^{-1} E(q_J^K(X_i)\text{Var}(Y_i|X_i)q_J^K(X_i)') \\ &\quad \cdot \left[E(q_J^K(X_i)q_J^K(X_i)')\right]^{-1} A' \end{aligned}$$

[11] We use capital letters here to distuingish this notation from the panel notation used throughout the remainder of the paper.

$$\begin{aligned}\hat{V}_K &= A\left[\sum_{i=1}^{N} q_J^K(X_i)q_J^K(X_i)'/N\right]^{-1}\left[\sum_{i=1}^{N} q_J^K(X_i)\hat{\varepsilon}_i\hat{\varepsilon}_i' q_J^K(X_i)'/N\right] \\ &\quad \cdot \left[\sum_{i=1}^{N} q_J^K(X_i)q_J^K(X_i)'/N\right]^{-1} A',\end{aligned}$$

where $\hat{\varepsilon}_i = Y_i - \hat{G}(X_i)$ and $\hat{G}(X_i) = (\hat{G}(X_{i1}), \ldots, \hat{G}(X_{iJ}))'$.

Given this multivariate series notation, we now give the assumptions for the asymptotic distribution theorem.

Assumption A-1 *$(Y_1, X_1), \ldots, (Y_N, X_N)$ are i.i.d., $E[\| Y_i - G_0(X_i) \|^4 |X]$ is bounded, and the smallest eigenvalue of $Var(Y_i|X_i)$ is bounded away from zero.*

Assumption A-2 *For every K there is a nonsingular constant matrix B_K such that for $Q^K(X_i) = B_K q^K(X_i)$; i) the smallest eigenvalue of $E[Q_J^K(X_i)Q_J^K(X_i)']$ is bounded away from zero uniformly in K; ii) there is a sequence of constants $\zeta_0(K)$ satisfying $\sup_{X\in\mathcal{X}} \| Q_J^K(X) \| \leq \zeta_{d0}(K)$.*

Assumption A-3 *There is $d, \alpha, (\beta_K)$ such that $|a(G)| \leq |G|_d$, $|G_0 - q^{K\prime}\beta_K|_d = O(K^{-\alpha})$.*

Assumption A-4 *$a(G)$ is a scalar functional and there exists $(\tilde{\beta}_K)$ such that for $G_K(X) = q^K(X)'\tilde{\beta}_K$, $E[G_K(X)^2] \longrightarrow 0$ and $|a(G_K)|$ is bounded away from zero.*

Lemma 1 *If Assumptions A-1 - A-4 are satisfied, $\zeta_0(K)^2 K/n \longrightarrow 0$, and $\sqrt{N}K^{-\alpha} \longrightarrow 0$, then*

$$\sqrt{N}V_K^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I), \quad \sqrt{N}\hat{V}_K^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I).$$

> **PROOF of Lemma 1:** This lemma is a simple extension of Newey (1994a) Theorem 1. With only slight modifications to include the multivariate notation, Newey's proof shows that the above assumptions are sufficient for the conclusion of the lemma. The eigenvalue part of Assumption A-1 is used to show $V_K = E(Aq_J^K(X_i)\mathrm{Var}(Y_i|X_i)q_J^K(X_i)'A') \geq CE(Aq_J^K(X_i)q_J^K(X_i)'A') = CAA' = C \| A \|^2$, since we have chosen $q_J^K(x)$ to be transformed by a nonsingular matrix so that $E(q_J^K(X_i)q_J^K(X_i)') = I$. Also, proving the second assertion for convergence with an estimated

variance requires some generalization. The step that requires significant modification is the one showing that, for a vector v with $\| v \| < C$, $\hat{u} = |v'[\frac{1}{N}\sum_i q_J^K(X_i)(\hat{\varepsilon}_i\hat{\varepsilon}_i' - \varepsilon_i\varepsilon_i')q_J^K(X_i)']v| \xrightarrow{p} 0$. Note that $\hat{\varepsilon}_i\hat{\varepsilon}_i' - \varepsilon_i\varepsilon_i' = \varepsilon_i(\hat{G}(X_i) - G_0(X_i))' + (\hat{G}(X_i) - G_0(X_i))\varepsilon_i' + (\hat{G}(X_i) - G_0(X_i))(\hat{G}(X_i) - G_0(X_i))'$. Let the absolute value of a vector denote the vector of absolute values of its elements. Define $\hat{S} = \frac{1}{N}\sum_i q_J^K(X_i)(|\varepsilon_i|e_J' + e_J|\varepsilon_i|')q_J^K(X_i)'$, $S = E[q_J^K(X_i)(|\varepsilon_i|e_J' + e_J|\varepsilon_i|')q_J^K(X_i)'] = E[q_J^K(X_i)E[(|\varepsilon_i|e_J' + e_J|\varepsilon_i|')|X_i] q_J^K(X_i)']$ and note the symmetry of the conditional expectation in the middle and the fact that it is bounded. Hence $|v'Sv| \leq C \| v \|^2$, and by an argument similar to the one showing that $\| \hat{Q} - I \| \xrightarrow{p} 0$, where $\hat{Q} = \frac{1}{N}\sum_i q_J^K(X_i)q_J^K(X_i)'$. It follows that $\| \hat{S} - S \| \xrightarrow{p} 0$. We already have $\max_{i \leq N} \| \hat{G}(X_i) - G_0(X_i) \| = o_p(1)$, so $\hat{u} \leq o_p(1)|v'(\hat{S} + \hat{Q} - S - I)v| + o_p(1)|v'Sv| + o_p(1) \| v \|^2 \leq o_p(1) \| v \|^2 (\| \hat{S} - S \| + \| \hat{Q} - I \|) + o_p(1) \xrightarrow{p} 0$. The remainder of Newey's proof holds as is. □

Now we modify the multivariate result to handle our first difference and difference from means series estimators. Let D represent either the first differencing operator ($DY_{ij} = Y_{ij} - Y_{i,j-1}$, $Df(X_{ij}) = f(X_{ij}) - f(X_{i,j-1})$, and $DY_{i1} = Df(X_{i1}) = 0$, for any function f of X) or the difference from means operator ($DY_{ij} = Y_{ij} - \frac{1}{J}\sum_{l=1}^J Y_{il}$, $Df(X_{ij}) = f(X_{ij}) - \frac{1}{J}\sum_{l=1}^J f(X_{is})$). The framework for Lemma 2 is more restrictive than Lemma 1's framework in the sense that interest here lies only in the linear functional equal to the function evaluated at a point. On the other hand, we have generalized the framework, since the series estimation in the first step estimates the function DG but the functional of interest is $G(x)$. The subtle point here is that, in general, there does not exist a functional a such that $G(X) = a(DG)$. Thus, Lemma 1 can not be applied.

Since a constant term will be eliminated by the differencing transform, suppose $q^K(X)$ does not contain a constant term and let $\tilde{q}^K(X) = (1, q^K(X)')'$. Now $\theta_0 = G_0(X)$, which will be estimated to within an additive constant, and $\tilde{\theta} = q^K(X)'\tilde{\beta}_K$, $\tilde{\beta}_K = [\sum_{i,j} Dq^K(X_{ij})\ Dq^K(X_{ij})']^{-1} \sum_{i,j} Dq^K(X_{ij})DY_{ij}$. Let $Dq_J^K(X_i) = (Dq^K(X_{i1}), \ldots, Dq^K(X_{iJ}))$ and $DY_i = (DY_{i1}, \ldots, DY_{iJ})'$. Variance estimation is as expected,

$$\Omega_K = q^K(X)'E[Dq_J^K(X_i)Dq_J^K(X_i)']E[Dq_J^K(X_i)\mathrm{Var}(DY_i|X_i)Dq_J^K(X_i)']$$
$$\cdot E[Dq_J^K(X_i)Dq_J^K(X_i)']q^K(X)$$

and

$$\tilde{\Omega}_K = q^K(X)'[\frac{1}{N}\sum_{i=1}^{N} Dq_J^K(X_i)Dq_J^K(X_i)'][\frac{1}{N}\sum_{i=1}^{N} Dq_J^K(X_i)D\tilde{\varepsilon}_i D\tilde{\varepsilon}_i' Dq_J^K(X_i)']$$
$$[\frac{1}{N}\sum_{i=1}^{N} Dq_J^K(X_i)Dq_J^K(X_i)']q^K(X)$$

where $D\tilde{\varepsilon}_{ij} = DY_{ij} - D\tilde{G}(X_{ij})$ and $D\tilde{\varepsilon}_i = (D\tilde{\varepsilon}_{i1}, \ldots, D\tilde{\varepsilon}_{iJ})'$. Assumptions A-1 - A-4 need only be modified to fit our slightly changed setting.

Assumption A-1' *$(Y_1, X_1), \ldots, (Y_N, X_N)$ are i.i.d., $E[\| DY_i - DG_0(X_i) \|^4 |X_i]$ is bounded, and the smallest eigenvalue of $Var(DY_i|X_i)$ is bounded away from zero.*

Assumption A-2' *For every K there is a nonsingular constant matrix B_K such that for $DQ^K(X_i) = B_K Dq^K(X_i)$; i) the smallest eigenvalue of $E[DQ_J^K(X_i)DQ_J^K(X_i)']$ is bounded away from zero uniformly in K; ii) there is a sequence of constants $\zeta_0(K)$ satisfying* $\sup_{X\in\mathcal{X}} \| DQ_J^K(X) \| \leq \zeta_0(K)$.

Assumption A-3' *There is $d, \alpha, (\gamma_K), (\beta_K)$ such that $|G_0 - \gamma_K - q^{K'}\beta_K|_d = O(K^{-\alpha})$.*

Assumption A-4' *There exists $(\tilde{\beta}_K)$ such that for $DG_K(X) = Dq^K(X)'\tilde{\beta}_K$, $E[DG_K(X)^2] \longrightarrow 0$ and $|DG_K(X)|$ is bounded away from zero.*

Lemma 2 *If Assumptions A-1', A-2', A-3', and A-4' are satisfied, $\zeta_0(K)^2K/n \longrightarrow 0$, and $\sqrt{N}K^{-\alpha} \longrightarrow 0$, then*

$$\sqrt{N}\Omega_K^{-1/2}(\tilde{\theta} - \theta_0 + \gamma_K) \xrightarrow{d} N(0, I), \quad \sqrt{N}\hat{V}_K^{-1/2}(\tilde{\theta} - \theta_0 + \gamma_K) \xrightarrow{d} N(0, I).$$

PROOF of Lemma 2: Under the Assumptions A-1',A-2', A-3', and A-4' the proof here follows the proof for Newey (1994a) Theorem 1 with only slight modifications. Along with the generalizations to the multivariate setting given in the proof of Lemma 1, the other part of the proof that requires significant modification is the decomposition of $\sqrt{N}V_K{}^{-1/2}(\hat{\theta}-\theta_0)$ into terms that converge to zero in probability and a term that converges in distribution to $N(0,1)$. In the present setting, that decomposition can also be accomplished. Let $DQ = (Dq^K(X_1), \ldots, Dq^K(X_N))'$, $DY = (DY_1', \ldots, DY_N')'$, and DG and $D\varepsilon$ similarly defined. Using the sequences

γ_K and β_K in Assumption A-3', define $G_K(X) = \gamma_K + q^K(X)'\beta_K$. Then, $DG_K(X) = Dq^K(X)'\beta_K$.

$$\begin{aligned}
&\sqrt{N}\Omega_K^{-1/2}(\tilde{\theta} - \theta_0 + \gamma_K) \\
&= \sqrt{N}\Omega_K^{-1/2}(q^K(X)'\tilde{\beta} - \theta_0 + \gamma_K) \\
&= \Omega_K^{-1/2}q^K(X)'(DQ'DQ/N)^{-1}DQ'(DG + D\varepsilon - DG_K)/\sqrt{N} \\
&\quad +\sqrt{N}\Omega_K^{-1/2}q^K(X)'\beta_K + \sqrt{N}\Omega_K^{-1/2}(\gamma_K - \theta_0) \\
&= \Omega_K^{-1/2}q^K(X)'DQ'D\varepsilon/\sqrt{N} + \Omega_K^{-1/2}q^K(X)'((DQ'DQ/N)^{-1} - I) \\
&\quad \cdot DQ'D\varepsilon/\sqrt{N} + \sqrt{N}\Omega_K^{-1/2}q^K(X)'(DQ'DQ/N)^{-1}DQ'(DG - DQ\beta_K) \\
&\quad +\sqrt{N}\Omega_K^{-1/2}(G_K(X) - G_0(X))
\end{aligned}$$

Note that the first term is asymptotically normal and the other terms converge to zero. □

PROOF of Theorem 1: We show that the Assumptions of Lemma 2 hold and then the conclusion follows. Assumption S-1 and Holder's Inequality show that the fourth moment condition of Assumption A-1' is satisfied. The smallest eigenvalue condition of A-1' is satisfied by the analogous condition in S-1 and by noting that there exists a matrix A such that $\text{Var}(DY_i|X_i) = A\text{Var}(Y_i|X_i)A'$. The smallest eigenvalue condition of A-2' follows similarly from S-2. Part ii) of A-2' is satisfied since $DQ_J^K(X)$ is a fixed linear transformation of $Q_J^K(X)$. Assumption A-3' follows from the smoothness assumption S-3 and the known approximating properties of splines and polynomials. Using polynomial or spline series and the method of proof given in Newey (1994a) for the functional defined as the difference of the function at two distinct points, it is straightforward to show the existence of a sequence $(\tilde{\beta}_{tK})$ satisfying Assumption A-4'. □

PROOF of Theorem 3: Here we use Lemma 1, so we must show that Assumptions A-1 - A-4 hold. The conditional variance assumption in A-1 follows directly from the second part of Assumption S-1. The smallest eigenvalue of $E[\lambda_t(x)P^K(x)P^K(x)']$ assumption follow as above. Assumption S-2 implies Assumption A-2 ii). Assumption S-3 implies the second part of Assumption A-3. The first part of Assumption A-3 follows with $d = 0$ from the choice of functional a. Using polynomial or spline series and the method of proof given in Newey (1994a) for the functional defined as the difference of the function at two distinct points, it is straightforward to show the existence of a sequence $(\tilde{\beta}_{tK})$ satisfying Assumption A-4. □

PROOF of Theorem 4: see proof of Theorem 1. □

Kernel Partial Means

To prove Theorem 2, the following three lemmas are useful. Suppose q is a vector of variables, and (x_t, x_{t-1}) is a $2k_1$-dimensional vector of variables with density denoted $f_{t1}(\cdot,\cdot)$. Define $\bar{f}_1 = \frac{1}{T-1}\sum_{t=2}^{T} f_{t1}$. Let $h_0(x_1, x_2) = E[q|x_1, x_2]\bar{f}_1(x_1, x_2)$ and μ is a functional to be defined below. To estimate h_0, define $\hat{h}(x_1, x_2) = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T} K_\sigma((x_1, x_2) - (x_{it}, x_{i,t-1}))q_{it}$.

Assumption B-1 *For all t and for some $p \geq 4$, $E[\| q_t \|^p] < \infty$, $E[\| q_t \|^p |x]f(x)$ is bounded.*

Assumption B-2 *Suppose that $k = 2k_1$; there are matrices of functions $\omega_1(a)$ and $\omega_2(a)$ with domains $\Re^{k_1}$, $k_1 > 0$, and a fixed vector $\bar{x} \in \Re^{k_1}$, such that (i) $\mu(h) = \sum_{j=1}^{2}\mu_j(h)$, where $\mu_j(h) = \int \omega_j(a)\partial^l h(x_j(a))/\partial x^l]da$, $x_1(a) = (\bar{x}', a')'$ and $x_2(a) = (a', \bar{x}')'$; (ii) For $j = 1, 2$, $\omega_j(a)$ is bounded and continuous almost everywhere and zero outside a compact set $\mathcal{T}$; (iii) For all s,t, $\Sigma_{st}(x) = E[q_s q_t'|x]$ is continuous a.e., and for $\varepsilon > 0$, $v_t(x) = E[\| q_t \|^4 |x]$, and a convex bounded set $\tilde{\mathcal{T}}$ containing $\mathcal{T}$, $\int_{\tilde{\mathcal{T}}}\sup_{\|\eta\|\leq\varepsilon}[v_t(\bar{x}+\eta, a)f_{t1}(\bar{x}+\eta, a)]da < \infty$ and $\int_{\tilde{\mathcal{T}}}\sup_{\|\eta\|\leq\varepsilon}[v_t(a, \bar{x}+\eta)f_{t1}(a, \bar{x}+\eta)]da < \infty$ hold for $t = 2, \ldots, T$.*

Let $\tilde{\mathcal{K}}_1(u) = \int \partial^l \mathcal{K}(u, v)/\partial u^l dv$ and $\tilde{\mathcal{K}}_2(u) = \int \partial^l \mathcal{K}(v, u)/\partial u^l dv$.

$$V_{jt} = \int \omega_j(a)\left[\Sigma_{tt}(x_j(a)) \otimes \left\{\int \tilde{\mathcal{K}}_j(u)\tilde{\mathcal{K}}_j(u)'du\right\}\right]\omega_j(a)' f_{t1}(x_j(a))da \qquad (1.10)$$

The following lemma is the key to showing the asymptotic normality of Theorem 2. The conclusion is the same as in Lemma 5.3 of Newey (1994b), but the hypotheses (Assumption B-2 in particular) allows for the more general integral representation necessary for the first difference estimator given in this paper.

