Pairwise Difference Estimators for Nonlinear Models*

Bo E. Honoré
Princeton University

James L. Powell
University of California, Berkeley

November, 1997

Revised July, 2001

Abstract

This paper uses insights from the literature on estimation of nonlinear panel data models to construct estimators of a number of semiparametric models with a partially linear index, including the partially linear logit model, the partially linear censored regression model, and the censored regression model with selection.. We develop the relevant asymptotic theory for these estimators and we apply the theory to derive the asymptotic distribution of the estimator for the partially linear logit model. We evaluate the finite sample behavior of this estimator using a Monte Carlo study.

*This research is supported by the National Science Foundation. Ekaterini Kyriazidou gave useful comment on earlier versions. Part of this research was done while Honoré visited the Economics Department at the University of Copenhagen.

"I have had my solutions for a long time, but I do not yet know how I am to arrive at them." (Carl Friedrich Gauss)

1 Introduction.

For the linear panel data regression model with fixed effects,

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}, \tag{1}$$

in which the individual-specific intercept ("fixed effect") α_i can be arbitrarily related to the regressors x_{it}, a standard estimation approach is based upon "pairwise differencing" of the dependent variable y_{it} across time for a given individual to eliminate the fixed effect:

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + (\varepsilon_{it} - \varepsilon_{is}),$$

a form which eliminates the nuisance parameters $\{\alpha_i\}$ and is amenable to the usual estimation methods for linear regression models under suitable conditions on the error terms $\{\varepsilon_{it}\}$. For nonlinear models – i.e., models which are not additively-separable in the fixed effect α_i – this pairwise differencing approach is generally not applicable, and identification and consistent estimation of the β coefficients is problematic at best. Still, for certain nonlinear panel data models, variations of the "pairwise comparison" or "matching" approach can be used to construct estimators which circumvent the incidental-parameters problem caused by the presence of fixed effects; such models include the binary logit model (Rasch 1960, Chamberlain 1984), the censored regression model (Honoré 1992), and the Poisson regression model (Hausman, Hall and Griliches 1984).

Powell (1987, 2000) and Ahn and Powell (1993) exploited an analogy between the linear panel data model (1) and semiparametric linear selection models for cross-section data to derive consistent estimators for the latter. These estimators treat the additive "selection correction term" as analogous to the fixed effect in the linear panel data model, and eliminate selectivity bias by differencing observations with approximately-equal selection correction terms. The object of this paper is to extend this analogy between linear panel data models and linear selection models to those nonlinear panel data models, cited above, for which pairwise comparisons can be used to eliminate the fixed effects. This extension will yield consistent and asymptotically normal estimators for the linear regression coefficients in binary logit, censored regression, and Poisson regression models with additively-separable nonparametric components – i.e., nonlinear extensions of the "semilinear regression model" – and also for the censored regression model with sample selection.

In the next section, more details of the analogy between linear panel data models and semiparametric regression and selection models are provided, and the resulting pairwise difference estimators for the various nonlinear models are precisely defined. These estimators are all defined as minimizers of "kernel-weighted U-statistics"; some general results for consistency and asymptotic normality of such estimators are provided in Section 3. One novel feature of the general asymptotic theory is a "generalized jackknife" method for direct bias reduction for the estimator, which is a computationally-convenient alternative to the usual requirement that the kernel weights be of "higher-order bias-reducing" form. The paper then specializes these general results to the pairwise-difference estimator for the partially linear logit model, and presents the results of a Monte Carlo study to evaluate the finite-sample performance of this estimator.

2 Motivation for the Proposed Estimators.

In order to motivate the estimation approach proposed here, it is useful to first consider the partially linear model[1]

$$y_i = x_i\beta + g(w_i) + \varepsilon_i \qquad i = 1, \ldots, n \tag{2}$$

where (y_i, x_i, w_i) are observed, β is the parameter of interest and $g(\cdot)$ is an unknown function which is assumed to be "sufficiently smooth". A number of estimators of β have been proposed for this model.[2] The term $g(w_i)$ in (2) can represent "true nonlinearity" or may be the result of sample selection. For example, in the sample selection model (Type 2 Tobit model, in the terminology of Amemiya (1985)), the data is generated from

$$y_i^\star = x_i\beta + \varepsilon_i \tag{3}$$

$$d_i = 1\{w_i\gamma + \eta_i > 0\} \tag{4}$$

and the data consists of $y_i = d_i y_i^\star$, d_i, x_i and w_i. If it is assumed that (ε_i, η_i) is independent of (x_i, w_i), then we can write

$$y_i = x_i\beta + g(w_i\gamma) + \nu_i, \qquad E[\nu_i | x_i, w_i, d_i = 1] = 0$$

where $g(w_i\gamma) = E[\varepsilon_i | w_i\gamma + \eta_i > 0]$ and $\nu_i = \varepsilon_i - g(w_i\gamma)$.

Powell (1987, 2000) proposed estimation of (3) (and implicitly also of (2)) which is based on the idea that if $w_i\gamma$ equals $w_j\gamma$ then for observations i and j the terms $g(w_i\gamma)$ and $g(w_i\gamma)$ are like a fixed effect which can be differenced away. Since γ is typically unknown, and $w_i\gamma$ typically continuously distributed, a feasible version of this idea uses all pairs of observations and gives bigger weight to pairs for which $w_i\hat{\gamma}$ is close to $w_j\hat{\gamma}$, where $\hat{\gamma}$ is an estimator of γ. The weights are chosen in such a way that asymptotically, only pairs with $w_i\gamma - w_j\gamma$ in a shrinking neighborhood of 0, will matter.

---

[1] This model is also called the semiparametric regression model and the semilinear regression model.

[2] References.

The insight in this paper is to observe that this pairwise difference idea can be applied to any model for which it is possible to "difference out" a fixed effect. Below we outline some examples of models in which the idea can be used.

2.1 Partially Linear Logit Model.

The logit model with fixed effects is given by

$$y_{it} = 1\{\alpha_i + x_{it}\beta + \varepsilon_{it} \geq 0\} \qquad t = 1, 2 \ i = 1, \ldots, n$$

where $\{\varepsilon_{it} : t = 1, 2; i = 1, \ldots, n\}$ are *i.i.d.* logistically distributed random variables. In this model, Rasch (1960) observed that β can be estimated by maximizing the conditional log–likelihood (see also Chamberlain (1984), page 1274)

$$L = \sum_{i: y_{i1} \neq y_{i2}} -y_{i1} \ln\left(1 + \exp((x_{i2} - x_{i1})b)\right) - y_{i2} \ln\left(1 + \exp((x_{i1} - x_{i2})b)\right).$$

Now consider the partially linear logit model

$$y_i = \{x_i\beta + g(w_i) + \varepsilon_i \geq 0\} \qquad i = 1, \ldots, n \tag{5}$$

For observations with w_i close to w_j, the terms $g(w_i)$ and $g(w_j)$ are almost like fixed effects, provided that g is smooth. This suggests estimating β by maximizing

$$L_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{\substack{i<j \\ y_i \neq y_j}} -K\left(\frac{w_i - w_j}{h_n}\right)(y_i \ln(1+\exp((x_j - x_i)b)) + y_j \ln(1+\exp((x_i - x_j)b))). \tag{6}$$

where $K(\cdot)$ is a kernel which gives the appropriate weight to the pair (i, j), and h_n is a bandwidth which shrinks as n increases. L denotes the dimensionality of w_i and the term $\binom{n}{2}^{-1} \frac{1}{h_n^L}$ in front of $\sum$ will ensure that the objective function converges to a nondegenerate function under appropriate regularity conditions. The asymptotic theory will require that $K(\cdot)$ is chosen so that a number of regularity conditions, such as $K(u) \to 0$ as $|u| \to \infty$, are satisfied. The effect of the term $K\left(\frac{w_i - w_j}{h_n}\right)$ is to give more weight to comparisons of observations (i, j) for which z_i is close to z_j.

2.2 Partially Linear Tobit Models.

The fixed effects censored regression model is given by

$$y_{it} = \max\{0, \alpha_i + x_{it}\beta + \varepsilon_{it}\}$$

Honoré (1992) showed that with two observations for each individual, and with the error terms being i.i.d. for a given individual[3],

$$\beta = \arg\min_b E[q(y_{i1}, y_{i2}, (x_{i1} - x_{i2}b)] \tag{7}$$

where

$$q(y_1, y_2, \delta) = \begin{cases} \Xi(y_1) - (y_2 + \delta)\xi(y_1) & \text{if} \quad \delta \le -y_2; \\ \Xi(y_1 - y_2 - \delta) & \text{if} \quad -y_2 < \delta < y_1; \\ \Xi(-y_2) - (\delta - y_1)\xi(-y_2) & \text{if} \quad y_1 \le \delta; \end{cases}$$

and $\Xi(d)$ is given by[4] either $\Xi(d) = |d|$ or $\Xi(d) = d^2$. The estimators for the fixed effect censored regression model presented in Honoré (1992) are based on sample analogs of (7):

$$\frac{1}{n} \sum_i q(y_{i1}, y_{i2}, (x_{i1} - x_{i2})b)$$

Applying the logic above, this suggests estimating β in the partially linear censored regression model

$$y_i = \max\{0, x_i\beta + g(w_i) + \varepsilon_i\},$$

by minimization

$$S_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{i<j} K\left(\frac{w_i - w_j}{h_n}\right) q(y_i, y_j, (x_i - x_j)b)$$

Honoré (1992) also proposed estimators of the truncated regression model with fixed effects. In the simplest version of the truncated regression model, (y, x) is observed only when $y > 0$, where $y = x\beta + \varepsilon$.

The idea in Honoré (1992) is that if $y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}$ and if ε_{it} satisfies certain regularity conditions, then

$$E[r(y_{i1}, y_{i2}, (x_{i1} - x_{i2}b))|y_{i1} > 0, y_{i2} > 0] \tag{8}$$

is uniquely minimized at $b = \beta$, where

$$r(y_1, y_2, \delta) = \begin{cases} \Xi(y_1) & \text{if} \quad \delta \le -y_2; \\ \Xi(y_1 - y_2 - \delta) & \text{if} \quad -y_2 < \delta < y_1; \\ \Xi(-y_2) & \text{if} \quad y_1 \le \delta; \end{cases}$$

and $\Xi\,()$ is as above.

This suggests that the partially linear truncated regression model, $y_i = x_i\beta + g(w_i) + \varepsilon_i$ with (y_i, x_i, w_i) observed only when $y_i > 0$, can be estimated by minimizing

$$T_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{i<j} K\left(\frac{w_i - w_j}{h_n}\right) t(y_i, y_j, (x_i - x_j)b) \tag{9}$$

---

[3] The assumption on the error terms made in Honoré (1992) allowed for very general serial correlation. However, for the discussion in this paper we will restrict ourselves to the i.i.d. assumption.

[4] Other convex loss functions, $\Xi\,()$, could be used as well.

2.3 Partially Linear Poisson Regression Models.

As a third example, consider the Poisson regression model with fixed effects:

$$y_{it} \sim \text{po}\left(\exp(\alpha_i + x_{it}\beta)\right) \qquad t = 1, 2 \; i = 1, \ldots, n.$$

This model can be estimated by maximizing (see, for example, Hausman, Hall and Griliches (1984)):

$$L = \sum_i -y_{i1} \ln\left(1 + \exp((x_{i2} - x_{i1})b)\right) - y_{i2} \ln\left(1 + \exp((x_{i1} - x_{i2})b)\right)$$

A partially linear Poisson regression model

$$y_i \sim \text{po}\left(\exp(x_i\beta + g(w_i))\right) \qquad t = 1, 2 \; i = 1, \ldots, n$$

can then be estimated by maximizing

$$L = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{i<j} K\left(\frac{w_i - w_j}{h_n}\right) \left(-y_i \ln\left(1 + \exp((x_j - x_i)b)\right) - y_j \ln\left(1 + \exp((x_i - x_j)b)\right)\right).$$

2.4 Partially Linear Duration Models.

Finally, Chamberlain (1985) has shown how to estimate a variety of duration models with fixed effects. Using the objective functions that he suggested, one can derive objective functions, the minimization of which will result in estimators of the linear part of partially linear versions of the same models.

2.5 Tobit Models with Selection.

As mentioned above, the nonlinear term can sometimes to caused by sample selection. Consider for example a modification of the model defined by (3) and (4), in which $y_i^\star$ is censored. One example of this would be a model of earnings. Here, the variable of interest is often censored from above by topcoding at some observable constant c_i, as well as subject to sample selection, because not everybody works. This model can be written as

$$y_i^\star = \min\{x_i\beta + \varepsilon_i, c_i\} \tag{10}$$

$$d_i = 1\{w_i\gamma + \eta_i > 0\} \tag{11}$$

where (ε_i, η_i) is independent of (x_i, w_i) and the data consists of $y_i = d_i y_i^\star$, d_i, x_i and w_i. As usual, we can translate this model, which has right-censoring at c_i, to the Tobit framework with left-censoring at zero by taking $\tilde{y}_i \equiv d_i c_i - y_i$ as the dependent variable and $\tilde{x}_i \equiv -x_i$ as the regressor.

For the observations for which $d_i = 1$, the distribution of ε_i is conditional on $\eta_i > -w_i\gamma$. For two observations, i and j with $w_i\gamma = w_j\gamma$ and $d_i = d_j = 1$, ε_i and ε_j will be independent and identically distributed (conditionally on (x_i, x_j, w_i, w_j)). Therefore

$$\beta = \arg\min_b E[q(\tilde{y}_i, \tilde{y}_j, (\tilde{x}_i - \tilde{x}_j)b)|d_i = d_j = 1, w_i\gamma = w_j\gamma].$$

This suggests estimating β by minimization of

$$Q_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{\substack{i<j \\ d_i=d_j=1}} K\left(\frac{w_i\hat{\gamma} - w_j\hat{\gamma}}{h_n}\right) q(\tilde{y}_i, \tilde{y}_j, (\tilde{x}_i - \tilde{x}_j)b), \tag{12}$$

where $\hat{\gamma}$ is an estimator of γ in (11) (numerous estimators of $\hat{\gamma}$ have been considered in the literature), and L denotes the dimensionality of $w_i\gamma$. If there is no censoring, and if quadratic loss ($\Xi(d) = d^2$) is used, then the minimizer of (12) is the estimator suggested by Powell (1989).

The estimator of the truncated regression model defined in (9) requires that the error, ε_i has a log–concave density. Whether it is possible to define an estimator for the truncated regression model with selection by

$$R_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n^L} \sum_{\substack{i<j \\ d_i=d_j=1}} K\left(\frac{w_i\hat{\gamma} - w_j\hat{\gamma}}{h_n}\right) r(\tilde{y}_i, \tilde{y}_j, (\tilde{x}_i - \tilde{x}_j)b), \tag{13}$$

depends on whether one is willing to assume that the conditional density of ε_i given $\eta_i > k$ is log–concave for all k. The estimators for the partially linear logit and partially linear Poisson regression models do not generalize in a straightforward manner to the case of selection, because the error–terms after selection will have non–logistic and non–Poisson distributions.

3 Asymptotic Properties of Estimators defined by Minimizing Kernel–Weighted U–Statistics.

The estimators defined in the previous section are all defined by minimizing objective functions of the form

$$Q_n(\widehat{\gamma}, b) = \binom{n}{2}^{-1} \sum_{i<j} q_n(z_i, z_j; \widehat{\gamma}, b) \tag{14}$$

with

$$q_n(z_i, z_j; \gamma, b) = \frac{1}{h_n^L} K\left(\frac{(w_i - w_j)\gamma}{h_n}\right) s(v_i, v_j; b) \tag{15}$$

$z_i = (y_i, x_i, w_i)$, and $\nu_i = (y_i, x_i)$. Note that for the estimators of the partially linear models, $\gamma = I$ (the identity matrix). Let $\theta = (\gamma', \beta')'$ and $L = \dim(w\gamma)$. In the next two subsections we give conditions under which such estimators are consistent, and asymptotically normal around some pseudo–true value. The third subsection will then show how to "jackknife" in such a way that the estimator is asymptotically normal around the true value. Finally, the last subsection will show how to consistently estimate the asymptotic variance of the estimator.

Throughout, we define $\triangle w_{ij} = w_i - w_j$.

3.1 Consistency.

We will present two sets of assumptions under which the estimators defined by minimization of (14) are consistent. One set of assumptions will require a compact parameter space, whereas the other will assume a convex objective function. In both cases we will use the theorems found in Newey and McFadden (1994) to prove consistency.

Let m be a function of z_i, z_j and b. It is useful to define two function k_m and ℓ_m by

$$k_m(a_1, a_2, b) = E\left[m(z_i, z_j; b) \middle| z_i = a_1, w_j'\gamma_0 = a_2 \right]$$

and

$$\ell_m(a_1, a_2, b) = E\left[m(z_i, z_j; b) \middle| z_i = a_1, w_j'\gamma_0 = a_2 \right] f_{w_i'\gamma_0}(a_2)$$

When m depends only on z_i, z_j and b, we will write

$$k_m(a_1, a_2, b) = E\left[m(z_i, z_j; b) \middle| v_i = a_1, w_j'\gamma_0 = a_2 \right]$$

and

$$\ell_m(a_1, a_2, b) = E\left[m(v_i, v_j; b) \middle| v_i = a_1, w_j'\gamma_0 = a_2 \right] f_{w_i'\gamma_0}(a_2).$$

Assumption 1. All of the following assumptions are made on the distribution of the data

1. $E\left[s(v_i, v_j; b)^2 \right] < \infty$;
2. $E\left[\|\triangle w_{ij}\|^2 \right] < \infty$; and
3. $w_i'\gamma_0$ is continuously distributed with bounded density, $f_{w_i'\gamma_0}(\cdot)$, and $k_s()$ defined above exists and is a continuous function of each of its arguments
4. for all b, $|\ell_s(a_1, a_2, b)| \leq k_1(a_1, b)$ with $E[k_1(v_i, b)] < \infty$.