Lemma 3 *If Assumptions P-1, P-2, B-1, and B-2 are satisfied with $d \geq l+\zeta$, and for $\alpha = \frac{k_1}{2}+l$, $\sqrt{n}\sigma^{k_1/2} \longrightarrow \infty$ and $\sqrt{n}\sigma^{\alpha+\zeta} \longrightarrow 0$, then $\sqrt{n}\sigma^{\alpha}[\mu(\hat{h}) - \mu(h_0)] \xrightarrow{d} N(0, V)$, where $V = \sum_{t=2}^{T}(V_{1t} + V_{2t})$.*

> **PROOF:** $E\mu(\hat{h}) = \mu(E\hat{h})$ by Newey (1994b) Lemma B.4 since $\mu(h)$ is linear. $\omega_j(a)$ is bounded and zero outside $\mathcal{T}$ and by $x_1(a)$ and $x_2(a)$ bounded

on $\mathcal{T}$, $\sqrt{N}\sigma^{\alpha}[E[\mu(\hat{h})] - \mu(h_0)] = \sqrt{N}\sigma^{\alpha}[\mu(E\hat{h}) - \mu(h_0)] \leq \sqrt{N}\sigma^{\alpha}(C_1 + C_2) \parallel E\hat{h}_t - h_{t0} \parallel_l = O(\sqrt{N}\sigma^{\alpha+\varsigma}) \longrightarrow 0$ (by Newey(1994b) Lemma B.2). Therefore it suffices to show $\sqrt{N}\sigma^{\alpha}[\mu(\hat{h}) - E\mu(\hat{h})] \xrightarrow{d} N(0, V)$. Let $\mathcal{K}^l(u)$ denote $\partial^l\mathcal{K}(u)/\partial u^l$ and $\rho_{\sigma}^{j}(x_1, x_2) = \sigma^{-k-l} \int \omega_j(a)[I \otimes \mathcal{K}^l((x_j(a) - (x_1, x_2))/\sigma)]da$, where I is the identity matrix with the same dimension as q. Then letting $\sum_t$, $\sum_i$, and $\sum_j$ denote $\sum_{t=2}^{T}$, $\sum_{i=1}^{N}$, and $\sum_{j=1}^{2}$, we have $\mu(\hat{h}) = \sum_j \mu_j(\hat{h}) = \sum_i \sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}/N(T-1)$. Thus to show $\sqrt{N}\sigma^{\alpha}[\mu(\hat{h}) - E\mu(\hat{h})] \xrightarrow{d} N(0, V)$, it suffices, by the Liapunov central limit theorem, that $\sigma^{2\alpha}\mathrm{var}(\frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}) \longrightarrow V$ and $\sigma^{4\alpha}E[\parallel \frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it} \parallel^4]/n \longrightarrow 0$. Since $\sigma^{\alpha} \parallel E[\frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}] - \mu(h_0) \parallel = \sigma^{\alpha} \parallel E\mu(\hat{h}) - \mu(h_0) \parallel \longrightarrow 0$, $\sigma^{\alpha} \parallel E[\frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}] \parallel \longrightarrow 0$. Therefore to show $\sigma^{2\alpha}\mathrm{var}(\frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}) \longrightarrow V$, it suffices to show $\sigma^{2\alpha}E[(\frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it})(\frac{1}{T-1}\sum_j \sum_t \rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it})'] \longrightarrow V$. It is straightforward to show that $\sigma^{2\alpha}E[\rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}q_{it}'\rho_{\sigma}^{j}(x_{it}, x_{i,t-1})'] \longrightarrow V_{jt}$. So it will remain only to show that $\sigma^{2\alpha}E[\rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}q_{i,t-p}'\rho_{\sigma}^{j}(x_{i,t-p}, x_{i,t-p-1})'] \longrightarrow 0$ for $p \geq 1$ and $\sigma^{2\alpha}E[\rho_{\sigma}^{1}(x_{it}, x_{i,t-1})q_{it}q_{i,t-p}'\rho_{\sigma}^{2}(x_{i,t-p}, x_{i,t-p-1})'] \longrightarrow 0$ for $p \geq 0$. $\mathcal{K}(u)$ has a bounded support, $\mathcal{K}(u_1, u_2)$ is zero for all u_1 outside a bounded set $\mathcal{V}_1$ and is zero for all u_2 outside a bounded set $\mathcal{V}_2$. Let $\bar{\mathcal{T}}$ be a compact, convex set containing $\mathcal{T}$ in its interior. For small enough σ, if $x \notin \bar{\mathcal{T}}$, then $x + \sigma v \notin \mathcal{T}$ for all $v \in \mathcal{V}_1$ or $v \in \mathcal{V}_2$.

Show $\sigma^{2\alpha}E[\rho_{\sigma}^{j}(x_{it}, x_{i,t-1})q_{it}q_{it}'\rho_{\sigma}^{j}(x_{it}, x_{i,t-1})'] \longrightarrow V_{jt}$. $\rho_{\sigma}^{1}(x_1, x_2) = \sigma^{-k_1-l} \int \omega_1(x_2 + \sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_1}{\sigma}, v)]dv$, where $v = \frac{a-x_2}{\sigma}$.

$$\begin{aligned}
& \sigma^{k_1+2l}E[\rho_{\sigma}^{1}(x_{it}, x_{i,t-1})q_{it}q_{it}'\rho_{\sigma}^{1}(x_{it}, x_{i,t-1})'] \\
= \; & \sigma^{2k_1+2l}\int\int \rho_{\sigma}^{1}(\bar{x} - \sigma u_1, x_2)\Sigma_{tt}(\bar{x} - \sigma u_1, x_2)\rho_{\sigma}^{1}(\bar{x} - \sigma u_1, x_2)' \\
& \qquad f_{t1}(\bar{x} - \sigma u_1, x_2)dx_2du_1 \\
& \text{where } u_1 = \frac{\bar{x} - x_1}{\sigma} \\
= \; & \int\int \left[\int \omega_1(x_2 + \sigma v)[I \otimes \mathcal{K}^l(u_1, v)]dv\right] \Sigma_{tt}(\bar{x} - \sigma u_1, x_2) \\
& \qquad \left[\int \omega_1(x_2 + \sigma v)[I \otimes \mathcal{K}^l(u_1, v)]dv\right]' f_{t1}(\bar{x} - \sigma u_1, x_2)dx_2du_1 \\
\longrightarrow \; & \int\int \left[\int \omega_1(x_2)[I \otimes \mathcal{K}^l(u_1, v)]dv\right] \Sigma_{tt}(\bar{x}, x_2) \\
& \qquad \left[\int \omega_1(x_2)[I \otimes \mathcal{K}^l(u_1, v)]dv\right]' f_{t1}(\bar{x}, x_2)dx_2du_1 \\
& \text{as } \sigma \longrightarrow 0 \text{ by DCT}
\end{aligned}$$

Thus, $\sigma^{2\alpha}E[\rho_{\sigma}^{1}(x_{it}, x_{i,t-1})q_{it}q_{it}'\rho_{\sigma}^{1}(x_{it}, x_{i,t-1})'] \longrightarrow V_{1t}$ as $\sigma \longrightarrow 0$. Note that the analogous result for $j = 2$ follows by symmetry.

To show the remaining terms converge to zero we use an appropriately chosen substitution for the ρ_σ terms (as above) and prove they converge to a finite constant at a slower rate. To show $\sigma^{2\alpha}E[\rho_\sigma^j(x_{it}, x_{i,t-1})q_{it}q'_{i,t-1}\rho_\sigma^j(x_{i,t-1}, x_{i,t-2})'] \longrightarrow 0$. We substitute $\rho_\sigma^1(x_1, x_2) = \sigma^{-k_1-l}\int \omega_1(\bar{x}+\sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_1}{\sigma}, v+ \frac{\bar{x}-x_2}{\sigma})]dv$, where $v = \frac{a-\bar{x}}{\sigma}$ and $\rho_\sigma^1(x_2, x_3) = \sigma^{-k_1-l}\int \omega_1(x_3 + \sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_2}{\sigma}, v)]dv$, where $v = \frac{a-x_3}{\sigma}$.

$$\begin{aligned}&\sigma^{2l}E[\rho_\sigma^1(X_{it}, X_{i,t-1})q_{it}q'_{i,t-1}\rho_\sigma^1(X_{i,t-1}, X_{i,t-2})'] \\ \longrightarrow\ &\int\int\int\left[\int \omega_1(\bar{x})[I \otimes \mathcal{K}^l(u_1, v+u_2)]dv\right]\Sigma_{tt-1}(\bar{x}, \bar{x}, x_3) \\ &\left[\int_{\mathcal{V}_2} \omega_1(x_3)[I \otimes \mathcal{K}^l(u_2, v)]dv\right]' f_{x_{it}, x_{i,t-1}, x_{i,t-2}}(\bar{x}, \bar{x}, x_3)dx_3du_2du_1 \\ &\text{as } \sigma \longrightarrow 0\end{aligned}$$

But note that the rate of convergence to the finite constant is σ^{2l}. Thus, $\sigma^{k_1+2l}\ E[\rho_\sigma^1(X_{it}, X_{i,t-1})q_{it}\ q'_{i,t-1}\rho_\sigma^1(X_{i,t-1}, X_{i,t-2})'] \longrightarrow 0$ as $\sigma \longrightarrow 0$. The analogous result for $j = 2$ follows similarly. The remaining terms also converge to a constant at the rate σ^{2l}. To show $\sigma^{2\alpha}\ E[\rho_\sigma^j(x_{it}, x_{i,t-1})q_{it}\ q'_{i,t-p}\rho_\sigma^j(x_{i,t-p}, x_{i,t-p-1})'] \longrightarrow 0$, for $p \geq 2$. We make the substitution $\rho_\sigma^1(x_1, x_2) = \sigma^{-k_1-l}\int \omega_1(x_2 + \sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_1}{\sigma}, v)]dv$, where $v = \frac{a-x_2}{\sigma}$, and $\rho_\sigma^1(x_3, x_4) = \sigma^{-k_1-l}\int \omega_1(x_4 + \sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_3}{\sigma}, v)]dv$, where $v = \frac{a-x_4}{\sigma}$. To show $\sigma^{2\alpha}E[\rho_\sigma^1(x_{it}, x_{i,t-1})q_{it}q'_{it}\rho_\sigma^2(x_{it}, x_{i,t-1})'] \longrightarrow 0$. We let $\rho_\sigma^1(x_1, x_2) = \sigma^{-k_1-l}\int \omega_1(\bar{x}+\sigma v)[I\otimes\mathcal{K}^l(\frac{\bar{x}-x_1}{\sigma}, v+\frac{\bar{x}-x_2}{\sigma})]dv$, where $v = \frac{a-\bar{x}}{\sigma}$, and $\rho_\sigma^2(x_1, x_2) = \sigma^{-k_1-l}\int \omega_2(\bar{x}+\sigma v)[I\otimes\mathcal{K}^l(v+\frac{\bar{x}-x_1}{\sigma}, \frac{\bar{x}-x_2}{\sigma})]dv$, where again $v = \frac{a-\bar{x}}{\sigma}$. To show $\sigma^{2\alpha}E[\rho_\sigma^1(x_{it}, x_{i,t-1})q_{it}q'_{i,t-1}\rho_\sigma^2(x_{i,t-1}, x_{i,t-2})'] \longrightarrow 0$. We let $\rho_\sigma^1(x_1, x_2) = \sigma^{-k_1-l}\int \omega_1(x_2 + \sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_1}{\sigma}, v)]dv$, where $v = \frac{a-x_2}{\sigma}$, and $\rho_\sigma^2(x_2, x_3) = \sigma^{-k_1-l}\int \omega_2(x_2 + \sigma v)[I \otimes \mathcal{K}^l(v, \frac{\bar{x}-x_3}{\sigma})]dv$, where $v = \frac{a-x_2}{\sigma}$. To show $\sigma^{2\alpha}E[\rho_\sigma^1(x_{it}, x_{i,t-1})q_{it}q'_{i,t-p}\rho_\sigma^2(x_{i,t-p}, x_{i,t-p-1})'] \longrightarrow 0$ for $p \geq 2$. $\rho_\sigma^1(x_1, x_2) = \sigma^{-k_1-l}\int \omega_1(x_2 + \sigma v)[I \otimes \mathcal{K}^l(\frac{\bar{x}-x_1}{\sigma}, v)]dv$, where $v = \frac{a-x_2}{\sigma}$, and $\rho_\sigma^2(x_3, x_4) = \sigma^{-k_1-l}\int \omega_2(x_3 + \sigma v)\ [I \otimes \mathcal{K}^l(v, \frac{\bar{x}-x_4}{\sigma})]dv$, where $v = \frac{a-x_3}{\sigma}$.

Finally, we must show that $\sigma^{4\alpha}E[\|\ \frac{1}{T-1}\sum_j\sum_t \rho_\sigma^j(x_{it}, x_{i,t-1})q_{it}\ \|^4]/N \longrightarrow 0$, note that by the triangle inequality, $\sigma^{4\alpha}\ E[\|\ \sum_j\sum_t\ \rho_\sigma^j(x_{it}, x_{i,t-1})q_{it}\ \|^4]/n \leq \sigma^{4\alpha}\ E[(\sum_j\sum_t\ \|\ \rho_\sigma^j(x_{it}, x_{i,t-1})q_{it}\ \|)^4]/n$ and by Hölder's Inequality applied to each of the terms in this sum it suffices to show that $\sigma^{4\alpha}E[\|\ \rho_\sigma^j(x_{it}, x_{i,t-1})q_{it}\ \|^4]/n \longrightarrow 0$ for all j and t. For $j = 1$ and for all t, $\sigma^{4\alpha}E[\|\ \rho_\sigma^1(X_{it}, X_{i,t-1})q_{it}\ \|^4]/N \leq \sigma^{4\alpha}E[\|\ \rho_\sigma^1(X_{it}, X_{i,t-1})\ \|^4\|\ \ q_{it}\ \|^4]/N \leq \sigma^{4\alpha}E[\|\ \rho_\sigma^1(X_{it}, X_{i,t-1})\ \|^4\ v_t(X_{it}, X_{i,t-1})]/N = C\sigma^{3k_1+4l}\int_{\mathcal{T}}\ \|\ \rho_\sigma^j(\bar{x} - \sigma u, a)\ \ \|^4\ v_t(\bar{x} - \sigma u, a)f_{t1}(\bar{x} - \sigma u, a)dadu/n \leq C/(n\sigma_{k_1}) \longrightarrow 0$. The analogous result for $j = 2$ holds similarly. The conclusion of the lemma now holds by the Liapunov central limit theorem. □

Lemma 4 *Suppose that Assumptions P-1, P-2, and B-1 are satisfied, $\mathcal{X}$ is compact, there is a vector of functions $D(z, h)$ and nonnegative constants α, $\Delta_l \leq \Delta$ $(l = 1, 2)$, $\epsilon > 0$ such that $d \geq \max\{\Delta + 1, \Delta_1 + \zeta, \Delta + \zeta\}$, and (i) $D(z, h)$ is linear in h on*

$\{h : \| h \|_\Delta < \infty\}$; *(ii) for all* h *with* $\| h - h_0 \|_\Delta < \epsilon$, $\| \mu(z,h) - \mu(z,h_0) - D(z,h-h_0) \| \le b(z) \| h - h_0 \|_{\Delta_1} \| h - h_0 \|_{\Delta_2}$; *(iii)* $\| D(z,h) \| \le \tilde{b}(z) \| h \|_{\Delta_1}$ *and* $E[\tilde{b}(z)^4] < \infty$; *(iv) for* $\varrho_N^j = [\ln N/(N\sigma^{k+2j})]^{1/2} + \sigma^\zeta$, $\varrho_N^\Delta \longrightarrow 0$, $\sqrt{N}\sigma^a E[b(z)]\varrho_N^{\Delta_1}\varrho_N^{\Delta_2} \longrightarrow 0$, *and* $\sqrt{N}\sigma^{k+\Delta_1-a} \longrightarrow \infty$. *Then for* $\mu(h) = \int D(z,h)dF(z)$,

$$\sqrt{N}\sigma^a \sum_{i=1}^{N}[\mu(z_i,\hat{h}) - \mu(z_i,h_0)]/N = \sqrt{N}\sigma^a[\mu(\hat{h}) - \mu(h_0)] + o_p(1).$$

Assumption B-3 *(i)* $\| \mu_j(z,\beta,\tilde{h}) - \mu_j(z,\beta_0,h_0) \| \le b_j(z)[\| \beta - \beta_0 \|^\epsilon + (\| \tilde{h} - h_0 \|_\Delta)^\epsilon]$ *and* $E[b_j(z)^2] < \infty$; *(ii) For* $\epsilon > 0$ *and* $\| \beta - \beta_0 \| < \epsilon$ *and* $\| \tilde{h} - h_0 \|_\Delta < \infty$, *there is* $D_j(z,h;\beta,\tilde{h})$ *that is linear on* $\| h \|_\Delta < \infty$ *satisfying* $|\mu_j(z,\beta,h) - \mu_j(z,\beta,\tilde{h}) - D_j(z,h-\tilde{h};\beta,\tilde{h})| = o(\| h - \tilde{h} \|_\Delta)$ *as* $\| h - \tilde{h} \|_\Delta \longrightarrow 0$ *for fixed* β *and* $\tilde{h}$; *(iii)* $\| D_j(z,h;\beta,\tilde{h}) - D_j(z,h;\beta_0,h_0) \| \le b_j(z) \| h \|_{\Delta_1} (\| \beta - \beta_0 \| + \| \tilde{h} - h_0 \|_{\Delta_2})$ *and* $\| D_j(z,h;\beta_0,h_0) \| \le \tilde{b}_j(z) \| h \|_{\Delta_3}$ *and* $E[\tilde{b}_j(z)^4] < \infty$; *(iv)* $\hat{\beta} = \beta_0 + O_p(\delta_{\beta N}) \longrightarrow 0$, $\sigma^{k_1/2-k-\Delta_1}\delta_{\beta N} \longrightarrow 0$, $k_1/2+\zeta > k+\Delta_1$, $N\sigma^{3k+2\Delta_1+2\Delta_2-k_1}/\ln N \longrightarrow \infty$, $N\sigma^{2k+2\Delta_3-k_1} \longrightarrow \infty$.

The following lemma is a generalization of Newey (1994b) Lemma 5.5.