Assumption 2. One of the following assumptions is made on the bandwidth

1. $h_n > 0$, $h_n = o(1)$ and $h_n^{-1} = O(n^{1/2L})$;
2. $h_n > 0$, $h_n = o(1)$ and $h_n^{-1} = o\left(n^{1/2(L+1)}\right)$

Assumption 3. K is bounded, differentiable with bounded derivative K', and of bounded variation. Furthermore, $\int K(u)\, du = 1$, $\int |K(u)|\, du < \infty$ and $\int |K(\eta)|\, \|\eta\|\, d\eta < \infty$.

The assumptions made on the kernel are satisfied for many commonly used kernels.

In some of the example γ will be estimated.

Assumption 4. One of the following assumptions is made on $\widehat{\gamma}$

1. $\widehat{\gamma} = \gamma_0$
2. $\widehat{\gamma} = \gamma_0 + O_p\left(\frac{1}{\sqrt{n}}\right)$

The following additional assumption will be made in the case where the objective function is not convex.

Assumption 5. $|s(v_i, v_j; b_1) - s(v_i, v_j; b_2)| \leq B_{ij} |b_1 - b_2|^{\alpha}$ for some $\alpha > 0$, where $E\left[B_{ij}^2\right] < \infty$.

3.1.1 Limiting Objective Function.

Consistency of extremum estimators is usually proved by studying the limiting objective function and the exact manner in which the objective function approaches its limit. In this case the limiting objective function is

$$Q(\gamma_0, b) = E\left[\ell_s\left(v_i, w_i'\gamma_0, b\right)\right]$$

which exists under Assumption 1.

3.1.2 Pointwise Convergence to Limiting Objective Function.

In this section we will state conditions under which the objective function converges pointwise to its limit. It is useful to distinguish between the case with and with out a preliminary estimator (i.e., between Assumptions 4(1) and 4(2)).

In the case of no preliminary estimator, we have

$$\begin{aligned} E\left[Q_n\left(\gamma_0, b\right)\right] &= E\left[\frac{1}{h_n^L} K\left(\frac{w_i'\gamma_0 - w_j'\gamma_0}{h_n}\right) s\left(v_i, v_j; b\right)\right] \\ &= E\left[E\left[\frac{1}{h_n^L} K\left(\frac{w_i'\gamma_0 - w_j'\gamma_0}{h_n}\right) k_s\left(v_i, w_j'\gamma_0, b\right)\middle| w_i, v_i\right]\right] \\ &= \int\int K(\eta)\, \ell_s\left(v_i, w_i'\gamma_0 - h_n\eta, b\right) d\eta dF\left(w_i, u_i\right) \\ &\longrightarrow Q(\gamma_0, b) \end{aligned}$$

by dominated convergence. Note that the first expectation exists because of Assumptions 1(1) and 3.

Under Assumptions 1, 2(1) and 3

$$E\left[\left\{\frac{1}{h_n^L}K\left(\frac{\triangle w_{ij}\gamma_0}{h_n}\right)s\left(v_i,v_j;b\right)\right\}^2\right]=O(n)$$

and therefore (Ahn and Powell (1993), Lemma A.3)

$$Q_n\left(\gamma_0,b\right)-E\left[Q_n\left(\gamma_0,b\right)\right]=o_p\left(1\right)$$

Combining,

$$Q_n\left(\gamma_0,b\right)\longrightarrow Q(\gamma_0,b)$$

In the case where $\widehat{\gamma}=\gamma_0+O_p\left(\frac{1}{\sqrt{n}}\right)$, we have

$$Q_n\left(\widehat{\gamma},b\right)=\left(Q_n\left(\widehat{\gamma},b\right)-Q_n\left(\gamma_0,b\right)\right)+Q_n\left(\gamma_0,b\right)$$

Pointwise convergence of $Q_n\left(\widehat{\gamma},b\right)$ to $Q(\gamma_0,b)$ then follows from

$$\begin{aligned}
Q_n\left(\widehat{\gamma},b\right)-Q_n\left(\gamma_0,b\right) &= \left|\binom{n}{2}^{-1}\sum_{i<j}\frac{1}{h_n^L}\left(K\left(\frac{\triangle w_{ij}\widehat{\gamma}}{h_n}\right)-K\left(\frac{\triangle w_{ij}\gamma_0}{h_n}\right)\right)s\left(v_i,v_j;b\right)\right| \\
&= \left|\binom{n}{2}^{-1}\sum_{i<j}\frac{1}{h_n^L}K'\left(c_{ij}^*\right)\frac{\triangle w_{ij}\left(\widehat{\gamma}-\gamma_0\right)}{h_n}s\left(v_i,v_j;b\right)\right| \\
&\leq \binom{n}{2}^{-1}\sum_{i<j}\frac{1}{h_n^L}\left|K'\left(c_{ij}^*\right)\right|\frac{\left\|\triangle w_{ij}\right\|\left\|\widehat{\gamma}-\gamma_0\right\|}{h_n}\left|s\left(v_i,v_j;b\right)\right| \\
&\leq \left\|\widehat{\gamma}-\gamma_0\right\|\frac{1}{h_n^{L+1}}C\binom{n}{2}^{-1}\sum_{i<j}\left\|\triangle w_{ij}\right\|\left|s\left(v_i,v_j;b\right)\right| \\
&= O_p\left(\frac{1}{n^{1/2}h_n^{L+1}}\right) \\
&= o_p\left(1\right)
\end{aligned}$$

where the last equality follows from Assumption 2(2).

3.1.3 Uniform Convergence to Limiting Objective Function.

With a convex objective function, the pointwise convergence suffices. This covers the cases where γ is estimated. Without convex objective functions, uniform convergence is the key ingredient in the proof of consistency of extremum estimators. Invoking Assumption 1a, uniform convergence follows from Lemma 2.9 of Newey and McFadden (1994) as follows.

With no preliminary estimation of γ we have

$$\begin{aligned}
\left|Q_n\left(\widehat{\gamma}, b_1\right) - Q_n\left(\widehat{\gamma}, b_2\right)\right| &= \left|Q_n\left(\gamma_0, b_1\right) - Q_n\left(\gamma_0, b_2\right)\right| \\
&= \left|\binom{n}{2}^{-1} \sum_{i<j} \frac{1}{h_n^L} K\left(\frac{\triangle w_{ij}\gamma_0}{h_n}\right)\left(s\left(v_i, v_j; b_1\right) - s\left(v_i, v_j; b_2\right)\right)\right| \\
&\leq \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{h_n^L}\left|K\left(\frac{\triangle w_{ij}\gamma_0}{h_n}\right)\right| B_{ij}\left|b_1 - b_2\right|^{\alpha} \\
&= O_p(1)\left|b_1 - b_2\right|^{\alpha}
\end{aligned}$$

When γ is estimated,

$$\begin{aligned}
& \left|Q_n\left(\widehat{\gamma}, b_1\right) - Q_n\left(\widehat{\gamma}, b_2\right)\right| \\
\leq\; & \left|Q_n\left(\gamma_0, b_1\right) - Q_n\left(\gamma_0, b_2\right)\right| + \left|\left(Q_n\left(\widehat{\gamma}, b_1\right) - Q_n\left(\widehat{\gamma}, b_2\right)\right) - \left(Q_n\left(\gamma_0, b_1\right) - Q_n\left(\gamma_0, b_2\right)\right)\right|
\end{aligned}$$

and

$$\begin{aligned}
& \left|\left(Q_n\left(\widehat{\gamma}, b_1\right) - Q_n\left(\widehat{\gamma}, b_2\right)\right) - \left(Q_n\left(\gamma_0, b_1\right) - Q_n\left(\gamma_0, b_2\right)\right)\right| \\
=\; & \left|\binom{n}{2}^{-1} \sum_{i<j} \frac{1}{h_n^L}\left(K\left(\frac{\triangle w_{ij}\widehat{\gamma}}{h_n}\right) - K\left(\frac{\triangle w_{ij}\gamma_0}{h_n}\right)\right)\left(s\left(v_i, v_j; b_1\right) - s\left(v_i, v_j; b_2\right)\right)\right| \\
=\; & \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{h_n^L}\left|K'\left(c_{ij}^*\right)\right| \frac{\left\|\triangle w_{ij}\right\|\left\|\widehat{\gamma} - \gamma_0\right\|}{h_n}\left(s\left(v_i, v_j; b_1\right) - s\left(v_i, v_j; b_2\right)\right) \\
\leq\; & \frac{1}{h_n^{L+1}}\left\|\widehat{\gamma} - \gamma_0\right\| \binom{n}{2}^{-1} \sum_{i<j}\left|K'\left(c_{ij}^*\right)\right|\left\|\triangle w_{ij}\right\| B_{ij}\left|b_1 - b_2\right|^{\alpha} \\
=\; & O_p\left(\frac{1}{h_n^{L+1} n^{1/2}}\right)\left|b_1 - b_2\right|^{\alpha}
\end{aligned}$$

3.1.4 Identification.

The limiting objective function is uniquely minimized at β_0 provided that

Assumption 6. $E\left[s\left(v_i, v_j; b\right) \mid \left(w_i - w_j\right)\gamma_0 = 0\right]$ is uniquely minimized at $b = \beta_0$.