Lemma 5 *Suppose that Assumption B-3 is satisfied. If for all* j,t, $\mu_j(h) = \int D_j(z,h;\beta_0,h_0)\, dF(z)$ *satisfies the conditions of Lemma 4, then for all* j,t, $\sigma^{2a}\hat{V}_{jt} \xrightarrow{r} V_{jt}$, *for* V_{jt} *in (1.10).*

PROOF: Let $\hat{D}_{itls,j} = D_j(z_{it}, q_{ls}K_\sigma(\cdot - x_{ls}); \hat{\beta}, \hat{h})$ and $D_{itls,j} = D_j(z_{it}, q_{ls}K_\sigma(\cdot - x_{ls}); \beta_0, h_0)$, $\tilde{\alpha}_{it,j} = \sum_{l=1}^N \sum_{s=1}^T D_{lsit,j}/NT$, and $\bar{\alpha}_{it,j} = \frac{1}{T}\sum_{l=1}^N \sum_{s=1}^T E[D_{lsit,j}|z_{it}]$ (note that $\hat{\alpha}_{it,j} = \sum_{l=1}^N \sum_{s=1}^T \hat{D}_{lsit,j}/NT$). Then using the method of proof for Newey (1994b) Lemma 5.5, it is shown that, for $t = 2,\dots,T$ and $j = 1,2$, $\sigma^{k_1}\sum_{i=1}^N \| \hat{\alpha}_{it,j} - \tilde{\alpha}_{it,j} \|^2 /N \xrightarrow{p} 0$ and $\sigma^{k_1}\sum_{i=1}^N \| \tilde{\alpha}_{it,j} - \bar{\alpha}_{it,j} \|^2 /N \xrightarrow{p} 0$. Note that $\bar{\alpha}_{it,j} = \rho_\sigma^j(X_{it}, X_{i,t-1})q_{it}$ as defined in the proof of Lemma 3. In that proof it was also shown that $\sigma^{k_1/2}E[\bar{\alpha}_{it,j}] \longrightarrow 0$, $\sigma^{k_1}E[\bar{\alpha}_{it,j}\bar{\alpha}'_{it,j}] \longrightarrow V_{jt}$, and $\sigma^{2k_1}E[\| \bar{\alpha}_{it,j} \|^4] \longrightarrow 0$. Hence by the Markov inequality, $\sigma^{k_1}\sum_{i=1}^N (\bar{\alpha}_{it,j} - \sum_{l=1}^N \bar{\alpha}_{lt,j}/N)(\bar{\alpha}_{it,j} - \sum_{l=1}^N \bar{\alpha}_{lt,j}/N)'/N \xrightarrow{p} V_{jt}$, the conclusion then follows by the triangle inequality. □

PROOF of Theorem 2: Let $\hat{m}_{P,\gamma}(x) = \gamma\hat{m}_{P,1}(x) + (1-\gamma)\hat{m}_{P,2}(x)$ and note that $\hat{m}_{P,\gamma=1} = \hat{m}_{P,1}$, $\hat{m}_{P,\gamma=0} = \hat{m}_{P,2}$ and $\hat{m}_{P,\gamma=1/2} = \hat{m}_P$; $\mu_1(z,h) =$

$\lambda(x_2)\frac{h_2(x,x_2)}{h_1(x,x_2)}$; $\mu_2(z,h) = -\lambda(x_1)\frac{h_2(x_1,x)}{h_1(x_1,x)}$; $D_1(z,h;\tilde{h}) = \lambda(x_2)\tilde{h}_1(x,x_2)^{-1}[h_2(x, x_2) - \frac{\tilde{h}_2(x,x_2)}{\tilde{h}_1(x,x_2)} h_1(x, x_2)]$; $D_2(z,h;\tilde{h}) = -\lambda_2(x_1)\tilde{h}_1(x_1,x)^{-1}$ $[h_2(x_1,x) -\frac{\tilde{h}_2(x_1,x)}{\tilde{h}_1(x_1,x)}h_1(x_1,x)]$; $\mu(z,h) = \gamma\mu_1(z,h) + (1-\gamma)\mu_2(z,h)$; $\mu(z,\beta,h) = \mu(z,h) - \beta$; $D(z,h;\tilde{h}) = \gamma D_1(z,h;\tilde{h}) + (1-\gamma)D_2(z,h;\tilde{h})]$; $D(z,h) = D(z,h;h_0)$. For $\| h - \tilde{h} \| < \epsilon$, $|\mu_j(z,h) - \mu_j(z,\tilde{h}) - D_j(z,h-\tilde{h};\tilde{h})| = |(\frac{\tilde{h}_1(x_j(x))}{h_1(x_j(x))} - 1)D_j(z,h-\tilde{h};\tilde{h})| \leq C \| h - \tilde{h} \|^2$ and $|D_j(z,h;\tilde{h})| \leq C \| h \|$. Hence, $|\mu(z,\beta,h) - \mu(z,\beta,\tilde{h}) - D(z,h-\tilde{h};\tilde{h})| \leq \sum_j((2-j)\gamma + (j-1)(1-\gamma))|\mu_j(z,h) - \mu_j(z,\tilde{h}) - D_j(z,h-\tilde{h};\tilde{h})| \leq \sum_j C_j \| h - \tilde{h} \|^2 \leq C \| h - \tilde{h} \|^2$ and $|D(z,h;\tilde{h})| \leq \sum_j((2-j)\gamma + (j-1)(1-\gamma))|D_j(z,h;\tilde{h})| \leq C \| h \|$. Checking the rate conditions, $\Delta = 0$, and note that $\sqrt{N}\sigma^{k_1/2} \longrightarrow \infty$ since $\frac{k_1}{2} \leq k - \frac{k_1}{2}$. $\varrho_N^0 = (\frac{\ln N}{N\sigma^k})^{1/2} + \sigma^\zeta \longrightarrow 0$ since $\frac{N^2\sigma^{2k}}{(\ln N)^2} = [\frac{N\sigma^{2k-k_1}}{(\ln N)^2}][N\sigma^{k_1}]$ and each of these two terms approaches infinity. $\sqrt{N}\sigma^{k_1/2}(\varrho_N^0)^2 = \frac{\ln N}{\sqrt{N}\sigma^{k-k_1/2}} + 2\sqrt{\ln N}\sigma^{\zeta+(k_1-k)/2} + \sqrt{N}\sigma^{2\zeta+k_1/2}$. So Lemma 4 holds with $\mu(h) = \sum_j \int((2-j)\gamma+(j-1)(1-\gamma))\lambda(x_j(a))\bar{f}_1(x_j(a))^{-1}$ $[h_2(x_j(a)) - l(x_j(a))h_1(x_j(a))]\bar{f}_0(a)da$, where $\bar{f}_0(\cdot) = \frac{1}{T}\sum_{s=1}^{T} f_{x_s}(\cdot)$. Now to check the assumptions of Lemma 3, we only need to consider the rate conditions and Assumption B-2. Let $\omega_j(a) = ((2-j)\gamma + (j-1)(1-\gamma))\lambda(x_j(a))\bar{f}_1(x_j(a))^{-1}\bar{f}_0(a)[-l(x_j(a)),1]$ which is bounded and continuous a.e. and zero outside a compact set. The rate conditions hold by hypothesis and by $N\sigma^{k_1/2} \longrightarrow \infty$.

Finally, note that for $\mu_j(z,h,\beta) = \mu_j(z,h)$ and $D_j(z,h;\beta,\tilde{h}) = D_j(z,h;\tilde{h})$ as given above, conditions (i)-(iii) of Assumption B-3 are satisfied, with $\Delta = \Delta_1 = \Delta_2 = \Delta_3 = 0$. Then with $\delta_{\beta N} = 0$ the rate conditions in (iv) hold and the consistent variance estimator conclusion follows by Lemma 5. □

References

Andrews, D. (1991) "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models." *Econometrica* 59, 307-345.

Bartlett, M. (1963) "Statistical Estimation of Density Estimation." *Sankhya*, Ser. A, 24, 145-154.

Breiman, L. and J. Friedman (1985) "Estimating Optimal Transformations for Multiple Regression and Correlation." *Journal of the American Statistical Association* 80, 580-598.

Chamberlain, G. (1982) "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18, 5-46.

Griliches, Z. and J. Hausman (1986) "Errors in Variables in Panel Data." *Journal of Econometrics* 31, 93-118.

Ichimura, H. (1993) "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models." *Journal of Econometrics* 58, 71-120.

Manski, C. (1984) "Adaptive Estimation of Non-linear Regression Models." *Econometric Reviews* 3, 145-194.

Nadaraya, E. (1965) "On Nonparametric Estimation of Density Functions and Regression Curves." *Theory Probability Applications* 10, 186-190.

Newey, W. (1994a) "Asymptotic Theory for Series Estimators." M.I.T. Department of Economics, Mimeo.

Newey, W. (1994b) "Kernel Estimation of Partial Means and a General Variance Estimator." *Econometric Theory* 10, 233-253.

Prakasa Rao, B. (1983) *Nonparametric Functional Estimation.* New York: Academic Press.

Robinson, P. (1983) "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis* 4, 185-207.

Robinson, P. (1988) "Root-N-Consistent Semiparametric Regression." *Econometrica* 56, 931-954.

Watson, G. (1964) "Smooth Regression Analysis" *Sankhya*, Ser A, 26, 359-372.

Chapter 2

A Nonparametric Empirical Analysis of the q Theory of Investment

2.1 Introduction

This chapter provides an application of the nonparametric panel estimation methods given in Chapter 1. We use these nonparametric methods to analyze nonlinearities in the investment - q relationship. The assumption of quadratic adjustment costs has received little attention in the empirical investment literature. We find that that assumption is too restrictive and that its acceptance explains some but not all of q's poor empirical performance. Additionally, cash flow is still found to be a significant regressor, and we are able to gain insight into the reason for its significance.

Brainard and Tobin (1968) and Tobin (1969) introduced the notion of average q, the ratio of a firm's market value to the replacement cost of its capital, as a determinant of the firm's optimal investment rate. Mussa (1977) and Abel (1983) then showed how the addition of adjustment costs to Jorgenson's (1963) neoclassical theory of investment leads to optimal investment that is a function only of marginal q, the marginal value of a firm's capital. Hayashi (1982) provided the theoretical foundation for the use of observable average q in the empirical literature by providing the conditions under which marginal q equals average q. Hayashi's conditions include the firm being a price-taker in its output market, its adjustment cost function being linearly homogenous in investment and capital, and its production function being linearly homogenous in capital and labor.

The empirical literature that followed began with regressions using aggregate investment data. Summers (1981) and Abel and Blanchard (1986) conclude that the estimated q coefficient was too low to explain the movements in investment and implied implausibly high costs of adjustment. The next wave of research used firm-level micro data, especially panel data, to test the q-theory. Papers like Schaller (1990) and Blundell, Bond, Devereux, and Schiantarelli (1992) (which also accounts for the potential endogeneity of q) find results somewhat more favorable to the q-theory than the aggregate data research, but still not completely satisfactory.

Fezzari, Hubbard, and Petersen (1988) focused attention on other variables that appear significant in investment-q regressions. In particular, they used a liquidity constraint story to explain the significance of cash flow, see also Hayashi and Inoue (1991). Other research, like Gilchrist and Himmelberg (1995), has considered the myriad of alternative explanations for cash flow's importance.

In section 2.2, we begin with a simple q model of investment to motivate our empirical approach. Section 2.3 describes the data used in estimation, and in section 2.4 we present the results of that estimation. Section 2.5 concludes.

2.2 A Simple q Model of Investment

We start with a simple example of the theory leading to our specification. Let $G_i(I, K, \eta)$ be the adjustment cost function for firm i with investment I and capital stock K. The exogenous shock to adjustment costs, η, is known to the firm but unobserved by the econometrician. Let Q be marginal q minus 1 times the ratio of the price of capital goods to the output price. Then the first-order condition from firm i's maximization of the expected present value of net profits is

$\left(\frac{\partial G_i}{\partial I}\right)_t = Q_{it}$. Thus, the marginal adjustment cost function is the key determinant of investment according to this theory. Conventional parametric methods assume a functional form for the marginal adjustment cost function, but with our nonparametric approach we can now directly estimate this function without restricting its shape. Traditionally, a quadratic parametrization for the adjustment cost function, such as $G_i(I, K, \eta) = \frac{b}{2}[(\frac{I}{K}) - \delta_i - \eta]^2 K$, leads to the popular linear specification $(\frac{I}{K})_{it} = \frac{1}{b}Q_{it} + \delta_i + \eta_{it}$. Here δ_i is a firm-specific parameter of the adjustment cost function. Instead we can allow for a much more flexible choice of adjustment cost function. Hayashi's conditions include linear homogeneity of G_i with respect to I and K, so $G_i(I, K, \eta) = KG_i(\frac{I}{K}, 1, \eta) = Kg_i(\frac{I}{K}, \eta)$. The only restriction we place on the model is that $g_i(u, \eta) = g(u - \delta_i - \eta)$, so that δ_i captures the differences in adjustment cost functions across firms and the exogenous shock enters additively as an argument of g. If G_i represents an adjustment cost function without fixed costs and not including purchasing costs, then $G_i(0, K) = 0$ so adjustment costs are zero at $\frac{I}{K} = \delta_i + \eta$. If η is a mean zero exogenous shock, then we can interpret δ_i as the average firm specific investment-capital ratio at which adjustment costs are zero. For example, we might think that each firm has a usual replacement ratio of investment to capital, which might depend directly on its depreciation rate. Adjustment costs are zero at this usual replacement ratio and increase for deviations from this ratio. For instance, higher than usual investment rates might entail paying overtime wages for installation. So we can interpret δ_i as the firm's usual replacement ratio and, in estimation, allow it to be different for different firms. Finally, solving for $\frac{I}{K}$, we have the nonlinear-in-Q specification

$$\left(\frac{I}{K}\right)_{it} = m(Q_{it}) + \delta_i + \eta_{it}$$

where $m(\cdot) = (g')^{-1}(\cdot)$.

2.3 Data

The results shown here use COMPUSTAT data on 425 manufacturing firms over the years 1981-86. Only manufacturing firms with data for all six years were included in the sample. We follow Fezzari, Hubbard, and Petersen (1988) in extracting and constructing our variables from the COMPUSTAT data set. The investment variable is defined as reported spending on property, plant, and equipment excluding spending on acquisitions. Tobin's q is calculated by the method outlined in Salinger and Summers (1983).

2.4 Empirical Results

2.4.1 Investment and q

In Table 2.1, we present results from estimation of the linear model ($m(Q_{it}) = \alpha + \beta Q_{it}$). Coefficient estimates are given along with standard errors corrected for heteroskedasticity and serial correlation in parentheses. The OLS estimate of the coefficient on Q ignoring the fixed effects is 0.0215, and the OLS constant term estimate (standard error) not shown in the table is 0.1766 (0.0209). The within and first difference estimates approximately double the OLS estimate, indicating the presence of fixed effects. These estimates (0.0403 and 0.0475) are typical of estimates found in the literature.

From Griliches and Hausman (1986), if errors in variables are present, we expect the first difference estimate to be biased more toward zero than the within estimate and the longest difference estimate to be the least biased toward zero. Here we see the opposite results. While this ordering could be explained within Griliches and Hausman's errors-in-variables framework by serial correlation in the errors, one might also consider the failure of linearity as a possible explanation. Though the robust standard errors for the within and first difference estimates are quite large, the difference of these two estimates is more precisely estimated. A chi-square test of the null hypothesis that the within and first difference estimates are equal has

Table 2.1: Linear Panel Estimation Results: Q Coefficient Estimates

OLS	Within	First Difference	Longest Difference
0.0215	0.0403	0.0475	0.0417
(0.0072)	(0.0147)	(0.0156)	(0.0175)

a value of 2.68 (degrees of freedom = 1), which has a p-value of 0.10. This result is not definitive but again points in the direction of the failure of the linear panel specification.

In Table 2.2, we turn to another test of the linear model, presenting the multivariate linear regression on leads and lags of Q as suggested by Chamberlain (1982). The null hypothesis of no heterogeneity bias is tested by imposing the restrictions that the off-diagonal elements of the matrix of coefficients are all equal to zero and the diagonal terms are equal. The chi-square test statistic of these restrictions has 35 degrees of freedom and has a value of 277.17, which gives a strong rejection of the null hypothesis.[1] Heterogeneity bias within the linear specification still implies restrictions on the coefficient estimate matrix in Table 2.2. Specifically, we expect equality of the off-diagonal terms within each column and equality of the diagonal elements. The chi-square test statistic of these restrictions has a value of 189.89 with 29 degrees of freedom, which strongly rejects the linear model.[2] Again, one possible conclusion would be the need for a more flexible nonlinear specification, so we turn our attention to the nonparametric estimators discussed in this paper.

In Figure 2-1 , we present two first difference type series estimators. The solid line is the simple first difference series estimator ($\hat{m}_F$) and the dashed line is the partial means first difference estimator ($\hat{m}_M$). In both cases, our estimators use polynomial

[1] Relaxing the restriction that the diagonal elements are equal to allow for time-varying β's and just testing that the off-diagonal elements are zero makes no difference in the results. The test statistic has a value of 201.23 with 30 degrees of freedom, so we still clearly reject.

[2] As before, we can allow for time-varying β's by ignoring the restriction of equality of the diagonal elements, and the results are unchanged. The chi-square statistic has 24 degrees of freedom and a value of 154.08, leading to rejection of the linear model with time-varying coefficients.