3.1.5 Consistency Theorem.

Combining these results, and referring to Newey and McFadden (1994) Theorem 2.7 and to Theorem 2.1 and Lemma 2.9, respectively, we have

Theorem 1 *If* $K\left(\frac{(w_i-w_j)\gamma}{h_n}\right) s\left(v_i, v_j; b\right)$ *is a continuous and convex function of* b *and the parameter space for* β *is a convex set with the true value,* β_0*, it its interior, then the minimizer of* (14) *over the parameter space is consistent under random sampling and assumptions 1, 3, 6 and either 2(1) and (4(1) or 2(2) and 4(2).*

Theorem 2 *If* $K\left(\frac{(w_i-w_j)\gamma}{h_n}\right) s\left(v_i, v_j; b\right)$ *is a continuous function of* b *and the parameter space for* β *is compact and includes the true value,* β_0*, then the minimizer of* (14) *over the parameter space is consistent under random sampling and assumptions 1, 3, 5, 6 and either 2(1) and (4(1) or 2(2) and 4(2).*

3.2 Asymptotic Normality.

In this section we derive the asymptotic distribution of the estimator defined by minimizing (14). Specifically, we will derive the limiting distribution of $\sqrt{n}\left(\widehat{\beta}-\beta_h\right)$ where β_h is the minimizer of $E\left[\frac{1}{h^L} K\left(\frac{(w_i-w_j)'\gamma_0}{h}\right) s\left(v_i, v_j; \beta\right)\right]$. Note that the argument that lead to consistency of $\widehat{\beta}$ implies that $\beta_h \to \beta_0$.Also note that β_h is non–stochastic. In section 3.3 we will discuss conditions under which

$$\beta_h = \beta_0 + \sum_{l=1}^{p} b_l h^l + o(h^p).$$

In this case, the estimator will have an asymptotic bias term, which we will eliminate via a jackknife approach. The advantage of the approach taken in this case is that it is not necessary to employ a bias–reducing kernel. This means that if s in equation (14) is convex, then so is the objective function Q_n in (15). At first sight, is seems that the disadvantage is that it is necessary to calculate the estimator a number of times. However, as we will see in section 3.3, it is often possible to do the optimization only once. Estimators that are asymptotically equivalent to the remaining estimators can then be defined as the result of performing a finite number of Newton_Raphson steps from the original estimator.

The following assumption is standard.

Assumption 7. The true parameter, β_0, is an interior point of the parameter space

In all the applications considered here, the objective function will be left– and right– differentiable. We therefore define

$$G_n\left(\widehat{\gamma}, \beta\right) = \binom{n}{2}^{-1} \sum_{i<j} p_n\left(z_i, z_j; \widehat{\gamma}, \beta\right)$$

where

$$p_n\left(z_i, z_j; \gamma, \beta\right) = \frac{1}{h_n^L} K\left(\frac{\left(w_i - w_j\right)' \gamma}{h_n}\right) t\left(v_i, v_j; \beta\right)$$

and $t(v_i, v_j; \beta)$ is a convex combination of the left– and right derivatives of $s(v_i, v_j; \beta)$.

It is useful to define

$$\begin{aligned} p_{1n}(z_i; \gamma, \beta) &= E[p_n(z_i, z_j; \gamma, \beta) | z_i] - E[p_n(z_i, z_j; \gamma, \beta)] \\ p_{0n}(\gamma, \beta) &= E[p_n(z_i, z_j; \gamma, \beta)] \end{aligned}$$

where the Assumption 8 below will guarantee that the expectations exist. We can then write

$$\begin{aligned} \binom{n}{2}^{-1} \sum_{i<j} p_n(z_i, z_j; \gamma, \beta) &= p_{0n}(\gamma, \beta) + \frac{2}{n} \sum_{i=1}^{n} p_{1n}(z_i; \gamma, \beta) \\ &\quad + \binom{n}{2}^{-1} \sum_{i<j} p_{2n}(z_i, z_j; \gamma, \beta) \end{aligned} \tag{16}$$

where p_{1n} and p_{2n} are P–degenerate (with P denoting the distribution of z_i). p_{2n} is defined implicitly by (16).

We will assume

Assumption 8. $\{t(\cdot, \cdot; \beta) : \beta \in B\}$ is Euclidean for an envelope F, i.e.

$$\sup_{n,\beta} |t(z_i, z_j; \beta)| \leq F(z_i, z_j),$$

satisfying $E\left[F^2\right] < \infty$. B need not be the whole parameter space, but could be some other set with β_0 in its interior.

Assumption 8 is satisfied for all of the examples considered in section 2. Assumptions 3 and 8 imply that $h_n^L p_n$ is Euclidean (for some envelope CF with $E\left[(CF)^2\right] < \infty$). This, in turn, implies that $h_n^L p_{2n}$ is Euclidean for an envelope with finite second moments (see Sherman (1994b), Lemma 6). This will be important for the derivation below,because it allows us to ignore the "error" when we approximate the U–statistic, $\binom{n}{2}^{-1} \sum_{i<j} p_n(z_i, z_j; \gamma, \beta)$, by its projection, $p_{0n}(\gamma, \beta) + \frac{2}{n} \sum_{i=1}^{n} p_{1n}(z_i; \gamma, \beta)$.

We also define

$$\widetilde{p}_n(z_i, \gamma, \beta) = p_{0n}(\gamma, \beta) + 2p_{1n}(z_i; \gamma, \beta)$$

Note that this implies that $\widetilde{p}_n(z_i, \gamma_0, \beta_h) = 2p_{1n}(z_i, \gamma_0, \beta_h)$. It is convenient to assume:

Assumption 9.

1. $\widetilde{p}_n(z_i, \gamma, \beta)$ is continuously differentiable in (γ, β) with a derivative $\widetilde{p}'_n(z_i, \gamma, \beta)$ with the property that for any sequence (γ^*, β^*) that converges in probability to (γ_0, β_0), $\widetilde{p}'_n(z_i, \gamma^*, \beta^*)$ converges to a matrix $\widetilde{p}'_0(\gamma_0, \beta_0)$, the lower part of which (i.e., the part that corresponds to differentiation with respect to β) is non–singular.

2. for some function $p_1(z_i;\gamma_0,\beta_0)$ with $E\left[\|p_1(z_i;\gamma_0,\beta_0)\|^2\right]<\infty$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{p}_n(z_i;\gamma_0,\beta_h)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_n(z_i;\gamma_0,\beta_0)=o_p(1).$$

In the next subsection, we give result that are useful for verifying Assumption 9.

We can write

$$\begin{aligned}
G_n(\gamma,\beta) &= \binom{n}{2}^{-1}\sum_{i<j}p_n(z_i,z_j;\gamma,\beta)\\
&= p_{0n}(\gamma,\beta)+\frac{2}{n}\sum_i p_{1n}(z_i;\gamma,\beta)+\binom{n}{2}^{-1}\sum_{i<j}p_{2n}(z_i,z_j;\gamma,\beta)\\
&= \frac{1}{n}\sum_i\widetilde{p}_n(z_i;\gamma,\beta)+\binom{n}{2}^{-1}\sum_{i<j}p_{2n}(z_i,z_j;\gamma,\beta)\\
&= \left(\frac{1}{n}\sum_i\widetilde{p}_n'(z_i;\theta^*)\right)\begin{pmatrix}\gamma-\gamma_0\\ \beta-\beta_h\end{pmatrix}+\frac{1}{n}\sum_i\widetilde{p}_n(z_i;\gamma_0,\beta_h)\\
&\quad+\binom{n}{2}^{-1}\sum_{i<j}p_{2n}(z_i,z_j;\gamma,\beta)
\end{aligned}$$

where, as usual, $\widetilde{p}_n'(z_i;\theta^*)$ should be interpreted as the derivative of $\widetilde{p}_n(z_i;\cdot)$ evaluated at a point (which may be different for different rows of $\widetilde{p}_n'$) between (γ,β) and (γ_0,β_h).

Since $\{p_{2n}(z_i,z_j;\gamma,\beta)\}$ is Euclidean, Sherman (1994), Theorem 3, can be applied to the function $h^L p_{2n}(z_i,z_j;g,b)$ with $\Theta_n=\{\theta:\|\theta-\theta_0\|\le c\}$ for some constant c, $\gamma_n=1$ (and noting that $k=2$) to get

$$\sup_{\Theta_n}\binom{n}{2}^{-1}\sum_{i<j}h^L p_{2n}(z_i,z_j;g,b)=O_p\left(\frac{1}{n}\right)$$

or

$$\sup_{\Theta_n}\binom{n}{2}^{-1}\sum_{i<j}p_{2n}(z_i,z_j;g,b)=O_p\left(\frac{1}{h^L n}\right).$$

where the assumption on the envelope guarantee that $E\left[\sup_{\Theta_n}p_{2n}(z_i,z_j;\gamma,\beta)^2\right]<\infty$ (this is condition (ii) for Sherman's Theorem)

This yields

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta}-\beta_h\right) &= \left(-\frac{1}{n}\sum_i\widetilde{p}_n^{\beta}(z_i;\theta^*)\right)^{-1}\left[\left(\frac{1}{n}\sum_i\widetilde{p}_n^{\gamma}(z_i;\theta^*)\right)\sqrt{n}(\widehat{\gamma}-\gamma_0)\right.\\
&\quad\left.+\frac{2}{\sqrt{n}}\sum_i p_{1n}(z_i,\gamma_0,\beta_h)+O_p\left(\frac{1}{h^L\sqrt{n}}\right)-\sqrt{n}G_n(\gamma,\beta)\right]
\end{aligned}$$

We therefore have

Theorem 3 *If* $\widehat{\beta}$ *is a consistent estimator of* β, $G_n\left(\widehat{\gamma},\widehat{\beta}\right) = o_p\left(n^{-1/2}\right)$, $1/h_n = o_p\left(n^{1/2L}\right)$, $\sqrt{n}\left(\widehat{\gamma}-\gamma_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega_i + o_p\left(1\right)$ *and Assumptions 7, 8 and 9 are satisfied, then*

$$\sqrt{n}\left(\widehat{\beta}-\beta_h\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i + o_p\left(1\right)$$

where

$$\psi_i = -p_0^{\beta}\left(\gamma_0,\beta_0\right)^{-1}p_0^{\gamma}\left(\gamma_0,\beta_0\right)\omega_i - 2p_0^{\beta}\left(\gamma_0,\beta_0\right)^{-1}p_1\left(z_i;\gamma_0,\beta_0\right)$$

3.2.1 Verifying some of the conditions.

Theorem 3 makes some high level assumptions. In this section we will present some results which will be useful in verifying these assumptions.