Table 2.2: Multivariate Linear Regression Results

Dependent Variable	Q_{1981}	Q_{1982}	Q_{1983}	Q_{1984}	Q_{1985}	Q_{1986}
I_{1981}	0.0853 (0.0226)	-0.1174 (0.0344)	0.0241 (0.0218)	-0.0184 (0.0235)	-0.0071 (0.0293)	0.0184 (0.0163)
I_{1982}	-0.0107 (0.0067)	0.0643 (0.0269)	-0.0186 (0.0148)	-0.0057 (0.0110)	-0.0022 (0.0146)	-0.0035 (0.0070)
I_{1983}	-0.0006 (0.0031)	0.0105 (0.0102)	0.0185 (0.0134)	-0.0113 (0.0097)	-0.0105 (0.0125)	-0.0000 (0.0067)
I_{1984}	0.0017 (0.0030)	-0.0038 ((0.0067)	0.0041 (0.0065)	0.0050 (0.0085)	0.0051 (0.0082)	-0.0049 (0.0044)
I_{1985}	-0.0026 (0.0013)	0.0061 (0.0057)	-0.0029 (0.0095)	-0.0015 (0.0085)	0.0355 (0.0229)	-0.0160 (0.0125)
I_{1986}	-0.0037 (0.0012)	-0.0011 (0.0034)	0.0137 (0.0062)	-0.0091 (0.0058)	0.0013 (0.0119)	0.0105 (0.0077)

$\chi^2_a(35) = 277.17 \quad \chi^2_b(29) = 189.89$

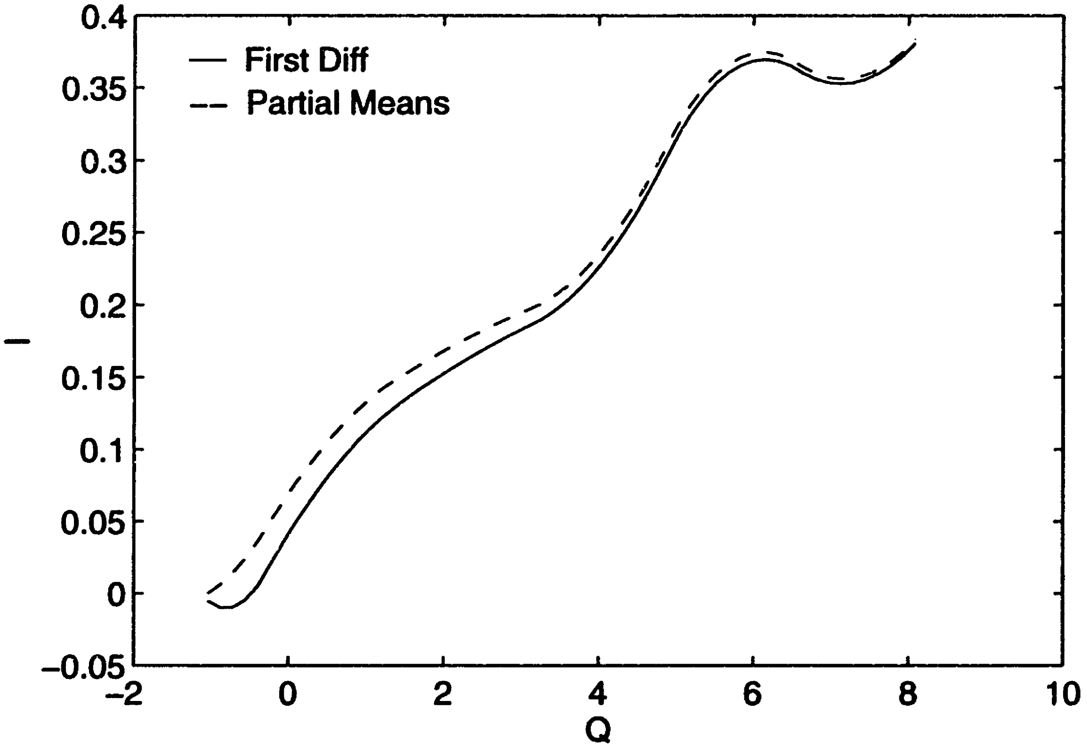
spline approximating functions, a second degree polynomial with six evenly spaced knots.[3]The difference in choice of approximating functions for these two estimators exists because the simple first difference estimator is essentially doing univariate regression, while the partial means estimator first estimates the two-dimensional regression function and then partials out with respect to one of the arguments. The two-dimensional estimation for the partial means estimator includes interactions of the leading polynomial terms but not interactions of the piecewise second degree polynomial terms. The latter interaction terms were never found to have statistically significant coefficient estimates and their addition usually resulted in near singularity of $q^K(x)q^K(x)'$ (corresponding to too much undersmoothing). It is clear that these two estimators are very close. It is encouraging that the partial means method, which nonparametrically regresses on a two-dimensional space and then partials out, so closely mirrors the simple first difference estimator, which uses a restriction to reduce the problem to nonparametric regression on a single dimension. This example is also reassuring for applications using kernel partial means, for which a similar comparison is not possible.

The nonlinearities in this figure are not striking, but are strong enough to result in the rejection of the linear model above. The flattening of the curve for higher values of Q seems to be a persistent feature of nonparametric panel estimation of investment on Q. Also, the overall slope roughly corresponds to the slope coefficent from first difference linear estimation in Table 2.1.

Next we consider difference from means type estimation using kernel methods. Difference from means type estimation is convenient for making comparisons to nonparametric estimation ignoring fixed effects. The within type estimate (using a bandwidth of 1.6) along with a standard kernel regression estimate (ignoring the fixed effects) are shown in Figure 2-2. A second-order Epanechnikov kernel is used, though the estimates do not seem very sensitive to kernel choice. Several features of this fig-

[3] The specification was selected by adding a single knot to the number chosen by a leave-one-out cross-validation criterion. This cross-validation criterion is known to lead to the mean square error minimizing choice of number of terms. We add a term to "undersmooth" in accordance with the theory.

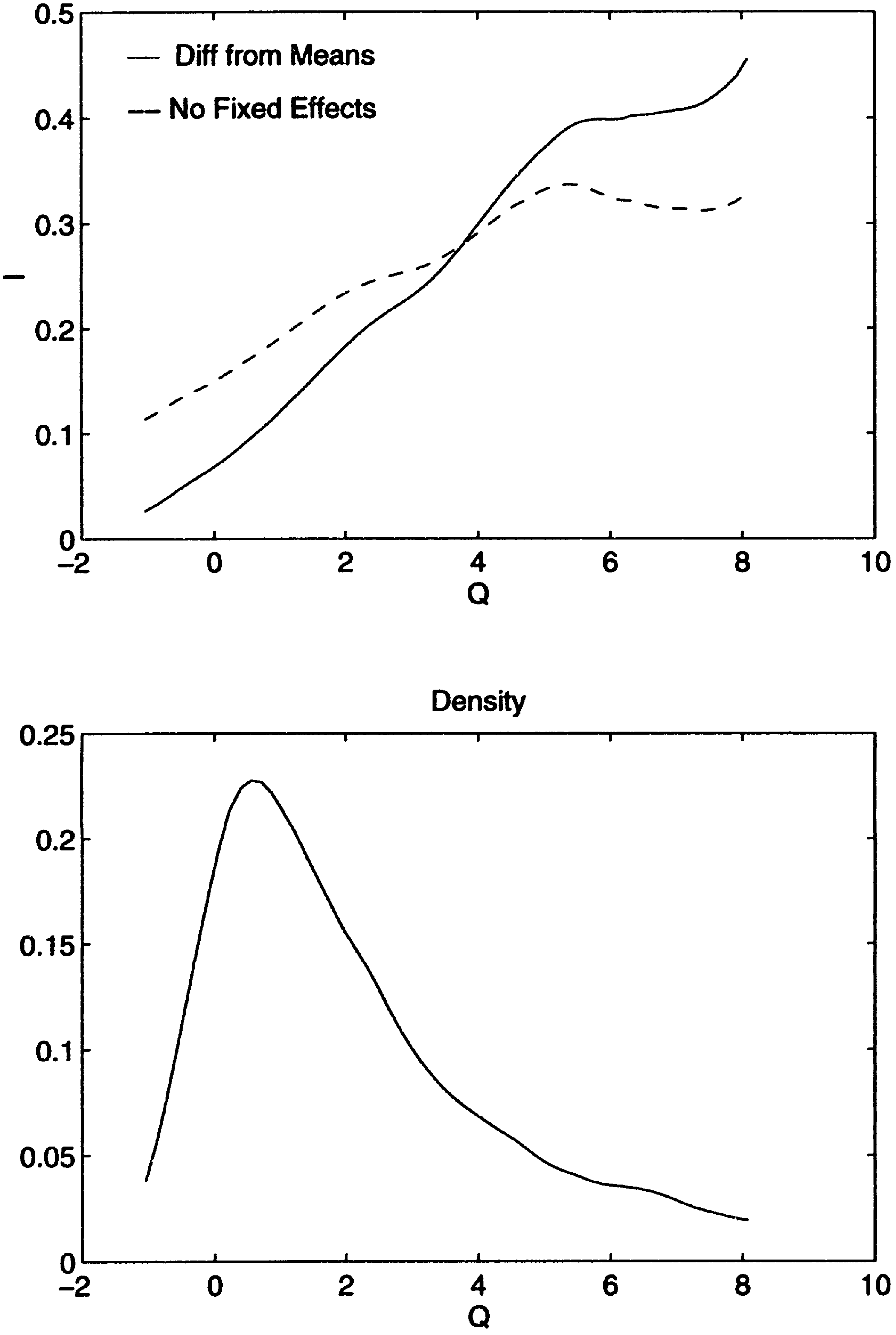
Figure 2-1: Investment on Q: Series Estimation



ure are interesting. First, the role of the fixed effects appears to be quite important even after we have allowed for a flexible nonlinear marginal adjustment cost function. A comparison of these two kernel regression estimates is analogous to the comparison of OLS and within estimates. Also in Figure 2-2, we see the role of the nonlinearites in the marginal adjustment cost function. In particular, at lower values of Q where most of the density is, the function seems steeper than indicated by linear methods. Again there is some flattening of the curve at high values of Q which accounts for the lower linear slope estimates.

Linear panel estimates of the coefficient on Q, like those given in Table 2.1, are generally considered too low (to account for the movements in investment that are seen empirically). Our nonparametric specifications give us some additional insight into the investment-q relationship. First, we find that even allowing for nonlinearities in the regression function, individual firm effects are still playing an important role (i.e. heterogeneity bias is still a prominent feature of our nonparametric specification). Overall, the curve is fairly linear, but for large values of Q there seems to be a

Figure 2-2: Investment on Q: Difference from Means Estimation



flattening of the investment-q curve, and correspondingly the curve is steeper for lower values of q than is found by conventional linear panel methods. Traditionally, we would associate this result with a finding of more convex adjustment costs than what is commonly found, but Whited (1994) cautions against pushing that link too far. Finally, the work of Abel and Eberly (1994) introduces flow fixed costs and kinks in the adjustment cost function into the theory of firm investment. The result is a threshold model of investment, which is estimated by Barnett and Sakellaris (1995). As in Barnett and Sakellaris, our figures present only mild evidence of flat thresholds and do indicate varying slopes over different regions of Q.

2.4.2 Testing for Measurement Error

To further substantiate the interesting results of the last section, we consider briefly the potential measurement error in Q. All of the nonparametric graphs of investment vs. Q seem to suggest a particular parametric relationship. Specifically, we will assume (for this subsection) a piecewise linear relationship with one breakpoint. By moving to this parametric nonlinear functional form, we will be able to easily compare traditional errors in variables test statistics in the linear and nonlinear cases.

In Table 2.3, we give the piecewise linear results analogous to Table 2.1. The estimated breakpoint 5.09 will be used throughout.[4] As expected, the Q coefficient is lower in all cases for high Q values, corresponding to our nonparametric graphs in the previous section. With the linear functional form, we tested for equality of the within and first difference Q coefficient estimates from Table 2.1, as a simple measurement error test. In that case, the chi-square test statistic was 2.68 which is not definitive but has a p-value of 0.10.

We now compute the chi-square test statistic for equality of the within and difference estimates of the Q coefficient for both high and low Q values. The resulting chi-square statistic with two degrees of freedom has a value of 1.43, which will not allow us to reject the null hypothesis of equality. This test is an indication that

[4] We will not account for the estimation of the breakpoint in the calculation of standard errors for other parameter estimates.

Table 2.3: Nonlinear Panel Estimation Results: Q Coefficient Estimates

Q	OLS	Within	Difference
$Q \leq 5.09$	0.0460 (0.0061)	0.0737 (0.0109)	0.0537 (0.0135)
$Q > 5.09$	0.0203 (0.0102)	0.0382 (0.0176)	0.0469 (0.0173)

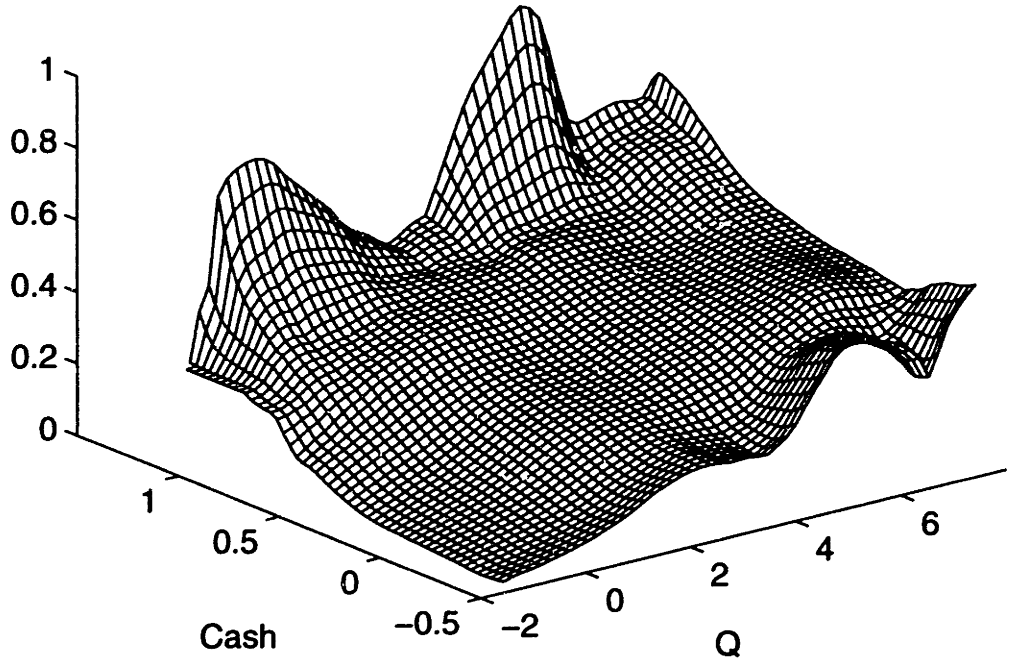
once the nonlinearities in the investment-Q relationship are accounted for the role of measurement error is diminished to statistical insignificance.

Following Griliches and Hausman (1986), in Table 2.4 we use instumental variables to test again for measurement error. We allow for an MA(2) measurement error process, so that some of the leads and lags of the Q's can be used as instruments in certain individual differenced equations. For each of these equations we estimate a Q coefficient in the linear model and two Q coefficients corresponding to high and low Q values for the piecewise linear model. Finally we impose the restriction of equality across the individual equations and compute the corresponding test staistic. For the linear model, the chi-square test of restrictions has five degrees of freedom and a value of 12.66. Thus we can reject the hypothesis of equality, indicating either measurement error or a specification error. In the nonlinear model, the chi-square test of restrictions has ten degrees of freedom and a value of 8.88. Thus in this nonlinear specification, our test does not show evidence of significant measurement error. Together the tests of this section point toward a specification error in the linear model, which is accounted for by the piecewise linear functional form. With this new nonlinear specification the role of measurement error in Q is statistically insignificant.

Table 2.4: Linear IV Panel Estimation Results (Q Coefficient Estimates) Allowing MA(2) Measurement Error Process

Difference	Q	$Q \leq 5.09$	$Q > 5.09$
$dy_2 - dy_1$	0.0451 (0.0159)	0.8390 (2.1250)	0.0922 (0.1217)
$dy_3 - dy_2$	-0.0178 (0.0126)	-0.7570 (1.6113)	-0.0276 (0.0384)
$dy_5 - dy_4$	0.0068 (0.0065)	-0.1612 (0.1895)	-0.0019 (0.0121)
$dy_6 - dy_5$	0.3068 (1.2081)	0.3052 (0.5126)	-0.1199 (0.1934)
$dy_3 - dy_1$	0.2099 (0.2256)	-0.7858 (1.4407)	0.0648 (0.1615)
$dy_6 - dy_4$	0.0205 (0.0100)	0.0904 (0.1242)	0.0220 (0.0108)
Restricted	0.0052 (0.0062)	0.0652 (0.0993)	0.0117 (0.0070)
χ^2 Test	12.6614 (5 d.f)	8.8799 (10 d.f.)	

Figure 2-3: Investment on (Q, Cash)

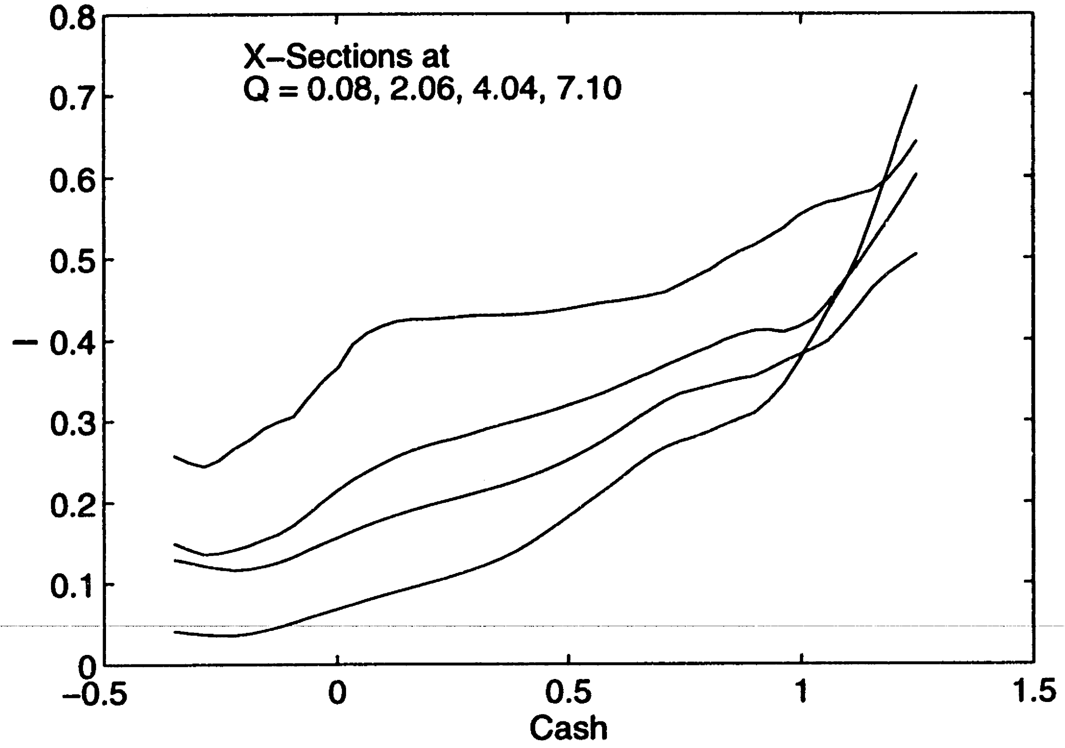


2.4.3 Cash Flow

Next we explore the role of cash flow as an explanatory variable for investment. In Figure 2-3, the kernel difference from means method with bandwidths 1.3 for Q and 0.35 for cash is used to regress investment on Q and cash flow. From the three dimensional picture it is difficult to discern the details of the figure, so in Figure 2-4, we present the investment - cash cross sections at various Q values. We find that the investment - cash relationship is fairly linear, except for the extreme Q's where data is more scarce. More importantly, the function is upward sloping over the full range of cash flow values with no signs of flattening out.

Interestingly, the addition of cash flow to our nonparametric regression has very little effect on the slope of the investment - Q relationship. For instance, since the investment - cash flow cross sections were quite parallel and linear suggesting an additive linear-in-cash-flow term, we estimated the partially linear model from section 1.5.3. We graph the resulting nonparametric curve estimate in Figure 2-5 along

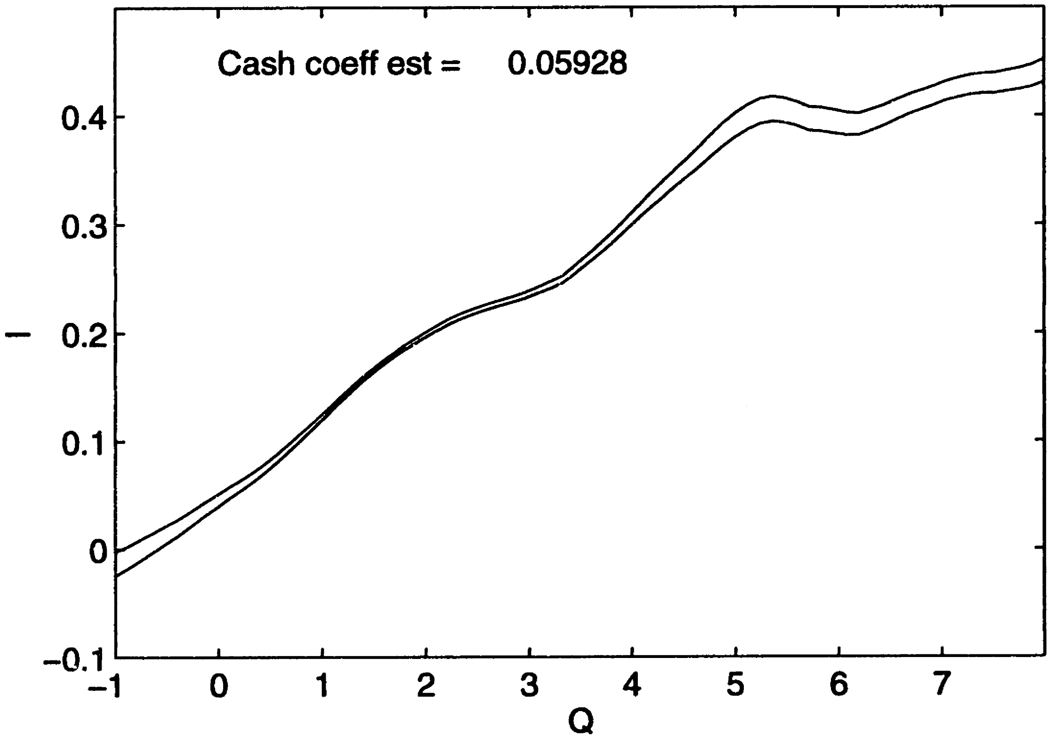
Figure 2-4: Investment on Cash Cross-Sections: Kernel Difference from Means



with the original investment - Q curve that does not partial out cash. Clearly, cash flow has a minimal effect along the Q-dimension.

Most importantly, the investment - cash flow relationship suggests that liquidity constraints may not be the overriding reason for cash flow's significance in the investment equation. Specifically, we graph the relationship for quite high values of cash flow (recalling that the cash flow axis is expressed in terms of the ratio of cash flow to capital stock), and never find a flattening of the curve even for very high cash values as would be expected with the liquidity constraint explanation. Instead, the shape of the investment - cash curve suggests that the free cash flow model where managers have an incentive to increase firms size and spend all cash on investment projects regardless of the return is a more likely explanation. Alternatively, one might think that cash flow is simply acting as a signal of firm investment opportunities.

Figure 2-5: Investment on Q with Patialled Out Cash



2.5 Conclusion

Using a panel of firm investment data and the estimators of Chapter 1, we analyze nonlinearities in the investment - q relationship. The main conclusion is that we find a steeper slope than usually found over lower Q valu, we turn to another test of the linear model, presenting the multivariate linear regression on leads and lags of Q as suggested by Chamberlain (1982). The null hypothesis of no heterogeneity bias is tested by imposing the restrictions that the off-diagonal elements of the matrix of coefficients are all equal to zero and the diagonal terms are equal. The chi-square test statistic of these restrictions has 35 degrees of freedom and has a value of 277.17, which gives a strong rejection of the null hypothesis.[5] Heterogeneity bias within the linear specification still implies restrictions on the coefficient estimate matrix in Table 2.2. Specifically, we expect equality of the off-diagonal terms within each column and

[5] Relaxing the restriction that the diagonal elements are equal to allow for time-varying β's and just testing that the off-diagonal elements are zero makes no difference in the results. The test statistic has a value of 201.23 with 30 degrees of freedom, so we still clearly reject.

equality of the diagonal elements. The chi-square test statistic of these restrictions has a value of 189.89 with 29 degrees of freedom, which strongly rejects the linear model.[6] Again, one possible conclusion would be the need for a more flexible nonlinear specification, so we turn our attention to the nonparametric estimators discussed in this paper.

[6] As before, we can allow for time-varying β's by ignoring the restriction of equality of the diagonal elements, and the results are unchanged. The chi-square statistic has 24 degrees of freedom and a value of 154.08, leading to rejecticn of the linear model with time-varying coefficients.

References

Abel, A. (1983) "Optimal Investment Under Uncertainty." *American Economic Review* 73, 228-233.

Abel, A. and O. Blanchard (1986) "The Present Value of Profits and Cyclical Movements in Investment." *Econometrica* 54, 249-273.

Abel, A. and J. Eberly (1994) "A Unified Model of Investment Under Uncertainty." *American Economic Review* 84, 1369-1384.

Barnett, S. and P. Sakellaris (1995) "Nonlinear Response of Firm Investment to Q: Testing a Model of Convex and Non-convex Adjustment Costs." University of Maryland Department of Economics, Mimeo.

Blundell, R., S. Bond, M. Devereux, F. Shiantarelli (1992) "Investment and Tobin's *Q*." *Journal of Econometrics* 51, 233-257.

Brainard, W. and J. Tobin (1968) "Pitfalls in Financial Model Building" *American Economic Review* 58, 99-122.

Chamberlain, G. (1982) "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18, 5-46.

Fezzari, S., R. Hubbard, and B. Petersen (1988) "Financing Constraints and Corporate Investment." *Brookings Papers on Economic Activity*, 141-195.

Gilchrist, S. and C. Himmelberg (1995) "Evidence on the Role of Cash Flow for Investment." Boston University Department of Economics, Mimeo.

Griliches, Z. and J. Hausman (1986) "Errors in Variables in Panel Data." *Journal of Econometrics* 31, 93-118.

Hayashi, F. (1982) "Tobin's Marginal and Average q: A Neoclassical Interpretation." *Econometrica* 50, 213-24.

Hayashi, F. and T. Inoue (1991) "The Relation Between Firm Growth and *Q* with Multiple Capital Goods: Theory and Evidence from Panel Data on Japanese Firms." *Econometrica* 59, 731-753.

Jorgenson, D. (1963) "Capital Theory and Investment Behavior." *American Economic Review (Papers and Proceedings)* 53, 247-259.

Mussa, M. (1977) "External and Internal Adjustment Costs and the Theory of Aggregate and Firm Investment." *Econometrica* 44, 163-178.

Salinger, M. and L. Summers (1983) "Tax Reform and Corporate Investment: A Microeconomic Simulation Study." In *Behaviorial Simulation Methods in Tax Policy Analysis*, ed. M. Feldstein, Chicago UP.

Schaller, H. (1990) "A Re-examination of the Q Theory of Investment Using U.S. Firm Data." *Journal of Applied Econometric* 5, 309-325.

Summers, L. (1981) "Taxation and Corporate Investment: A q-theory Approach." *Brookings Papers on Economic Activity*, 67-140.

Tobin, J. (1969) "A General Equilibrium Approach to Monetary Theory." *Journal of Money, Credit, and Banking* 1, 15-29.

Whited, T. (1994) "Problems with Identifying Adjustment Costs from Regressions of Investment on Q." *Economics Letters* 46, 339-344.

Chapter 3

Feasible Simulation Variance Estimation for Maximum Likelihood

3.1 Introduction

In general the asymptotic variance of a maximum likelihood estimator is $H^{-1}JH^{-1}$, where H is the Hessian of the log-likelihood and J is the Fisher information matrix. Under correct specification of the distribution and certain regularity conditions, the information matrix equality holds, $H = -J$, and the asymptotic variance collapses to its familiar form, J^{-1}. Two common estimators of J are the sample average of the outer product of the scores and the sample average of the negative derivative of the score, both using a parameter estimate in place of the unknown true parameter value.

A third less frequently used estimator is the sample average of the conditional information matrix (conditioning on the regressors). There is some evidence to suggest that the conditional information matrix [c.i.m.] estimator often outperforms the above estimators in terms of approximating the asymptotic distribution of test statistics, see Davidson and MacKinnon (1984). Further evidence in favor of the c.i.m. estimator is presented in Section 3.2 of this paper. Specifically, this estimator achieves the semiparametric efficiency bound and thus is at least as efficient as any other ML variance estimator. Unfortunately, the integral involved in computation of the conditional information matrix often makes this approach infeasible with complicated likelihoods. We follow Lerman and Manski (1981), McFadden (1989), and Pakes and Pollard (1989) in using simulation to overcome the problem of evaluating a burdensome integral. The conditional density evaluated at the given data points and the ML parameter estimate can be used to simulate an approximation to the conditional information matrix for each observation. The sample average of these approximations then gives a consistent estimate of the asymptotic variance. We analyze the performance of these simulation variance estimators through a comparison of their asymptotic distributions and through Monte Carlo evidence of their error coverage rates in confidence intervals for the maximum likelihood estimates.

In Section 3.2, we present the commonly used variance estimators for maximum likelihood, and conditions for consistency of each estimator are given. Their relative performances in estimating the asymptotic variance is discussed, and the efficiency advantage of the c.i.m. estimator is established formally in a semiparametric estimation setting. The simulation estimators are presented in Section 3.3. A theorem giving consistency conditions for the simulation estimators is stated. In Section 3.4, the influence functions for the variance estimators are used to derive their asymptotic variances which are, in turn, analyzed and compared. Section 3.5 contains Monte Carlo results using these variance estimators in computing confidence intervals for maximum likelihood estimates. In Section 3.5.1, all of the variance estimators presented in the paper are implemented in a probit model. In Section 3.5.2, a sample selection model is used to give further Monte Carlo evidence in favor of simulation estimation. In this model, not all of the variance estimators are available so the simulation estimators become additionally useful. Section 3.6 concludes.

3.2 Standard MLE Variance Estimators

Suppose we are given i.i.d. data $(y_1, x_1), \ldots, (y_n, x_n)$ where $y_1|x_1, \ldots, y_n|x_n$ are drawn from a probability density function $f(y|x, \beta_0)$, which is a member of the family of p.d.f.'s $f(y|x, \beta)$ indexed by β. Here, as often occurs in econometrics, we have a set of "regressors" x whose marginal distribution is unspecified,[1] and so estimation proceeds by conditional maximum likelihood. In particular, the log likelihood function is $L_n(\beta) = n^{-1} \sum_{i=1}^{n} \ln f(y_i|x_i, \beta)$ and the maximum likelihood estimate of β_0 is given by

$$\hat{\beta} = \underset{\beta \in B}{\text{argmax}} \; L_n(\beta)$$

where B is a (possibly restricted) parameter space which contains β_0.

Conditions for consistency and asymptotic normality of $\hat{\beta}$ can be found in standard references such as Amemiya (1985). If $H = E[\partial^2 \ln f(y|x, \beta_0)/\partial\beta\partial\beta']$ and $J = E[\partial \ln f(y|x, \beta_0)/\partial\beta \; (\partial \ln f(y|x, \beta_0)/\partial\beta)']$, then $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, H^{-1}JH^{-1})$ under the conditions of a standard asymptotic normality result for extremum estimators. We can also apply the information matrix equality $H = -J$ to obtain the simplified expression $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, J^{-1})$.

This last result leaves us with many potential estimates of the asymptotic variance. By continuity of matrix inversion and the Slutsky Theorem, attention can center on estimation of J itself rather than J^{-1}. The most convenient estimator is the sample average of the outer product of the scores. Let $s(y, x, \beta) = \partial \ln f(y|x, \beta)/\partial\beta'$, then

$$\hat{J}_1 = \frac{1}{n} \sum_{i=1}^{n} s(y_i, x_i, \hat{\beta})' s(y_i, x_i, \hat{\beta}).$$

This estimator is just the sample analog of the Fisher Information Matrix.

Using the information matrix equality, the sample analog of the negative Hessian of the log likelihood provides a second estimator. Let $h(y, x, \beta) = \partial^2 \ln f(y|x, \beta)/\partial\beta\partial\beta'$, then

$$\hat{J}_2 = -\frac{1}{n} \sum_{i=1}^{n} h(y_i, x_i, \hat{\beta}).$$

[1] The unconditional case can be considered as a special case in this framework.

In some contexts, the computational difficulty of taking second derivatives renders $\hat{J}_2$ less suitable than $\hat{J}_1$.

Consider next the conditional information matrix

$$J(x,\beta) \;=\; \int \frac{\partial \ln f(y|x,\beta)}{\partial \beta}\left(\frac{\partial \ln f(y|x,\beta)}{\partial \beta}\right)' f(y|x,\beta)dy. \tag{3.1}$$

Using the law of iterated expectations, we have another identity, $J = E[J(x,\beta_0)]$, which leads to a third estimator,

$$\hat{J}_3 = \frac{1}{n}\sum_{i=1}^{n} J(x_i,\hat{\beta}). \tag{3.2}$$

Analogous to the conditional information matrix estimator is the conditional Hessian estimator, where

$$H(x,\beta) = E\left[h(y,x,\beta)|x,\beta\right] = \int h(y,x,\beta)f(y|x,\beta)dy.$$

However, under a generalization of the information matrix equality, $J(x,\beta) = -H(x,\beta)$.

Theorem 5 *Suppose $(y_1,x_1),\ldots,(y_n,x_n)$ are i.i.d. and i) $f(y|x,\beta)$ is twice continuously differentiable and $f(y|x,\beta) > 0$ a.s. $dxdy$ in a neighborhood $\mathcal{N}$of β_0; ii) $\int \sup_{\beta\in\mathcal{N}} \parallel \partial f(y|x,\beta)/\partial\beta \parallel dy < \infty$ a.s. dx; iii) $\int \sup_{\beta\in\mathcal{N}} \parallel \partial^2 f(y|x,\beta)/\partial\beta\partial\beta' \parallel dy < \infty$ a.s. dx. Then $H(x,\beta) = -J(x,\beta)$ a.s. dx for $\beta \in \mathcal{N}$.*

PROOF: see Appendix

Hence $\hat{J}_3 = n^{-1}\sum_i J(x_i,\hat{\beta}) = -n^{-1}\sum_i H(x_i,\hat{\beta})$, i.e. the c.i.m. estimator and the conditional Hessian estimator are numerically identical estimators. This estimator is used even less than $\hat{J}_1$ and $\hat{J}_2$ as it requires an evaluation of the integral in equation (3.1). In complicated models, such a solution may be very difficult to obtain. In the next section simulation estimators that avoid computation of this potentially complicated integral are suggested. Before presenting those estimators, we need to establish the conditions necessary to achieve consistency of $\hat{J}_1$, $\hat{J}_2$, and $\hat{J}_3$.

The following lemma taken from Newey and McFadden (1993) is the key to proving consistency of the usual covariance estimators above. The sufficient conditions are continuity at the true parameter and a dominance condition.

Lemma 6 *If $z_1, \ldots, z_n$ are i.i.d., $a(z, \beta)$ is continuous at β_0 with probability one and there exists a neighborhood N of β_0 such that $E[\sup_{\beta \in \mathcal{N}} \parallel a(z, \beta) \parallel] < \infty$, then for any $\hat{\beta} \xrightarrow{p} \beta_0$, $\frac{1}{n} \sum_{i=1}^{n} a(z_i, \hat{\beta}) \xrightarrow{p} E[a(z, \beta_0)]$.*[2]

Then the following theorem gives the conditions for consistency of the covariance estimators discussed so far.

Theorem 6 *Suppose (i) $\hat{\beta} \xrightarrow{p} \beta_0$; (ii) $f(y|x, \beta)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of β_0; and (iii) J exists and is nonsingular.*
a) *If (iv) $E[\sup_{\beta \in \mathcal{N}} \parallel \partial f(y|x, \beta)/\partial \beta \parallel^2] < \infty$, then $\hat{J}_1^{-1} \xrightarrow{p} J^{-1}$.*
b) *If (v) $\int \sup_{\beta \in \mathcal{N}} \parallel \partial f(y|x, \beta)/\partial \beta \parallel dy < \infty$ and $\int \sup_{\beta \in \mathcal{N}} \parallel \partial^2 f(y|x, \beta)/\partial \beta \partial \beta' \parallel dy < \infty$ a.s. dx and (vi) $E[\sup_{\beta \in \mathcal{N}} \parallel \partial^2 f(y|x, \beta)/\partial \beta \partial \beta' \parallel < \infty$, then $\hat{J}_2^{-1} \xrightarrow{p} J^{-1}$.*
c) *If (vii) $J(x, \beta)$ is continuous at β_0 with probability one and (viii) $E[\sup_{\beta \in \mathcal{N}} \parallel J(x, \beta) \parallel] < \infty$, then $\hat{J}_3^{-1} \xrightarrow{p} J^{-1}$.*

> **PROOF:** The conclusions in (a) and (c) follow by a straigtforward application of Lemma 6. Conditions (ii) and (v) are sufficient to switch the order of differentiation and integration twice in the expression $\int \frac{\partial^2 f(y|x, \beta_0)}{\partial \beta \partial \beta'} dy$ which in turn is enough to prove the information matrix equality, $H = -J$ (see the proof of Theorem 5 for details). The conclusion in (b) then follows by an application of Lemma 6.