The following Lemma, which follows immediately from Lemma 1 in Honoré and Powell (1994), is useful for verifying that $G_n\left(\widehat{\gamma},\widehat{\beta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$

Lemma 4 *If the true parameter value,* β_0, *is an interior point in the parameter space, and*

1. $s\left(v_i,v_j;\beta\right)$ *is left and right differentiable in each component of* β *in some open neighborhood of the true parameter* β_0

2. *in an open neighborhood* B_0 *of* β_0,

$$\sup_{\beta\in B_0}\sum_{i<j}1\left\{\frac{\partial^{-}s\left(v_i,v_j;\beta\right)}{\partial\beta_\ell} \neq \frac{\partial^{+}s\left(v_i,v_j;\beta\right)}{\partial\beta_\ell}\right\} = O_p\left(1\right)$$

3. *in an open neighborhood of* β_0

$$\left|\frac{\partial^{-}s\left(v_i,v_j;\beta\right)}{\partial\beta_\ell} - \frac{\partial^{+}s\left(v_i,v_j;\beta\right)}{\partial\beta_\ell}\right| \leq h\left(v_i,v_j\right)$$

for some function h *with* $E\left[h\left(v_i,v_j\right)^{1+\delta}\right] < \infty$ *for some* δ.

and if K *is bounded, then*

$$G_n\left(\widehat{\gamma},\widehat{\beta}\right) = o_p\left(n^{-2+2/(1+\delta)}h_n^{-L}\right).$$

We next turn to some assumptions under which the conditions of Assumption 9 are satisfied. Recall that by definition of k_m and ℓ_m

$$\begin{aligned} k_t\left(z_i,a,b\right) &= E\left[t\left(v_i,v_j,b\right)\middle| z_i, w_j'\gamma_0 = a\right] \\ \ell_t\left(z_i,a,b\right) &= E\left[t\left(v_i,v_j,b\right)\middle| z_i, w_j'\gamma_0 = a\right] f_{w_j'\gamma_0}\left(a\right) \end{aligned}$$

In addition, define

$$t_1(z_i, z_j, \beta) = (w_i - w_j)\, t(v_i, v_j, \beta)$$

then k_{t_1} and ℓ_{t_1}, evaluated at $\beta = \beta_0$, become

$$k_{t_1}(z_i, a_2, \beta_0) = E\left[(w_i - w_j)\, t(v_i, v_j, \beta_0)\right| z_i, w_j'\gamma_0 = a_2\right]$$

$$\ell_{t_1}(z_i, a_2, \beta_0) = E\left[(w_i - w_j)\, t(v_i, v_j, \beta_0)\right| z_i, w_j'\gamma_0 = a_2\right] f_{w_i'\gamma_0}(a_2).$$

The following assumptions will be made.

Assumption 10.

1. ℓ_t is differentiable with respect to its third argument, and there is a function g with $E[g(z_i)] < \infty$, such that $\left|\ell_t^{(3)}(v_i, w_i'\gamma_0 - h\eta, \beta_0)\right| \le g(z_i)$.
2. ℓ_{t_1} is differentiable with respect to its second argument, and there is a function g with $E\left[g(z_i)^2\right] < \infty$, such that $\left|\ell_{t_1}^{(2)}(z_i, w_i'\gamma_0 - h\eta, \beta_0)\right| \le g(z_i)$. Furthermore,

   $$K(\eta)\, \ell_{t_1}^{(2)}\left(z_i, w_i'\gamma_0 - h\eta, \beta_0\right) \to 0 \quad \text{as} \quad \eta \to \pm\infty.$$

   Finally $E[(w_i - w_j)\, t(v_i, v_j, \beta)] < \infty$.
3. ℓ_t is differentiable with respect to its second argument, and there is a function g with $E[g(z_i)] < \infty$, such that $\left|\ell_t^{(2)}(v_i, w_i'\gamma_0 - h\eta, \beta_0)\right| \le g(z_i)$.

A number of results can be used to verify the convergence in Assumption 9.1. For example, Amemiya (1984), Theorem 4.1.4 gives conditions that can be used to verify that $\widetilde{p}_n'(z_i, \gamma^*, \beta^*)$ converges to $\lim \widetilde{p}_n'(z_i, \gamma_0, \beta_0)$ The following two lemmata give the expressions for the $\widetilde{p}_1'$ that appear in Assumption 9.1 and in Theorem 3.

Lemma 5 *Let*

$$p_0^\beta(\gamma_0, \beta_0) = E\left[\ell_t^{(3)}\left(v_i, w_i'\gamma_0, \beta_0\right)\right]$$

Then under assumptions 3 and 10(1)

$$p_{0n}^\beta(\gamma_0, \beta_0) \to p_0^\beta(\gamma_0, \beta_0)$$

Lemma 6 *Let*

$$p_0^\gamma(\gamma_0, \beta_0) = -E\left[\ell_{t_1}^{(2)}\left(z_i, w_i'\gamma_0, \beta_0\right)\right]$$

Then under assumptions 3 and 10(2)

$$p_{0n}^\gamma(\gamma_0, \beta_0) \to p_0^\gamma(\gamma_0, \beta_0)$$

Combined, the next two lemmata will give conditions under which Assumption 9.2 is satisfied.

Lemma 7 *Suppose that $p_{1n}(z_i;\gamma_0,\cdot)$ is continuously differentiable in a neighborhood, $N(\beta_0)$ of β_0, and that there is a function $h(z_i)$ with $E\left[\|h(z_i)\|^2\right]<\infty$, such that $\|p_{1n}(z_i;\gamma_0,b)\|\leq h(z_i)$ for all b in $N(\beta_0)$. Then*

$$\frac{1}{\sqrt{n}}\sum p_{1n}(z_i;\gamma_0,\beta_h)-p_{1n}(z_i;\gamma_0,\beta_0)=o_p(1)$$

Lemma 8 *If*

$$\begin{aligned}p_1(z_i;\gamma_0,\beta_0)&=E\left[t(v_i,v_j;\beta)|\,v_i,w_i,w_i'\gamma_0=w_j'\gamma_0\right]f_{w_j'\gamma_0}\left(w_i'\gamma_0\right)\\&=\ell_t\left(z_i,w_i'\gamma_0,\beta_0\right)\end{aligned}$$

then under assumptions 3 and 10.3

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_{1n}(z_i;\gamma_0,\beta_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}p_n(z_i;\gamma_0,\beta_0)=o_p(1).$$

3.3 Bias-reduction.

In this section, we will discuss conditions under which it is possible to get a Taylor series expansion of β_h arond β_0 as a function of h. To see why such a taylor series expansion is useful, suppose that one could write

$$\beta_h=\beta_0+\sum_{l=1}^{p}b_lh^l+o\left(n^{-1/2}\right)\tag{17}$$

Now let $c_1,c_2,...c_{p+1}$ be any sequence of positive numbers, and choose $a_1,a_2,...a_{p+1}$ such that

$$\begin{pmatrix}1&1&\cdots&1\\c_1&c_2&\cdots&c_{p+1}\\\vdots&\vdots&\ddots&\vdots\\c_1^p&c_2^p&\cdots&c_{p+1}^p\end{pmatrix}\begin{pmatrix}a_1\\a_2\\\vdots\\a_{p+1}\end{pmatrix}=\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$$

or

$$\begin{pmatrix}a_1\\a_2\\\vdots\\a_{p+1}\end{pmatrix}=\begin{pmatrix}1&1&\cdots&1\\c_1&c_2&\cdots&c_{p+1}\\\vdots&\vdots&\ddots&\vdots\\c_1^p&c_2^p&\cdots&c_{p+1}^p\end{pmatrix}^{-1}\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$$

Consider $p+1$ estimators based on bandwidths $c_1\cdot h$, $c_2\cdot h,\ldots c_{p+1}\cdot h$ with corresponding pseudo–true values $\beta_{c_1h},\beta_{c_2h},\ldots\beta_{c_{p+1}h}$, then the Taylor series expression in(17) implies that

$$\sum_{k=1}^{p+1}a_k\beta_{c_kh}=\beta_0+o\left(n^{-1/2}\right)\tag{18}$$

Indeed, we will see later that if $\int uK(u)\,du = 0$, then b_1 in (17) equals 0. In that case, the first order bias is 0 no matter what convex combination is chosen, and it therefore suffices to choose $a_1, a_2, ...a_p$ such that

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1^2 & c_2^2 & \cdots & c_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ c_1^p & c_2^p & \cdots & c_p^p \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

or

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1^2 & c_2^2 & \cdots & c_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ c_1^p & c_2^p & \cdots & c_p^p \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

in order to get

$$\sum_{k=1}^{p+1} a_k \beta_{c_k h} = \beta_0 + o\left(n^{-1/2}\right) \tag{19}$$

and hence

$$\sum_{k=1}^{p+1} a_k \widehat{\beta}_{c_k h} = \beta_0 + \frac{1}{n}\sum_{i=1}^{n} \psi_i + o_p\left(n^{-1/2}\right) \tag{20}$$

This line of argument assumes that (17) holds. To see why this is the case, note that with "enough differentiability" (defined below), β_h is defined by the relationship

$$p_{0n}(\gamma_0, \beta_h) = E\left[\int K(\eta)\,\ell_t\left(v_i, w_i'\gamma_0 - h\eta, \beta\right) d\eta\right] = 0$$

Since p_{0n} depends on n only through h, this could be written as

$$F(h, \beta_h) \overset{def}{=} E\left[\int K(\eta)\,\ell_t\left(v_i, w_i'\gamma_0 - h\eta, \beta_h\right) d\eta\right] = 0$$

By differentiation of implicit functions we then have

$$\begin{aligned} \frac{d\beta_h}{dh} &= -\left(\frac{\partial F(h, \beta_h)}{\partial \beta_h}\right)^{-1} \frac{\partial F(h, \beta_h)}{\partial h} \\ &= E\left[\int K(\eta)\,\ell_t^{(3)}\left(v_i, w_i'\gamma_0 - h\eta, \beta_h\right) d\eta\right]^{-1} E\left[\int \eta K(\eta)\,\ell_t^{(3)}\left(v_i, w_i'\gamma_0 - h\eta, \beta_h\right) d\eta\right] \end{aligned}$$

Evaluated at $h = 0$ (and hence $\beta_h = \beta_0$), we have

$$\begin{aligned} \left.\frac{d\beta_h}{dh}\right|_{h=0} &= E\left[\int K(\eta)\,\ell_t^{(3)}\left(v_i, w_i'\gamma_0, \beta_h\right) d\eta\right]^{-1} E\left[\int \eta K(\eta)\,\ell_t^{(3)}\left(v_i, w_i'\gamma_0, \beta_h\right) d\eta\right] \\ &= E\left[\int K(\eta)\,\ell_t^{(3)}\left(v_i, w_i'\gamma_0, \beta_h\right) d\eta\right]^{-1} E\left[\ell_t^{(3)}\left(v_i, w_i'\gamma_0, \beta_h\right)\right] \int \eta K(\eta)\, d\eta \\ &= 0 \end{aligned}$$

4 Asymptotic Properties of Partially Linear Logit Estimator.

In this section we discuss how the general results of section 3 might be used to derive the asymptotic properties of one of the estimators defined in section 2, the partially linear logit model. For this model, the terms in the objective function (6) are convex if K is positive. We can therefore use Theorem 1 to prove consistency. With the notation in section 3,, and with $\Lambda(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$, we have

$$s\left((y_i, x_i), (y_j, x_j); b\right) = -1\left\{y_i \neq y_j\right\}\left(y_i \ln \Lambda\left((x_i - x_j)'b\right) + y_j \ln \Lambda\left(-(x_i - x_j)'b\right)\right)$$

and $\widehat{\gamma} = I$. Also

$$t\left((y_i, x_i), (y_j, x_j); b\right) = 1\left\{y_i \neq y_j\right\}\left(y_i \Lambda\left(-(x_i - x_j)'b\right) - y_j \Lambda\left((x_i - x_j)'b\right)\right)(x_i - x_j)$$

Theorem 9 *Assume a random sample $\{(y_i, x_i, w_i)\}_{i=1}^{n}$ from (5) ε_i logistically distributed. The estimator defined by minimizing (6) where $h_n^{-1} = o\left(n^{-L/2}\right)$ and K satisfies Assumptions , 3, is consistent and asymptotically normal with*

$$\sqrt{n}\left(\widehat{\beta} - \beta_h\right) \widetilde{\longrightarrow} N\left(0, 4\Gamma^{-1} V \Gamma^{-1}\right)$$

with

$$V = V\left[r\left(y_i, x_i, w_i\right)\right]$$

where

$$r\left(y_i, x_i, w_i\right) = E\left[1\left\{y_i \neq y_j\right\}\left(y_i - \frac{\exp((x_i - x_j)'\beta)}{1 + \exp((x_i - x_j)'\beta)}\right)(x_i - x_j)\middle| y_i, x_i, w_i, w_j = w_i\right] f_w\left(w_i\right)$$

and

$$\Gamma = E\left[E\left[1\left\{y_i \neq y_j\right\} \frac{\exp\left((x_i - x_j)'\beta\right)}{\left(1 + \exp\left((x_i - x_j)'\beta\right)\right)^2}(x_i - x_j)(x_i - x_j)'\middle| y_i, x_i, w_i, w_j = w_i\right] f_{w_i}\left(w_i\right)\right]$$

provided that

1. $E\left[\|x_i\|^2\right] < \infty$.
2. *w_i is continuously distributed with a bounded density,f_{w_i}. Also $E\left[\|w_i\|^2\right] < \infty$.*
3. *$E\left[\|x_i\| \,|w_i = a\right] f_{w_i}(a)$ is a bounded function of a.*
4. *$(x_i - x_j)$ has full rank conditional on $z_i = z_j$.*

PROOF: First note that $|\log \Lambda(\eta)| \leq \log(2) + |\eta|$. Therefore $|s((y_i, x_i), (y_j, x_j); b)| \leq \log(2) + |(x_i - x_j)'b| \leq \log(2) + (\|x_i\| + \|x_j\|)\|b\|$. Assumptions 1.1, 1.2 and 1.3 are therefore satisfied. To verify 1.4, let $a_1 = (a_1^y, a_1^x)$ $|E[s(a_1, v_j, b)| w_j = a_2]| \leq E[|s(a_1, v_j, b)|| w_j = a_2] \leq E[\log(2) + (\|a_1^x\| + \|x_j\|)\|b\|| w_j = a_2]$, from which 1.4 follows. Assumption 6 then follows from consistency of the maximizer of the conditional likelihood for logit models with fixed effects. The remaining assumptions for consistency follow by assumption.

The asymptotic normality of the jackknifed estimator follows:

Corollary 10 *Under the assumptions of Theorem 9, the jackknifed estimator is consistent and asymptotically normal*

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \widetilde{\longrightarrow} N\left(0, 4\Gamma^{-1}V\Gamma^{-1}\right)$$

with V and Γ defined as in Theorem 9.

5 A Monte Carlo Illustration.

To get an idea about the small sample properties of the estimators of the partially linear models described above, we have performed a Monte Carlo investigation for the partially linear logit model for a particular design. The design is chosen to illustrate the method and is not meant to mimic a design that one would expect in a particular data set.. The model is

$$y_i = 1\{x_{1i}\beta_1 + x_{2i}\beta_2 + g(z_i) + \varepsilon_i\} \qquad i = 1, 2, \ldots, n, \tag{21}$$

where $(\beta_1, \beta_2) = (1, 1)$, $g(z) = z^2 - 2$, x_{2i} has a discrete distribution with $P(x_{2i} = -1) = P(x_{2i} = 1) = \frac{1}{2}$, $z_i \sim N(0, 1)$ and $x_{i1} = v_i + z_i^2$ where $v_i \sim N(0, 1)$. With this design, $P(y_i = 1) \simeq 0.44$. With the design used here, ignoring the non–linearity of $g(z)$ is expected to result in a bias in the estimators of both $\beta_1$and β_2, although we expect the bias to be bigger for β_1, because $g(z_i)$ is uncorrelated with x_{1i}.

For each replication of the model, we calculate a number of estimators. First, we calculate the logit maximum likelihood estimator using a constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors. This estimator would be asymptotically efficient if one knew $g(\cdot)$; comparing the estimators proposed here to that estimator will therefore give a measure of the cost of not knowing g (and using the estimators proposed here). Secondly, we calculate three estimators, $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_3$, based on (6) with K being the biweight (quartic)[5] kernel and $h_n = c * std(z) * n^{-1/5}$ where c takes the values 0.3, 0.9 and 2.7. These bandwidths are somewhat arbitrary. the middle one is motivated to the rule of thumb suggested by Silverman (1986, page 48) for estimation of densities (using normal kernel). That bandwidth is supposed to illustrate what happens if one uses a "reasonable" bandwidth. The

[5]Throughout, the kernel was normalized to have mean 0 and variance 1.

two other bandwidths are supposed to be "small" and "big". We also calculate four jackknifed estimators. The first, $\widehat{\beta}_{123}$ combines the three estimators according to (18). This ignores the fact that $\int uK(u)\,du = 0$, and we therefore also consider the three jack–knifed estimators based on combining two of the three estimators according to (20). These three estimators are denoted $\widehat{\beta}_{12}$, $\widehat{\beta}_{13}$, and $\widehat{\beta}_{23}$.