Condition (iii) in Theorem 6 is needed to apply the Slutsky Theorem. Condition (ii) is stronger than necessary for the result in (a), but is usually a condition for asymptotic normality anyway (to obtain uniform convergence of the Hessian).

If we are simply interested in obtaining the best possible estimate of J or J^{-1} given data and the MLE $\hat{\beta}$, then Theorem 6 provides no guide as to which estimator to use. Theorem 6 is an asymptotic result that does not address the small sample properties of these estimators.

[2] A proof of this lemma is given in Section 4 of Newey and McFadden (1993).

The summations in $\hat{J}_1$ and $\hat{J}_2$ are approximations to integrals in the $Y \times X$ space taken with respect to the joint density $f(y, x|\beta_0)$. The summation in $\hat{J}_3$ is an approximation to an integral in the X space with respect to the marginal density of x. Compared to $\hat{J}_1$ and $\hat{J}_2$, $\hat{J}_3$ reduces the dimension (by dim$[Y]$) of the space over which the integral it is approximating is taken. This reduction is accomplished by making use of the fact that the conditional density $f(y|x, \hat{\beta})$ is known. $J(x, \beta)$ is computed by taking an integral with respect to the given conditional density. Then we need only approximate the integral of $J(x, \hat{\beta})$ with respect to the marginal density of x which is unspecified. So the data $y_1, \ldots, y_n$ is not used because we have estimates of the conditional densities from which the y_i's are drawn, specifically $f(y|x_1, \hat{\beta}), \ldots, f(y|x_n, \hat{\beta})$. The next theorem formalizes the intuition that $\hat{J}_3$ has an advantage over $\hat{J}_1$ and $\hat{J}_2$ because it makes more use of the structure of the problem.

Before stating the assumptions for Theorem 7, we need some additional notation. All of the notation given here corresponds directly to the notation in Brown and Newey (1992) [BN]. An s subscript denotes the stacked vector version of a matrix. We can think of the given variance estimators as method of moments estimators with $\hat{J} = n^{-1} \sum_{i=1}^{n} m(y_i, x_i, \hat{\beta})$, where for example $m(y, x, \beta) = h(y, x, \beta)$ in $\hat{J}_2$. So we will present the assumptions below for general m and then note that they must hold for $m(y, x, \beta) = [s(y, x, \beta)' s(y, x, \beta)]$ and $m(y, x, \beta) = h(y, x, \beta)$. Let $\mu^*(\beta) = \int \int m(y, x, \beta)_s f(y, x|\beta_0) dy dx$ and $J^c(\beta) = E[J(x, \beta)_s]$. As in BN, $\theta = (\beta, \eta)$ is used to index parametric submodels, where η is a finite vector of shape parameters for the distribution of x. Let $\mu(\theta) = \int \int m(y, x, \beta(\theta)) f(y|x, \beta(\theta)) dy f(x, \eta(\theta)) dx$ and $S = s(y, x, \beta_0)$.

Assumptions 1-5 correspond to Assumptions 1-6 in BN, and Assumptions 6-7 are analogs to Assumptions 2-3 for the efficient influence function case.[3] Assumption 1 and the finite second moment condition in Assumption 4 ensure that $\hat{\beta}$ is consistent and asymptotically normal with variance J. Assumptions 2 and 3 and the finite second moment condition in Assumption 5 are smoothness conditions which allow us

[3] Lemma A in the Appendix could be used to develop sufficient conditions t o replace Assumptions 6 and 7.

to obtain the asymptotic expansions of $\hat{J}_1$ and $\hat{J}_2$. Assumptions 6 and 7 are similar smoothness conditions for expanding $\hat{J}_3$. Finally Assumptions 4 and 5 establish the regularity (in the sense of Newey (1990)) of the estimators with which we are working.

Assumption 1 *$\hat{\beta}$ is asymptotically linear with influence function $\psi_\beta(y,x)$, $E[\psi_\beta(y,x)] = 0$.*

Assumption 2 *$M(\beta) = E(\partial m(y,x,\beta)_s/\partial\beta'))$ exists and is continuous in a neighborhood of β_0.*

Assumption 3 *$n^{-1/2}\sum_{i=1}^{n}[\{m(y_i,x_i,\beta)_s - \mu^*(\beta)\} - \{m(y_i,x_i,\beta_0)_s - J_s\}]$ is stochastically equicontinuous at $\beta = \beta_0$.*

Assumption 4 *For all parametric submodels $E_\theta[\| \ \psi_\beta(y,x) \ \|^2]$ exists and is continuous in a neighborhood of θ_0.*

Assumption 5 *For all parametric submodels, $\mu(\theta)$ is differentiable and $E_\theta[\| \ m(y,x,\beta_0)_s - J_s \ \|^2]$ exists and is continuous in a neighborhood of θ_0.*

Assumption 6 *$Q(\beta) = E[\partial J(x,\beta)_s/\partial\beta']$ exists and is continuous in a neighboihood of β_0.*

Assumption 7 *$n^{-1/2}\sum_{i=1}^{n}[\{J(x_i,\beta)_s - J^c(\beta)\} - \{J(x_i,\beta_0)_s - J_s\}]$ is stochastically equicontinuous at $\beta = \beta_0$.*

Theorem 7 *Under Assumptions 1-7 (and 1-5 with $m(y,x,\beta) = [s(y,x,\beta)'s(y,x,\beta)]$ and $m(y,x,\beta) = h(y,x,\beta)$), $\hat{J}_3$ is asymptotically at least as efficient as $\hat{J}_1$ and $\hat{J}_2$ as an estimator of J. Moreover, $\hat{J}_3$ attains the semiparametric efficiency bound.*

PROOF: see Appendix

This result is proven by noting that $\hat{J}_1$, $\hat{J}_2$, and $\hat{J}_3$ are semiparametric estimators of an expectation and then applying results from Brown and Newey (1992). The nonparametric component in our setup is the marginal distribution of x, and the main point of the proof is that $J(x,\beta_0) = E[(S'S)_s|x] = -E[h_s|x]$, making $\hat{J}_3$ the efficient estimator. In other words, $J(x, \beta)$ is equal to the projection of $s(y, x, \beta)'s(y, x, \beta)$ and $h(y, x, \beta)$ on the nonparametric tangent set.

Of course, in most applications, we are ultimately interested in something beyond accurate estimation of J. If we are using a variance estimate to compute a confidence interval for $\hat{\beta}$, then ultimately we want as low coverage error as possible. Obviously accurate estimation of J is only one component that might help in this regard, but as we will see in Section 3.5 it appears to be a potentially important component.

3.3 Simulation Estimators

Often the integration required to compute $J(x, \beta)$ in equation (3.1) is complicated, making the use of $\hat{J}_3$ unwieldy. However, closer inspection of equation (3.2) reveals that we are actually interested in $J(x_i, \hat{\beta})$ for $i = 1, \dots, n$. Since we already have our maximum likelihood estimate $\hat{\beta}$ and we are given $f(y|x, \beta)$, we have the exact density $f(y|x, \hat{\beta})$ used in computation of $J(x, \hat{\beta})$. Thus we can approximate $J(x_i, \hat{\beta})$ for each $i = 1, \dots, n$ by simulation without analytically solving the integral in equation (3.1). Let

$$\hat{J}_m(x_i, \hat{\beta}) = \frac{1}{m} \sum_{j=1}^{m} s(\hat{y}_{ij}, x_i, \hat{\beta})' s(\hat{y}_{ij}, x_i, \hat{\beta}),$$

where $\hat{y}_{i1}, \dots, \hat{y}_{im}$ are drawn randomly and independently for each i from the known distribution $f(y|x_i, \hat{\beta})$.[4] Using Theorem 5, we have a second way to approximate $J(x_i, \hat{\beta}) = -H(x_i, \hat{\beta})$,

$$\hat{H}_m(x_i, \hat{\beta}) = \frac{1}{m} \sum_{j=1}^{m} h(\hat{y}_{ij}, x_i, \hat{\beta}).$$

So two feasible simulation variance estimators for MLE (both analogous to $\hat{J}_3$) are

$$\hat{J}_4 = \frac{1}{n} \sum_{i=1}^{n} \hat{J}_m(x_i, \hat{\beta}) \tag{3.3}$$

$$\hat{J}_5 = -\frac{1}{n} \sum_{i=1}^{n} \hat{H}_m(x_i, \hat{\beta}). \tag{3.4}$$

[4] If the scores themselves must be simulated, then the following results could be extended possibly putting to use the techniques suggested in Hajivassiliou and McFadden (1992).

These simulation estimators are small sample approximations to $\hat{J}_3$ in the sense that for a fixed sample size n, $\hat{J}_4 \xrightarrow{p} \hat{J}_3$ and $\hat{J}_5 \xrightarrow{p} \hat{J}_3$ as $m \longrightarrow \infty$. Still consistency of these estimators does not require that the number of simulations increases to infinity. The next theorem gives consistency for any fixed positive number of simulations per observation.

Following Pakes and Pollard (1989), if we can express y as a function of x, β, and a random variable from a known distribution, then consistency of $\hat{J}_4$ and $\hat{J}_5$ is given by the following result.

Theorem 8 *Suppose conditions (i) and (iii) of* Theorem 6 *are satisfied, and suppose there is a random variable* $u \sim P(\cdot|x)$, P *a known distribution, and a known function* g *such that* $g(x,u,\beta) \sim f(\cdot|x,\beta)$. *Let* $k_s(x,u,\beta) = s(g(x,u,\beta),x,\beta)$ *and* $k_h(x,u,\beta) = h(g(x,u,\beta),x,\beta)$.
a) *If* $k_s(x,u,\beta)$ *is continuous at* β_0 *with probability one and if there exists a neighborhood* $\mathcal{N}$ *of* β_0 *such that* $E[\sup_{\beta\in\mathcal{N}} \| k_s(x,u,\beta) \|^2] < \infty$, *then for any fixed positive integer* m, $\hat{J}_4^{-1} \xrightarrow{p} J^{-1}$.
b) *Suppose conditions (ii) and (v) of Theorem 6 are satisfied. If* $k_h(x,u,\beta)$ *is continuous at* β_0 *with probability one and if there exists a neighborhood* $\mathcal{N}$ *of* β_0 *such that* $E[\sup_{\beta\in\mathcal{N}} \| k_h(x,u,\beta) \|] < \infty$, *then for any fixed positive integer* m, $\hat{J}_5^{-1} \xrightarrow{p} J^{-1}$.

PROOF: see Appendix

Verification of the dominance conditions in Theorem 8 usually can be accomplished by showing equivalence to the analogous dominance conditions needed for consistency of the outer product of the scores [o.p.s.] and the negative Hessian estimators. An example of this approach in a probit model follows.

Example 1: PROBIT

$$\begin{aligned} y^* &= X\beta + \epsilon \qquad \text{where } \epsilon \sim \mathrm{N}(0,1) \\ y &= \mathbf{1}(y^* > 0) \end{aligned}$$

Then,

$$\begin{aligned}
s(y,x,\beta) &= \frac{y-\Phi(x'\beta)}{\Phi(x'\beta)\Phi(-x'\beta)}\phi(x'\beta)x \\
g(x,u,\beta) &= \mathbf{1}(x'\beta+u) \qquad \text{where } u \sim \mathrm{N}(0,1) \\
k_s(x,u,\beta) &= \frac{\mathbf{1}(x'\beta+u)-\Phi(x'\beta)}{\Phi(x'\beta)\Phi(-x'\beta)}\phi(x'\beta)x
\end{aligned}$$

Now we usually prove $E[\sup_{\beta\in\mathcal{N}} \parallel s(y,x,\beta) \parallel^2] < \infty$ by noting that

$$\parallel s(y,x,\beta) \parallel \leq |y-\Phi(x'\beta)| \parallel \frac{\phi(x'\beta)x}{\Phi(x'\beta)\Phi(-x'\beta)} \parallel \leq 2 \parallel \frac{\phi(x'\beta)x}{\Phi(x'\beta)\Phi(-x'\beta)} \parallel$$

and showing $E[\sup_{\beta\in\mathcal{N}} \parallel \{\phi(x'\beta)x\}/\{\Phi(x'\beta)\Phi(-x'\beta)\} \parallel^2] < \infty$.[5]

Similarly $|\mathbf{1}(x'\beta+u)-\Phi(x'\beta)| \leq 2$ so $E[\sup_{\beta\in\mathcal{N}} \parallel \{\phi(x'\beta)x\}/\{\Phi(x'\beta)\Phi(-x'\beta)\} \parallel^2] < \infty$ is also sufficient to prove $E[\sup_{\beta\in\mathcal{N}} \parallel k_s(x,u,\beta) \parallel^2] < \infty$. Verification of the dominance condition for the Hessian simulation estimator follows similarly.

In Theorem 8, we see that both simulation variance estimators are consistent for *any* fixed positive m. If, for example, we look at $\hat{J}_4$ with $m = 1$, we find a formula ($\hat{J}_4 = n^{-1}\sum_{i=1}^n s(\hat{y}_{i1},x_i,\hat{\beta}))'s(\hat{y}_{i1},x_i,\hat{\beta})$) almost identical to the definition of $\hat{J}_1$. The only difference is that $\hat{J}_4$ uses values $\hat{y}_{i1}$ drawn from the conditional densities $f(y|x_i,\hat{\beta})$ whereas $\hat{J}_1$ uses the given data y_i drawn from the true conditional density $f(y|x_i,\beta_0)$. On the other hand, for large m, $\hat{J}_m(x_i,\hat{\beta})$ converges in probability to $J(x_i,\hat{\beta})$ by the Weak Law of Large Numbers. Hence, for large values of m, we expect $\hat{J}_4$ (and $\hat{J}_5$) to closely follow $\hat{J}_3$. These comparisons are formalized in the next section by looking at the asymptotic distribution of these estimators.

3.4 Asymptotic Behavior Comparisons

Before presenting a Monte Carlo comparison of the performance of the variance estimators discussed in this paper, we present a comparison of their influence functions

[5] $E[\parallel x \parallel^4] < \infty$ is sufficient to prove this step, see Newey and McFadden (1993)

and their asymptotic variances. The influence function for $\hat{J}_3$ is derived in the proof of Theorem 7. The derivation for the other estimators is similar, involving a standard asymptotic expansion and a stochastic equicontinuity assumption. The assumptions of Theorem 7 are sufficient for the validity of the influence functions for $\hat{J}_1$, $\hat{J}_2$, and $\hat{J}_3$ given below. The only additional assumptions needed for the simulation estimators are stochastic equicontinuity conditions.

Assumption 8 $n^{-1/2}\sum_{i=1}^{n}[\{\hat{J}_r(x_i,\beta)_s - J^c(\beta)\} - \{\hat{J}_r(x_i,\beta_0)_s - J_s\}]$ *is stochastically equicontinuous at* $\beta = \beta_0$.

Assumption 9 $n^{-1/2}\sum_{i=1}^{n}[\{\hat{H}_r(x_i,\beta)_s - H^c(\beta)\} - \{\hat{H}_r(x_i,\beta_0)_s - H_s\}]$ *is stochastically equicontinuous at* $\beta = \beta_0$, *where* $H^c(\beta) = E[H(x,\beta)_s] = -J^c(\beta)$.

The influence functions for the variance estimators follow. Let $D_1 = E(\partial[s(y,x,\beta_0)' s(y,x,\beta_0)]_s/\partial\beta')$ and $D_2 = E(\partial h(y,x,\beta_0)_s/\partial\beta')$.

$$\begin{aligned}
\phi_{J_1}(y,x) &= [S'S]_s - J_s + D_1 J^{-1}S \\
\phi_{J_2}(y,x) &= -[h_s - H_s + D_2 J^{-1}S] \\
\phi_{J_3}(y,x) &= J(x,\beta_0) - J_s + \{D_1 + E([S'S]_s S')\}J^{-1}S \\
&= -[H(x,\beta_0) -- H_s + \{D_2 + E(h_s S')\}J^{-1}S] \\
\phi_{J_4}(y,x) &= [\hat{J}_r(x,\beta_0) - J]_s + \{D_1 + E([S'S]_s S')\}J^{-1}S \\
\phi_{J_5}(y,x) &= -[\hat{H}_r(x,\beta_0) - H]_s - \{D_2 + E(h_s S')\}J^{-1}S
\end{aligned}$$

The derivations for these influence functions are given in the Appendix.

In light of the effiency result of Theorem 7, the influence functions can be used to determine the conditions under which the variance estimators are efficient. Setting $\phi_{J_1} = \phi_{J_3}$ $[\phi_{J_2} = \phi_{J_3}]$ and noting that $E(m|x) = J(x,\beta_0)$ $[H(x,\beta_0)]$, we find that $\hat{J}_1$ $[\hat{J}_2]$ is efficient when $m_s - E(m_s|x) = E(m_s S')J^{-1}S$ for $m = S'S$ $[m = h]$. The condition that $m_s - E(m_s|x)$ be a linear combination of the scores seems both difficult to interpret and unlikely to happen in practice.[6] Of course, efficiency of $\hat{J}_4$ $[\hat{J}_5]$

[6] When this condition is satisfied it often seems to be when $m_s = E(m_s|x)$. For example, if given i.i.d. data from a normal distribution and estimating unknown mean and variance parameters, then

occurs only if its simulation term is exactly equal to the integral it is simulating, ie $\hat{J}_r(x,\beta) = J(x,\beta)$ $[\hat{H}_r(x,\beta) = H(x,\beta)]$ which again is unlikely unless the conditional density or the moment used in estimation is degenerate.

The asymptotic variances of the variance estimators can also be obtained from the influence functions.

$$\begin{aligned}
\text{Var}([\hat{J}_1]_s) &= \{E([S'S]_s[S'S]_s{}') - J_sJ_s'\} + [D_1 + E([S'S]_sS')]J^{-1}[D_1 + E([S'S]_sS')]' \\
&\quad -E([S'S]_sS')J^{-1}E([S'S]_sS')' \\
\text{Var}([\hat{J}_2]_s) &= \{E(h_sh_s{}') - J_sJ_s'\} + [D_2 + E(h_sS')]J^{-1}[D_2 + E(h_sS')]' \\
&\quad -E(h_sS')J^{-1}E(h_sS')' \\
\text{Var}([\hat{J}_3]_s) &= \{E[E([S'S]_s|x)E([S'S]_s|x)'] - J_sJ_s'\} \\
&\quad +[D_1 + E([S'S]_sS')]J^{-1}[D_1 + E([S'S]_sS')]' \\
&= \{E[E(h_s|x)E(h_s|x)'] - J_sJ_s'\} + [D_2 + E(h_sS')]J^{-1}[D_2 + E(h_sS')]' \\
\text{Var}([\hat{J}_4]_s) &= \frac{1}{r}\{E([S'S]_s[S'S]_s{}') - J_sJ_s'\} + \frac{r-1}{r}\{E[E([S'S]_s|x)E([S'S]_s|x)'] - J_sJ_s'\} \\
&\quad +[D_1 + E([S'S]_sS')]J^{-1}[D_1 + E([S'S]_sS')]' \\
\text{Var}([\hat{J}_5]_s) &= \frac{1}{r}(E(h_sh_s{}') - J_sJ_s') + \frac{r-1}{r}(E[E(h_s|x)E(h_s|x)'] - J_sJ_s') \\
&\quad +[D_2 + E(h_sS')]J^{-1}[D_2 + E(h_sS')]'
\end{aligned}$$

The first thing to notice is that as the number of simulations, r, approaches infinity, the variances of $\hat{J}_4$ and $\hat{J}_5$ approach the effiency bound given by the variance of $\hat{J}_3$. Hence given enough simulation draws $\hat{J}_4$ and $\hat{J}_5$ are more efficient estimators of J than $\hat{J}_1$ and $\hat{J}_2$ (the exact number of simulations required depending on the particular problem). On the other hand if only one simulation draw is used the ranking goes in the opposite direction, so that $\hat{J}_1$ $[\hat{J}_2]$ is more efficient than $\hat{J}_4(1)$ $[\hat{J}_5(1)]$. This result is seen easily by noting that $\text{Var}([\hat{J}_4]_s) = (1/r)\text{Var}([\hat{J}_1]_s) + ((r-1)/r)\text{Var}([\hat{J}_3]_s) + (1/r)E([S'S]_sS')J^{-1}E([S'S]_sS')'$. Hence when $r = 1$, $\text{Var}([\hat{J}_4]_s)$ $[\text{Var}([\hat{J}_5]_s)]$ and $\text{Var}([\hat{J}_1]_s)$ $[\text{Var}([\hat{J}_2]_s)]$ differ by the nonnegative definite matrix $E(m_sS')J^{-1}E(m_sS')'$,

$\hat{J}_2 = \hat{J}_3 = \hat{J}_5$.

where $m = (S'S)_s$ $[m = h_s]$. Intuitively, $\hat{J}_1$ and $\hat{J}_2$ are using a single draw from the true conditional distribution of y given x, while $\hat{J}_4(1)$ and $\hat{J}_5(1)$ are using a single draw from the estimated conditional distribution $f(y|x, \hat{\beta})$. We would expect the estimators using the truth to do better. Finally, we see that $\hat{J}_3$ gains efficiency over $\hat{J}_1$ and $\hat{J}_2$ since $E(mm') - E[E(m|x)E(m|x)']$ is nonnegative definite.[7] Still, $\hat{J}_1$ and $\hat{J}_2$ gain relative to $\hat{J}_3$ through the nonpositive definite term $-E(m_s S')J^{-1}E(m_s S')'$. Theorem 7 tells us that asymptotically the former gains outweigh the latter, since $\hat{J}_3$ is efficient.

In the next section, Monte Carlo results reveal how these theoretical comparisons translate into confidence interval performance in specific example models.

3.5 Monte Carlo Results

3.5.1 Probit Model

In this section, results from application of the above variance estimators in the formation of confidence intervals for probit model parameters are presented. Because the model is simple, all of the estimators can be used and compared.

The first model estimated, model (P-A), is a very simple probit with a constant term and one regressor.

$$\begin{aligned} y^* &= \beta_0 + x\beta_1 + \epsilon \qquad \text{where } \epsilon \sim \mathrm{N}(0,1) \\ y &= \mathbf{1}(y^* > 0) \end{aligned}$$

The generated independent variables $x_1, \ldots, x_n$ are drawn from a standard normal (as the ϵ's are). Here and in the model that follows the x's and the ϵ's will be i.i.d. and independent of each other. Also, $\beta_0 = 1$ and $\beta_1 = 0.5$ will be the true parameter values used throughout.[8] A Newton-Raphson method is used to obtain the maximum likelihood estimates. Starting values for the parameters are zero.

[8] Additional Monte Carlo runs have shown that the general conclusions noted here are robust to the exact parametrization or confidence interval size.

Model (P-B) is a probit with heteroskedasticity of known form.

$$\begin{aligned} y^* &= \beta_0 + x\beta_1 + \epsilon \qquad \text{where } \epsilon \sim \mathrm{N}(0, \exp(2\gamma z)) \\ y &= \mathbf{1}(y^* > 0) \end{aligned}$$

The variables $z_1, \ldots, z_n$ are drawn independently of the x's from a standard normal. The model is generated with $\gamma = 0.1$ as the true parameter value. Starting values are obtained for β_0 and β_1 by estimating the homoskedastic model (P-A), and the starting value for γ is zero.

Model (P-A) is estimated for $n = 50$ and $n = 150$. Model (P-B) is estimated only for $n = 150$ because a sample size of 50 is not sufficient for estimation. Parameters are estimated using maximum likelihood. Then all of the variance estimators discussed so far are computed and used to form confidence intervals using the asymptotic normality approximation. For example the 95% confidence interval for β_0 is $\hat{\beta}_0 \pm 1.96\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_0)}$. In Tables 3.1 and 3.2, coverage rates are shown for the 90% and 95% confidence intervals of the parameters in models (P-A) and (P-B) using 10,000 Monte Carlo replications.

The results for confidence intervals using $\hat{J}_1^{-1}$, $\hat{J}_2^{-1}$, $\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$, and $\hat{J}_3^{-1}$ are in the fourth through seventh columns of Tables 3.1 and 3.2. Performance of these estimators will be discussed in terms of the error coverage rates, defined as the empirical coverage rate minus the nominal coverage rate. $\hat{J}_1^{-1}$, the o.p.s. estimator, performs relatively poorly in all the models. Gourieroux, Monfort, and Trognon (1984) discuss the robustness of the $H^{-1}JH^{-1}$ estimator, but for models (P-A) and (P-B) we find it performs rather poorly in terms of size. It seems to be particularly poor in providing good confidence intervals for slope parameters. $\hat{J}_2^{-1}$, the Hessian estimator, and $\hat{J}_3^{-1}$, the c.i.m. estimator, generally have the lowest error coverage rates. In the simplest model (P-A), $\hat{J}_3^{-1}$ performs better than $\hat{J}_2^{-1}$ uniformly, although the margin of improvement is often small. In model (P-B), the results are mixed. The Hessian estimator performs best in Table 3.1 iwth 90% confidence intervals, but $\hat{J}_3^{-1}$ outperforms $\hat{J}_2^{-1}$ for the constant term and the heteroskedasticity parameter in Table 3.2. Also $\hat{J}_1^{-1}$ improves its relative performance for the slope parameter, while maintaining relatively large error coverage rates for the other parameters.

Tables 3.1 and 3.2 also show that using $\hat{J}_1^{-1}$ leads to too large confidence intervals. The Hessian estimator seems to have the same tendency but not as extreme. Conversely, using $\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$ results in too small confidence intervals. $\hat{J}_3^{-1}$ has both positive and negative error coverage rates so that no similar general conclusion can be drawn.

The last four columns of Tables 3.1 and 3.2 contain the results using the simulation estimators $\hat{J}_4^{-1}$ and $\hat{J}_5^{-1}$. The numbers in the parentheses of the column headings are the values of m used in the simulation estimators. The coverage rates for $\hat{J}_4(1)^{-1}$ and $\hat{J}_5(1)^{-1}$ mimic closely the coverage rates for $\hat{J}_1^{-1}$ and $\hat{J}_2^{-1}$, respectively, suggesting that the term $E(m_sS')J^{-1}E(m_sS')'$, which represents the difference of these estimators as discussed in the last section, is close to zero. Both $\hat{J}_4(300)^{-1}$ and $\hat{J}_5(300)^{-1}$ use enough simulations to act much like $\hat{J}_3^{-1}$ although $\hat{J}_5(300)^{-1}$ seems to do a better job. As m increases the coverage rates for $\hat{J}_4^{-1}$ seem to change fairly monotonically from the neighborhood of the $\hat{J}_1^{-1}$ coverage rates to the neighborhood of the $\hat{J}_3^{-1}$ coverage rates. Such a change suggests optimizing the empirical coverage rate (to match the nominal coverage rate) with respect to m, although an attempt is not made here.

Since the true marginal distribution from which the x's are generated is known, the true asymptotic variance J^{-1} is obtained for model (P-A) using a numerical integration routine. The average matrix distance from the variance estimators to the true asymptotic variance is then reported in Table 3.3. Matrix distance is defined by the usual $\Re^k$-metric, $(\sum_{i=1}^k(a_i-b_i)^2)^{1/2}$. For the nonsimulation estimators, the results generally correspond to our conclusions from Tables 3.1 and 3.2. The exception is that $\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$ has the smallest average distance to J^{-1} for $n = 50$, while not having the smallest corresponding error coverage rates. For $n = 150$, the distance ranking matches the error coverage rate ranking for the standard variance estimators.

Table 3.3 provides a different perspective for the simulation estimators. As measured by distance to the true asymptotic variance, $\hat{J}_4^{-1}$ and $\hat{J}_5^{-1}$ are very similar to $\hat{J}_1^{-1}$ and $\hat{J}_2^{-1}$. However, for large $m = 300$, the distance seems to move to about halfway toward the $\hat{J}_3^{-1}$ average distance. Perhaps surprisingly, for $n = 50$, the $\hat{J}_4(300)^{-1}$ average distance to J^{-1} is twice the $\hat{J}_3^{-1}$ average distance and yet $\hat{J}_4(300)^{-1}$ still performs remarkably similar to $\hat{J}_3^{-1}$ in terms of coverage rates.

The probit Monte Carlo results seem to substantiate the two main points of the paper. First, the efficiency of the conditional information matrix estimator makes it the most attractive of the standard variance estimators. That efficiency advantage is exemplified by the relatively low error coverage rates of $\hat{J}_3^{-1}$. Second, the simulation estimators are acting as desired. In particular, for large m, they would seem to provide a suitable replacement for the c.i.m. estimator when such an estimator is not available, as in the next section.

3.5.2 Sample Selection Model

As previously noted, the choice of variance estimator is often determined by which is easiest or most feasible. Sample selection models are complicated enough that feasibility becomes an important concern. Most studies use Heckman's (1976, 1979) two-step estimation procedure. However, maximum likelihood estimation can provide significant efficiency gains as seen in Nelson (1984).

The exact model considered here is

$$\begin{aligned} y_1 &= \mathbf{1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1) \\ y_2 &= y_1[\alpha_0 + \alpha_1 x_1 + \varepsilon_2] \end{aligned} \tag{3.5}$$

$$\text{where} \begin{pmatrix} \varepsilon_2 \\ \varepsilon_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right).$$

This model is complicated enough that obtaining an analytical solution to the integral in equation (3.1) leaves the c.i.m. estimator, $\hat{J}_3^{-1}$, infeasible. Additionally, estimation of the correlation parameter ρ is difficult. Using Heckman's approach, estimates of ρ often end up outside the interval $[-1, 1]$, causing the restriction $-1 \leq \rho \leq 1$ to bind.

Correspondingly, maximum likelihood estimation inevitably entails a grid search to obtain correlation parameter estimates. The result of this difficulty is that the Hessian estimator $\hat{J}_2^{-1}$ becomes difficult to use since it is not always positive definite. For a given set of data, it may be possible to come up with an acceptable value of $\hat{J}_2^{-1}$, but for a Monte Carlo study where estimation occurs for hundreds of data sets, $\hat{J}_2^{-1}$ cannot be used. Thus of the estimators discussed in section 3.2, only the o.p.s. estimator, $\hat{J}_1^{-1}$, is available. For the reasons above, $\hat{J}_5^{-1}$, which corresponds to the Hessian estimator, is also unavailable, but the simulation estimator $\hat{J}_4^{-1}$ is always feasible (and easy to compute).

Estimation proceeds as suggested by Nawata (1993). Heckman's procedure is used to obtain initial estimates. Then for each value of ρ in the grid $(-.97, -.96, \ldots, .96, .97)$, a Newton-Raphson type method is employed to obtain estimates for all the parameters but ρ. The maximum of the likelihoods for this grid gives an approximation $\tilde{\rho}$ to ρ. The process is repeated with a finer grid (.001 between values) in a neighborhood of $\tilde{\rho}$, and we use this result as our estimate, $\hat{\rho}$. Tables 3.4 and 3.5 present the results for estimation of the sample selection model (3.5) with various parameter values. Confidence intervals are formed using the variance estimators $\hat{J}_1^{-1}$ and $\hat{J}_4^{-1}$, and coverage rates are obtained by using 1000 Monte Carlos.

Parametrization (S-A) is defined by the following: $\beta_0 = 1, \beta_1 = .5, \beta_2 = .5, \alpha_0 = 0, \alpha_1 = .5, \sigma^2 = 1, \rho = .3$. In parametrization (S-B), we change the slope parameters of the selection equation and increase the correlation between error terms in the two equations. The true parameter values used are $\beta_0 = 1, \beta_1 = .75, \beta_2 = .25, \alpha_0 = 0, \alpha_1 = .5, \sigma^2 = 1, \rho = .7$. Tables 3.4 and 3.5 show the 90% and 95% confidence interval coverage rates for the constant term and slope parameters of both equations. $\hat{J}_4(1)^{-1}$ follows $\hat{J}_1^{-1}$ quite closely. For the constant term confidence intervals, $\hat{J}_4(100)^{-1}$ makes inconsequential improvement upon the performance of $\hat{J}_1^{-1}$. However, the error coverage rates for the slope parameters using the large m simulation estimator are much smaller than the error coverage rates for the o.p.s. estimator. The gains from using a simulation estimator in this complicated model are clear and quite significant.

3.6 Conclusion

The usual consistency results for maximum likelihood variance estimators provide no guidance in choosing an estimator, so we first note that the conditional information matrix variance estimator achieves the semiparametric efficiency bound for the variance estimation problem. Monte Carlo results from a probit model are used to exemplify the efficiency advantage. In terms of confidence interval error coverage rates, the c.i.m. estimator performs considerably better than the outer product of the scores estimator or the $H^{-1}JH^{-1}$ estimator. It also seems to hold a slight advantage over the Hessian estimator.

In more complicated models, the c.i.m. estimator is not available since it requires evaluation of a burdensome integral. We suggest two simulation variance estimators that avoid direct computation of that integral but provide a close approximation to it. Conditions for consistency of these estimators are easy to check, and they turn out to be quite simple to use, especially if one is already using the o.p.s estimator or the Hessian estimator.

A sample selection model provides an example of a "complicated" model. In this model, the c.i.m. estimator is not available, and the Hessian estimator may not even always be practical. Hence confidence interval coverage rates are obtained using only the o.p.s. estimator and the corresponding simulation estimator. The Monte Carlo results strongly favor the simulation estimator.

Appendix

PROOF of Theorem 5: For $\beta \in \mathcal{N}$ and for all x a.s.,

$$\begin{aligned}
\int \frac{\partial^2 f(y|x,\beta)}{\partial\beta\partial\beta'}dy &= \frac{\partial \int \partial f(y|x,\beta)/\partial\beta dy}{\partial\beta'} \\
&= \frac{\partial^2}{\partial\beta\partial\beta'}\int f(y|x,\beta)dy \\
&= \frac{\partial^2}{\partial\beta\partial\beta'}(1) \\
&= 0
\end{aligned}$$

The first equality following by iii) and $\partial f/\partial\beta$ having a continuous derivative by i). The second equality follows similarly by ii) and f having a continuous derivative.[9] Since $f(y|x,\beta)$ is a member of a family of p.d.f.'s for given β and a.s.dx, the third equality follows.

Now note that

$$\begin{aligned}
H(x,\beta)+J(x,\beta) &= \int \left[\frac{\partial^2 \ln f(y|x,\beta)}{\partial\beta\partial\beta'} + \frac{\partial \ln f(y|x,\beta)}{\partial\beta}\right. \\
&\quad \left.\cdot\left(\frac{\partial \ln f(y|x,\beta)}{\partial\beta}\right)'\right] f(y|x,\beta)dy \\
&= \int \left[\left\{\frac{-1}{f(y|x,\beta)^2}\frac{\partial f(y|x,\beta)}{\partial\beta}\left(\frac{\partial f(y|x,\beta)}{\partial\beta}\right)'\right.\right. \\
&\quad \left. + \frac{1}{f(y|x,\beta)}\frac{\partial^2 f(y|x,\beta)}{\partial\beta\partial\beta'}\right\} + \frac{1}{f(y|x,\beta)^2}\frac{\partial f(y|x,\beta)}{\partial\beta} \\
&\quad \left.\cdot\left(\frac{\partial f(y|x,\beta)}{\partial\beta}\right)'\right] f(y|x,\beta)dy \\
&= \int \frac{\partial^2 f(y|x,\beta)}{\partial\beta\partial\beta'}dy \\
&= 0
\end{aligned}$$

so $H(x,\beta) = -J(x,\beta)$.

PROOF of Theorem 7: Under Assumptions 1-5, the efficient influence function $\psi_J^*(y,x)$ can be obtained from the projection of $m(y,x,\beta_0)$ on the tangent set for the problem. In Theorem 3 of BN, the efficiency bound obtained from the efficient influence function is then given. We will proceed by showing that $\hat{J}_3$ has the efficient influence function and

[9] see Newey and McFadden (1993) Lemma 3.6 for a basic result on switching the order of differentiation and integration.

thus attains the bound. Then the check that asymptotic covariance of $\hat{J}_3$ and $\hat{J}_1$ is equal to the asymptotic variance of $\hat{J}_3$ will complete the argument.