The results from 1000 replications with sample sized 100, 200, 400, 800, 1600 and 3200 are given in Table 1A through 1F. In addition to the true parameter values, each table also reports bias, standard deviation and root mean square error of the estimator, as well as the corresponding robust measures, the median bias, the median absolute deviation from the median and the median absolute error. Since all the estimators discussed here are likely to have fat tails (they are not even finite with probability 1), the discussion below will focus on these robust measures. The sample sizes are not chosen because we think that they are realistic given the small number of explanatory variables, but rather because we want to confirm that for large samples the estimator behaves as predicted by the asymptotic theory. As expected, the (correctly specified) maximum likelihood estimator that uses x_{1i}, x_{2i} and $g(z_i)$ as regressors outperforms the semiparametric estimators. However, the jackknifed estimators perform almost as well. For example, the median absolute error of the estimator based on jack-knifing using $\widehat{\beta}_2$ and $\widehat{\beta}_3$ is within 10% of the median absolute error of the maximum likelihood estimator (and often closer).

The patterns of the bias and the dispersion of the three estimators based on (6) are the expected — lower values of the bandwidth, h, give less bias but higher dispersion.

The proposed jack-knife procedure generally succeed in removing the bias of the proposed estimators. For example, focusing on the coefficient on x_{1i} (which has the bigger bias), the estimator that removes bias by comparing $\widehat{\beta}_2$ and $\widehat{\beta}_3$ has lower bias than either $\widehat{\beta}_2$ or $\widehat{\beta}_3$ for all sample sizes. Finally, for the largest sample sizes, there is almost no difference between the four bias reduced estimators, which corresponds to the predictions of the asymptotic theory.

Table 2 presents evidence about the effect of the bias term (as the bias reduction) on the performance of the test–statistics calculated on the basis of the estimators discussed here. For each sample and for each of the semiparametric estimators, we calculated 80, 90 and 95 percent confidence intervals. In order to do this, we estimated the (asymptotic variance) of the three non—bias reduced estimators by $4\widehat{\Gamma}_k^{-1}\widehat{V}_k\widehat{\Gamma}_k^{-1}$ where $k = 1, 2, 3$, denotes the estimator and $\widehat{\Gamma}_k$ is the sample variance of r_i^k defined by

$$r_i^k = \frac{1}{(n-1)\,h_n} \sum_{j \neq i} 1\{y_i \neq y_j\} \cdot K\left(\frac{w_i - w_j}{h_h}\right) \left(y_i - \frac{\exp\left((x_i - x_j)'\widehat{\beta}\right)}{1 + \exp\left((x_i - x_j)'\widehat{\beta}\right)} \right) (x_i - x_j)$$

and

$$\widehat{V}_k = \frac{2}{n\,(n-1)\,h_n} \sum_{i<j} 1\{y_i \neq y_j\} \cdot K\left(\frac{w_i - w_j}{h_h}\right) \frac{\exp\left((x_i - x_j)'\widehat{\beta}\right)}{\left(1 + \exp\left((x_i - x_j)'\widehat{\beta}\right)\right)^2} (x_i - x_j)(x_i - x_j)'$$

The estimated variance of any of the three estimators could be used to estimate the asymptotic variance of the jack–knifed estimators. However, in order to avoid arbitrarily choosing one variance estimator over an other, we estimated the joint asymptotic distribution of $\left(\widehat{\beta}_1', \widehat{\beta}_2', \widehat{\beta}_3'\right)'$ by $4\widehat{\Gamma}^{-1}\widehat{V}\widehat{\Gamma}^{-1}$ where $\widehat{\Gamma}$ is the sample variance of $\left(r_i^{1\prime}, r_i^{2\prime}, r_i^{3\prime}\right)'$ and

$$\widehat{\Gamma} = \begin{pmatrix} \widehat{\Gamma}_1 & 0 & 0 \\ 0 & \widehat{\Gamma}_2 & 0 \\ 0 & 0 & \widehat{\Gamma}_3 \end{pmatrix}$$

Table 2 gives the fraction of the replications for which these confidence intervals covered the true parameter.. Because the biases are more dramatic for β_1, we only present the results for that parameter. For all three sample sizes we see that the confidence interval that is based on the estimator, $\widehat{\beta}_1$, that is based on a very small bandwidth, has coverage probabilities that are close the 80, 90 and 95 percent, whereas the coverage probabilities are smaller for the two other non-bias reduced estimators, $\widehat{\beta}_2$ and $\widehat{\beta}_3$. While the discrepancies are not enormous (except for $\widehat{\beta}_3$ in large samples), it is interesting to note that all the bias corrected estimators perform better that both $\widehat{\beta}_2$ and $\widehat{\beta}_3$ for all sample sizes.

The sample sizes discussed so far are unrealitically large relative to the number of parameters, and there is little reason to think that the design mimics designs that one might encounter in applications. In order to investigate whether the good performance of the proposed estimators is an artifact of the very simple Monte Carlo design, we performed an additional experiment using the labor force participation data given in Berndt (1991, see also Mroz, 1987). Using a constant, log hourly earnings[6], number of kids under 6, number of kids between 6 and 18, age, age–squared, age–cubed, education, local unemployment rate, a dummy variable for whether the person lived in a large city, and other family income as explanatory variables, we estimated a logit for whether a woman worked in 1975. The sample size was 753 (of whom 428 worked). Using the original 753 vectors of explanatory variables, we generated 1000 data sets from this model. We then estimated the parameters using the correctly specified logit maximum likelihood estimator and the semiparametric estimator that treats the functional form for the effect of age as unknown. We calculate the bandwidths as for design 1, and the results are presented in Table 3.

6 Extensions and Future Research.

Ahn and Powell (1990) extended the model given in (3) and (4) by allowing the $z\gamma$ part to be an unknown function $p(z)$. Making the same extension in (10) and (11) leads to an estimator that

[6] This variable was imputed for the individuals who did not work.

minimizes a function of the form

$$Q_n(b) = \binom{n}{2}^{-1} \frac{1}{h_n} \sum_{\substack{i<j \\ d_i=d_j=1}} K\left(\frac{\hat{p}(z_i) - \hat{p}(z_j)}{h_n}\right) s(y_i, y_j, (x_i - x_j)b). \tag{22}$$

The estimator proposed by Ahn and Powell (1990) minimizes Q_n in (22), if there is no censoring and if quadratic loss ($\Xi(d) = d^2$) is used. In a companion paper in progress, Honoré and Powell (2002), we investigate the properties of the estimator defined by (22). That paper also presents the results of an empirical example.

7 References.

Amemiya, T. (1973): "Regression Analysis when the Dependent Variable is Truncated Normal," *Econometrica*, 41(6): 997–1016.

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

Barrodale, I. and, F. D. K. R. (1974): "Solution of an Overdetermined System of Equations in the l_1 Norm," *Communications ACM*, 17: 319-320.

Berndt, E. R. (1991): *The Practice of Econometrics: Classic and Contemporary*. Reading, Mass.: Addison-Wesley,

Cavanagh, C. and, R. Sherman. (1992): "Rank Estimators for Monotonic Regression Models," manuscript, Bellcore Economic Discussion Paper #84.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol. II, edited by Z. Griliches and M. Intriligator. Amsterdam: North Holland.

Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, edited by J. J.. Heckman and B. Singer. Cambridge University Press.

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35(2/3): 303-316.

Hausman, J., Hall, B. and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52(4), pp. 909–38.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60(3): 533–565.

Honoré, B. E., and J. L. Powell (1994): "Pairwise Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics*, 64(2): 241–278.

Mroz, T. A. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55(4), pp. 765–99.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. IV, edited by R. F. Engle and D. L. McFadden.

Powell, J. L. (1989): "Semiparametric Estimation of Censored Selection Models." Unpublished.

Powell, J. L., J. Stock and T. Stoker (1989): "Semiparametric Estimation of Weighted Average Derivatives", *Econometrica*, 57, pp. 1403-1430.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1986): *Numerical Recipes: The Art of Scientific Computing* , Cambridge: Cambridge University Press.

Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmarks Pædagogiske Institut, Copenhagen.

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics*, New York: John Wiley and Sons.

Sherman R. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61(1): 123–137.

Sherman, R (1994a): "U–Processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10, pp. 372–395.