Note that Assumption 1 entails no additional restrictions when we are using the MLE $\hat{\beta}$. We will prove the theorem using $m(y,x,\beta) = [s(y,x,\beta)' s(y,x,\beta)]$. Let $M = M(\beta_0)$ and note that $Q(\beta_0) = E[\frac{\partial}{\partial\beta'}J(x,\beta_0)_s] = \int\int \frac{\partial}{\partial\beta'}(m(y,x,\beta_0)_s)f(y,x|\beta_0)dydx + \int\int m(y,x,\beta_0)_s\frac{\partial}{\partial\beta'}f(y|x,\beta_0)dyf(x)dx = M + E[m_sS']$. First we find the efficient influence function from BN Theorem 3. The nonparametric tangent set for this conditional maximum likelihood setup is $\mathcal{J}^* = \{\sqcup(x) : E[\| \sqcup(x) \|^2] < \infty,\ E[\sqcup(x)] = 0\}$, so $\text{Proj}(m(y,x,\beta_0)|\mathcal{J}^*) = E[m|x] - J = J(x,\beta_0) - J$ by standard projection results. So Theorem 3 in BN provides the efficient influence function $\phi^* = (J(x,\beta_0)_s - J_s) + (E[m_sS'] + M)J^{-1}S$.

Now we derive the influence function for $\hat{J}_3$ which follows a standard asymptotic expansion argument. Consider the asymptotic expansion for $J^c(\hat{\beta})$. By Assumption A.7, $J^c(\hat{\beta}) = J_s + Q(\beta_0)(\hat{\beta}-\beta_0) + o_p(\| \hat{\beta}-\beta_0 \|)$. Then,

$$\begin{aligned}
&\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} J(x_i,\hat{\beta})_s - J_s) \\
&= \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} J(x_i,\beta_0)_s - J_s) + \sqrt{n}(J^c(\hat{\beta}) - J_s) \\
&\quad + n^{-1/2}\sum_{i=1}^{n}[\{J(x_i,\hat{\beta})_s - J^c(\hat{\beta})\} - \{J(x_i,\beta_0)_s - J_s\}] \\
&= \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} J(x_i,\beta_0)_s - J_s) + Q(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} J^{-1}S_i \\
&\quad + n^{-1/2}\sum_{i=1}^{n}[\{J(x_i,\hat{\beta})_s - J^c(\hat{\beta})\} - \{J(x_i,\beta_0)_s - J_s\}] \\
&\quad + Q(\beta_0)o_p(1) + o_p(1) \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}[(J(x_i,\beta_0)_s - J_s) + (E[m_sS'] + M)J^{-1}S_i] + o_p(1).
\end{aligned}$$

The second equality follows from substituting in the influence function for $\hat{\beta}$. The last equality follows from Assumption 7. This straightforward derivation completes the proof that the influence function for $\hat{J}_3$ is exactly the efficient influence function given in Theorem 3 of BN. Hence $\hat{J}_3$ attains the semiparametric efficiency bound. While this conclusion also gives the relative efficiency result, we present the standard covariance approach for completeness.

Under Assumptions 1-3 and 5, Theorem 1 of BN gives the influence function for $\hat{J}_1$, which can be derived from the asymptotic expansion similar to the above derivation for $\hat{J}_3$. The result is that $\psi_{J_1}(y,x) = m(y,x,\beta_0)_s - J_s + M\psi_\beta(y,x) = m(y,x,\beta_0)_s - J_s + MJ^{-1}S$. Now we can compute the asymptotic covariance of $\hat{J}_1$ and $\hat{J}_3$.

Note that $E[(m_s - J_s)(E(m_s|x) - J_s)'] = E[E[(m_s - J_s)(E(m_s|x) - J_s)'|x]] = E[(E(m_s|x) - J_s)(E(m_s|x) - J_s)']$. Since S is orthogonal to the elements of $\mathcal{J}^*$, $E[MJ^{-1}S(E(m_s|x) - J_s)'] = 0$. Finally, $E[(E[m_sS'] + M)J^{-1}SS'J^{-1}M'] = (E[m_sS']+M)J^{-1}M'$ and $E[(E[m_sS']+M)J^{-1}S(m_s - J_s)'] = (E[m_sS'] + M)J^{-1}E[m_sS']'$. Hence $E[\psi_{J_1}(y,x\}\psi_J^*(y,x)'] = E[\psi_J^*(y,x)\psi_J^*(y,x)'] = V_J^*$. So the asymptotic efficiency of $\hat{J}_3$ relative to $\hat{J}_1$ is established.

Finally we note that we could have let $m(y,x,\beta) = h(y,x,\beta)$, and the same proof holds.

PROOF of Theorem 8: Fix m. Define $K_m^s(x,u_1,\ldots,u_m,\beta) = m^{-1}\sum_{j=1}^m k_s(x,u_j,\beta)k_s(x,u_j,\beta)'$. Since $k_s(x,u,\beta)$ is continuous at β_0 with probability one, $K_m^s(x,u_1,\ldots,u_m,\beta)$ is continuous at β_0 with probability one. Also,

$$\begin{aligned}
E[\sup_{\beta\in\mathcal{N}} \| K_m^s(x,u_1,\ldots,u_m,\beta) \|] &\le E[\sup_{\beta\in\mathcal{N}} \frac{1}{m}\sum_{j=1}^m \| k_s(x,u_j,\beta) \|^2] \\
&\le E[\frac{1}{m}\sum_{j=1}^m \sup_{\beta\in\mathcal{N}} \| k_s(x,u_j,\beta) \|^2] \\
&= E[\sup_{\beta\in\mathcal{N}} \| k_s(x,u,\beta) \|^2] \\
&< \infty
\end{aligned}$$

Now let $u_{i1},\ldots,u_{im}$ be a random sample from $P(u|x_i)$ and set $\hat{y}_{ij} = g(x_i,u_{ij},\hat{\beta})$ Then,

$$\begin{aligned}
\hat{J}_4 &= \frac{1}{n}\sum_{i=1}^n \frac{1}{m}\sum_{j=1}^m s(\hat{y}_{ij},x_i,\hat{\beta})s(\hat{y}_{ij},x_i,\hat{\beta})' \\
&= \frac{1}{n}\sum_{i=1}^n K_m^s(x_i,u_{i1},\ldots,u_{im},\hat{\beta}) \\
&\xrightarrow{p} E[K_m^s(x,u_1,\ldots,u_m,\beta_0)] \text{ by Lemma 6} \\
&= E[E[K_m^s(x,u_1,\ldots,u_m,\beta_0)|x]] \\
&= E[E[k_s(x,u_1,\beta_0)k_s(x,u_1,\beta_0)'|x]] \\
&= E[\int k_s(x,u,\beta_0)k_s(x,u,\beta_0)'P(u|x)du] \\
&= E[\int s(y,x,\beta_0)s(y,x,\beta_0)'f(y|x,\beta_0)dy] \\
&= E[J(x,\beta_0)] \\
&= J
\end{aligned}$$

And so by the Slutsky Theorem, $\hat{J}_4^{-1} \xrightarrow{p} J^{-1}$

Similarly, it can be shown that $\hat{J}_5 \xrightarrow{p} -H$. Conditions (ii) and (v) of Theorem 6 are sufficient to give the information matrix equality, which in turn gives the desired result, $\hat{J}_5^{-1} \xrightarrow{p} J^{-1}$.

Derivation of Influence Functions:

When $m(y,x,\beta) = [s(y,x,\beta)'s(y,x,\beta)]_s$, let $\mu_J(\beta) = \mu(\beta)$, and when $m(y,x,\beta) = h(y,x,\beta)_s$, let $\mu_H(\beta) = \mu(\beta)$.

$$\begin{aligned}
&[s(y,x,\hat\beta)'s(y,x,\hat\beta)]_s - J_s \\
&\quad = [S'S]_s - J_s + [\mu_J(\hat\beta) - J]_s + [s(y,x,\hat\beta)'s(y,x,\hat\beta)]_s \\
&\qquad -[S'S]_s - [\mu_J(\hat\beta) - J]_s \\
&\quad = [S'S]_s - J_s + \left\{\int\!\!\int \left[\frac{\partial[S'S]_s}{\partial\beta}\right] f(y|x,\beta_0)f(x)dydx\right\}(\hat\beta - \beta_0) + o_p(1) \\
&\quad = [S'S]_s - J_s + D_1 J^{-1} S + o_p(1)
\end{aligned}$$

And similarly for ϕ_{J_2}.

The first equality after ϕ_{J_3} is shown in the proof of Theorem 7. To obtain the second equality remember that we noted above that a generalization of the information matrix equality gives $J(x,\beta_0) = -H(x,\beta_0)$, so integration with respect to x gives $J^c(\beta) = -H^c(\beta)$. If we differentiate both sides of this last equality with respect to β and evaluate at β_0, then we get $D_1 + E([S'S]_s S') = -[D_2 + E(h_s S')]$, as desired.

$$\begin{aligned}
[\hat{J}_r(x,\hat\beta) - J]_s &= [\hat{J}_r(x,\beta_0) - J]_s + J^c(\hat\beta) - J_s \\
&\quad +\{[\hat{J}_r(x,\hat\beta) - \hat{J}_r(x,\beta_0)]_s - [J^c(\hat\beta) - J_s] \\
&= [\hat{J}_r(x,\beta_0) - J]_s + \{\int [\frac{\partial J(x,\beta_0)}{\partial\beta}] f(x)dx\}(\hat\beta - \beta_0) \\
&\quad +\{[\hat{J}_r(x,\hat\beta)_s - J^c(\hat\beta)] - [\hat{J}_r(x,\beta_0)_s - J_s]\} \\
&= [\hat{J}_r(x,\beta_0) - J]_s + \{D_1 + E([S'S]_s S')\} J^{-1} S + o_p(1)
\end{aligned}$$

And similarly for ϕ_{J_5}. □

Derivation of the Asymptotic Variances of the Variance Estimators: Var($\hat{J}_1$) and Var($\hat{J}_1$) come straight from their influence functions. Var($\hat{J}_3$) follows by noting that the score S is orthogonal to all elements of the nonparametric tangent set $\mathcal{J}^*$ and $J(x,\beta_0)_s - J_s = E([S'S]_s|x) - J_s \in \mathcal{J}^*$.

For Var($\hat{J}_4$) and Var($\hat{J}_5$) (with $m = [S'S]$ and $m = h$), we make the following observations,

$$\begin{aligned}
&E[m(\hat{y}_j, x, \beta_0)_s m(y, x, \beta_0)'_s] \\
&\quad = \int\int\int m(\hat{y}_j, x, \beta_0)_s m(y, x, \beta_0)'_s f(\hat{y}_j, y|x, \beta_0) f(x) d\hat{y}_j dy dx \\
&\quad = \int \left[\int m(\hat{y}_j, x, \beta_0)_s f(\hat{y}_j|x, \beta_0) d\hat{y}_j\right] \left[\int m(y, x, \beta_0)_s f(y|x, \beta_0) dy\right]' f(x) dx \\
&\quad = E[E(m_s|x)E(m_s|x)'] \qquad (3.6)
\end{aligned}$$

Thus, for example, with $m = [S'S]$,

$$\begin{aligned}
&E[\hat{J}_r(x, \beta_0)\hat{J}_r(x, \beta_0)'] \\
&\quad = E[(\frac{1}{r}\sum_{j=1}^{r} m(\hat{y}_j, x, \beta_0)_s)(\frac{1}{r}\sum_{k=1}^{r} m(\hat{y}_k, x, \beta_0)_s)'] \\
&\quad = \frac{1}{r^2}\sum_{j=1}^{r} E[m(\hat{y}_j, x, \beta_0)_s m(\hat{y}_j, x, \beta_0)'_s] \\
&\qquad + \frac{1}{r^2}\sum_{j \neq k} E[m(\hat{y}_j, x, \beta_0)_s m(\hat{y}_k, x, \beta_0)'_s] \\
&\quad = \frac{1}{r}E[m_s m'_s] + \frac{r-1}{r}E[E(m_s|x)E(m_s|x)'] \\
&\quad = \frac{1}{r}E([S'S]_s[S'S]_s{}') + \frac{r-1}{r}E[E([S'S]_s|x)E([S'S]_s|x)']
\end{aligned}$$

Also, $E[m(\hat{y}_j, x, \beta_0)_s s(y, x, \beta_0)'] = E[E(m_s|x)E(S|x)']$ follows as in (3.6), but the orthogonality of S and $E(m_s|x)$ along with the law of iterated expectations gives $0 = E[E(m_s|x)S'] = E[E(m_s|x)E(S|x)']$. □

References

Amemiya, T. (1985): *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

Brown, B. and W. Newey (1992): "Efficient Semiparametric Estimation of Expectations," manuscript.

Davidson, R. and J. MacKinnon (1984): "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, 25, 241-262.

Gourieroux, C., A. Monfort and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681-700.

Hajivassilou, V. and D. McFadden (1992): "The Method of Simulated Scores for the Estimation of LDV Models," manuscript.

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

_____ (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.

Lerman S. and C. Manski (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities." in *Structural Analysis of Discrete Data with Econometric Applications*, eds. C. Manski and D. McFadden. Cambridge, MA: MIT Press, 305-319.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995-1026.

Nawata, K. (1993) "A Monte Carlo Comparison of the Maximum Likelihood Estimator and Heckman's Two Step Estimator in Models with Sample Selection Biases," manuscript.

Nelson, F. (1984): "Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection," *Journal of Econometrics*, 24, 181-196.

Newey, W. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

Newey, W. and D. McFadden (1993): "Estimation and Hypothesis Testing in Large Samples," *Handbook of Econometrics.*

Pakes, A. and D. Pollard (1989): "The Asymptotic Distribution of Simulation Experiments," *Econometrica*, 57, 1027-1057.

Table 3.1: Coverage Rates for 90% Confidence Intervals of Probit Model Parameters[a]

Model	n	Parameter	$\hat{J}_1^{-1}$	$\hat{J}_2^{-1}$	$\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$	$\hat{J}_3^{-1}$	$\hat{J}_4(1)^{-1}$	$\hat{J}_4(300)^{-1}$	$\hat{J}_5(1)^{-1}$	$\hat{J}_5(300)^{-1}$
P-A	50	β_0	.9332	.9078	.8881	.9067	.9209	.9063	.9083	.9076
		β_1	.9319	.9048	.8612	.9015	.9221	.9017	.9007	.9012
	150	β_0	.9066	.9005	.8990	.9000	.9100	.9003	.9006	.9002
		β_1	.9099	.9030	.8895	.9028	.9086	.9022	.9028	.9030
P-B	150	β_0	.9101	.8976	.8916	.8967	.9202	.8970	.9022	.9000
		β_1	.9107	.8930	.8779	.8875	.9103	.8876	.8947	.8952
		γ	.9296	.9060	.8840	.8844	.9188	.8855	.9014	.8850

[a] When the optimization routine failed to converge for a generated data set, new pseudo-data was obtained. Model (P-A) $n = 50$ had three failures in 10000 Monte Carlo draws. Model (P-B) $n = 150$ had two failures. The other model had no failures.

Table 3.2: Coverage Rates for 95% Confidence Intervals of Probit Model Parameters

Model	n	Parameter	$\hat{J}_1^{-1}$	$\hat{J}_2^{-1}$	$\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$	$\hat{J}_3^{-1}$	$\hat{J}_4(1)^{-1}$	$\hat{J}_4(300)^{-1}$	$\hat{J}_5(1)^{-1}$	$\hat{J}_5(300)^{-1}$
P-A	50	β_0	.9694	.9668	.9415	.9655	.9578	.9655	.9658	.9653
		β_1	.9716	.9616	.9239	.9592	.9595	.9598	.9593	.9592
	150	β_0	.9562	.9525	.9487	.9523	.9534	.9528	.9525	.9524
		β_1	.9651	.9512	.9433	.9497	.9511	.9500	.9505	.9499
P-B	150	β_0	.9591	.9508	.9451	.9496	.9584	.9496	.9543	.9498
		β_1	.9561	.9473	.9342	.9435	.9553	.9439	.9482	.9434
		γ	.9710	.9578	.9416	.9426	.9583	.9428	.9505	.9426

Table 3.3: Matrix Distance From Variance Estimators To True Variance in Probit Model (P-A)

n	$\hat{J}_1^{-1}$	$\hat{J}_2^{-1}$	$\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$	$\hat{J}_3^{-1}$	$\hat{J}_4(1)^{-1}$	$\hat{J}_4(300)^{-1}$	$\hat{J}_5(1)^{-1}$	$\hat{J}_5(300)^{-1}$
50	51.0843	11.0917	4.1114	10.1868	50.7225	20.7142	10.9054	10.6406
150	1.39704	.942323	1.01752	.925955	2.08335	.930312	.956956	.926027

Table 3.4: Coverage Rates for 90% Confidence Intervals of Sample Selection Model Parameters[b]

Parametrization	n	Parameter	$\hat{J}_1^{-1}$	$\hat{J}_4(1)^{-1}$	$\hat{J}_4(100)^{-1}$
S-A	50	β_0	.946	.940	.918
		β_1	.952	.942	.890
		β_2	.940	.942	.872
		α_0	.996	.997	.994
		α_1	.931	.931	.899
S-B	50	β_0	.950	.941	.901
		β_1	.958	.943	.895
		β_2	.953	.949	.884
		α_0	.994	.993	.993
		α_1	.953	.944	.922

[b] There were zero failures in 1000 Monte Carlos for sample selection model parametrization (S-A) and one failure in 1000 runs for (S-B).

Table 3.5: Coverage Rates for 95% Confidence Intervals of Sample Selection Model Parameters

Parametrization	n	Parameter	$\hat{J}_1^{-1}$	$\hat{J}_4(1)^{-1}$	$\hat{J}_4(100)^{-1}$
S-A	50	β_0	.977	.974	.973
		β_1	.975	.969	.953
		β_2	.979	.968	.942
		α_0	.997	.997	.996
		α_1	.968	.966	.944
S-B	50	β_0	.981	.964	.960
		β_1	.982	.973	.948
		β_2	.982	.970	.954
		α_0	.994	.994	.994
		α_1	.981	.974	.967