Sherman, R. (1994b): "Maximal Inequalities for Degenerate U–Processes with Applications to Optimization Estimators," *Annals of Statistics*, 22(1), pp. 439–459.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

8 Appendix: The Most Dull Derivations.

Proof of Lemma 5:

$$\begin{aligned}
p_{0n}(\gamma_0,\beta) &= E\left[\frac{1}{h^L}K\left(\frac{w_i'\gamma_0-w_j'\gamma_0}{h}\right)t(v_i,v_j,\beta)\right]\\
&= E\left[\frac{1}{h^L}K\left(\frac{w_i'\gamma_0-w_j'\gamma_0}{h}\right)E\left[t(v_i,v_j,\beta)|(v_i,w_i),w_j'\gamma_0\right]\right]\\
&= E\left[\frac{1}{h^L}K\left(\frac{w_i'\gamma_0-w_j'\gamma_0}{h}\right)k_t\left(v_i,w_j'\gamma_0,\beta\right)\right]\\
&= E\left[\int\frac{1}{h^L}K\left(\frac{w_i'\gamma_0-\omega}{h}\right)k_t(v_i,\omega,\beta)f_{w_i'\gamma_0}(\omega)\,d\omega\right]\\
&= E\left[\int K(\eta)\,\ell_t\left(v_i,w_i'\gamma_0-h\eta,\beta\right)d\eta\right]
\end{aligned}$$

By assumption 10.1, we can differentiate under the expectation and integral (see e.g., Cramér (1946, page 68)):

$$\begin{aligned}
p_{0n}^{\beta}(\gamma_0,\beta_0) &= E\left[\int K(\eta)\,\ell_t^{(3)}\left(v_i,w_i'\gamma_0-h\eta,\beta_0\right)d\eta\right]\\
&\rightarrow E\left[\ell_t^{(3)}\left(v_i,w_i'\gamma_0,\beta_0\right)\right]
\end{aligned}$$

where the limit follows from dominated convergence.■

Proof of Lemma 6: Recall that

$$p_{0n}(\gamma,\beta)=E\left[\frac{1}{h^L}K\left(\frac{w_i'\gamma-w_j'\gamma}{h}\right)t(v_i,v_j,\beta)\right]$$

By assumptions 3 and 10.2, we can differentiate under the expectation:

$$p_{0n}^{\gamma}(\gamma,\beta)=E\left[\frac{1}{h^L}K'\left(\frac{w_i'\gamma-w_j'\gamma}{h}\right)\frac{w_i-w_j}{h}t(v_i,v_j,\beta)\right].$$

Evaluation this at(γ_0,β_0), we get

$$\begin{aligned}
p_{0n}^{\gamma}(\gamma_0,\beta_0) &= E\left[\frac{1}{h^L}K'\left(\frac{w_i'\gamma_0-w_j'\gamma_0}{h}\right)\frac{w_i-w_j}{h}t(v_i,v_j,\beta_0)\right]\\
&= E\left[E\left[E\left[\frac{1}{h^L}K'\left(\frac{w_i'\gamma_0-w_j'\gamma_0}{h}\right)\frac{w_i-w_j}{h}t(v_i,v_j,\beta_0)\middle|v_i,w_i,w_j'\gamma_0\right]\middle|v_i,w_i\right]\right]\\
&= E\left[E\left[\frac{1}{h^{L+1}}K'\left(\frac{w_i'\gamma_0-w_j'\gamma_0}{h}\right)E\left[(w_i-w_j)\,t(v_i,v_j,\beta_0)|\,v_i,w_i,w_j'\gamma_0\right]\middle|v_i,w_i\right]\right]
\end{aligned}$$

with the definitions of k_{t_1} and ℓ_{t_1}, and using integration by parts, we have

$$p_{0n}^{\gamma}(\gamma_0,\beta_0) = E\left[\int\frac{1}{h^{L+1}}K'\left(\frac{w_i'\gamma_0-\omega}{h}\right)k_{t_1}(z_i,\omega,\beta_0)f_{w_i'\gamma_0}(\omega)\,d\omega\right]$$

$$
\begin{aligned}
&= E\left[\int \frac{1}{h^{L+1}} K'\left(\frac{w_i'\gamma_0 - \omega}{h}\right) \ell_{t_1}\left(z_i, \omega, \beta_0\right) d\omega\right] \\
&= E\left[\int \frac{1}{h} K'\left(\eta\right) \ell_{t_1}\left(z_i, w_i'\gamma_0 - h\eta, \beta_0\right) d\eta\right] \\
&= -E\left[\int K\left(\eta\right) \ell_{t_1}^{(2)}\left(z_i, w_i'\gamma_0 - h\eta, \beta_0\right) d\eta\right] \\
&\to -E\left[\ell_{t_1}^{(2)}\left(z_i, w_i'\gamma_0, \beta_0\right)\right]
\end{aligned}
$$

where the limit follows from dominated convergence.■

PROOF OF LEMMA 8: Write

$$p_{1n}\left(z_i; \gamma_0, \beta_0\right) = r_n\left(z_i\right) - E\left[r_n\left(z_i\right)\right]$$

and

$$p_1\left(z_i\right) = \ell_t\left(z_i, w_i'\gamma_0, \beta_0\right) - E\left[\ell_t\left(z_i, w_i'\gamma_0, \beta_0\right)\right]$$

where

$$
\begin{aligned}
r_n\left(z_i\right) &= E\left[p_n\left(z_i, z_j; \gamma_0, \beta_0\right) \middle| z_i\right] \\
&= \int \frac{1}{h^L} K\left(\frac{w_i'\gamma_0 - \omega}{h}\right) \ell_t\left(v_i, \omega, \beta_0\right) d\omega \\
&= \int K\left(\eta\right) \ell_t\left(v_i, w_i'\gamma_0 - h\eta, \beta_0\right) d\eta
\end{aligned}
$$

and $E\left[\ell_t\left(z_i, w_i'\gamma_0, \beta_0\right)\right] = 0$. With $\delta_n\left(z_i\right) = r_n\left(z_i\right) - \ell_t\left(z_i, w_i'\gamma_0, \beta_0\right)$, we then have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_{1n}\left(z_i; \gamma_0, \beta_0\right) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} p_n\left(z_i; \gamma_0, \beta_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_n\left(z_i\right) - E\left[\delta_n\left(z_i\right)\right]$$

The right hand side has mean 0 and variance

$$
\begin{aligned}
V\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_n\left(z_i\right) - E\left[\delta_n\left(z_i\right)\right]\right] &= V\left[\delta_n\left(z_i\right) - E\left[\delta_n\left(z_i\right)\right]\right] \\
&\leq E\left[\delta_n\left(z_i\right)^2\right] \\
&= E\left[\left\{r_n\left(z_i\right) - \ell_t\left(z_i, w_i'\gamma_0, \beta_0\right)\right\}^2\right] \\
&= E\left[\left\{\int K\left(\eta\right) \ell_t\left(v_i, w_i'\gamma_0 - h\eta, \beta_0\right) d\eta - \ell_t\left(z_i, w_i'\gamma_0, \beta_0\right)\right\}^2\right] \\
&= E\left[\left\{\int K\left(\eta\right) \left(\ell_t\left(v_i, w_i'\gamma_0 - h\eta, \beta_0\right) - \ell_t\left(z_i, w_i'\gamma_0, \beta_0\right)\right) d\eta\right\}^2\right] \\
&\leq E\left[g\left(z_i\right)^2 h^2 \left\{\int \left|K\left(\eta\right)\right| \left\|\eta\right\| d\eta\right\}^2\right] \\
&= O\left(h^2\right) \to 0.
\end{aligned}
$$

where g is the function in Assumption 10.3.■

PROOF OF LEMMA 7: For the duration of this proof, let p_{1n} denote one of the elements of p_{1n}. By definition of p_{1n} and by random sampling, the mean of the left hand side is 0, while the variance is

$$E\left[\left(p_{1n}\left(z_i;\gamma_0,\beta_h\right)-p_{1n}\left(z_i;\gamma_0,\beta_0\right)\right)^2\right]\leq E\left[\left\|p_{1n}^{\beta}\left(z_i;\gamma_0,\beta_i^*\right)\right\|^2\right]\left\|\beta_h-\beta_0\right\|^2$$

where β_i^* is between β_h and β_0, but may depend on z_i (hence the subscript i). The result now follows from $\beta_h\to\beta_0$.■

TABLE 1A: Design 1. Sample size=100.

	value	bias	st.dev.	RMSE	m.bias	MAD	MAE
Logit MLE using constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors							
β_1	1.000	0.138	0.406	0.428	0.083	0.253	0.245
β_2	1.000	0.107	0.376	0.390	0.063	0.222	0.213
The estimator based on (6) with $h_n = 0.3 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.175	0.450	0.482	0.103	0.279	0.271
β_2	1.000	0.117	0.500	0.513	0.047	0.241	0.246
The estimator based on (6) with $h_n = 0.9 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.249	0.402	0.473	0.198	0.248	0.251
β_2	1.000	0.065	0.372	0.377	0.026	0.228	0.223
The estimator based on (6) with $h_n = 2.7 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.427	0.370	0.565	0.373	0.233	0.373
β_2	1.000	0.010	0.333	0.333	-0.021	0.202	0.206
Jack-knife using $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_2$							
β_1	1.000	0.136	0.504	0.521	0.057	0.309	0.307
β_2	1.000	0.150	0.621	0.638	0.071	0.265	0.271
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_2$							
β_1	1.000	0.166	0.459	0.488	0.091	0.283	0.275
β_2	1.000	0.124	0.521	0.535	0.049	0.247	0.251
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_3$							
β_1	1.000	0.172	0.451	0.483	0.099	0.281	0.272
β_2	1.000	0.119	0.503	0.516	0.049	0.243	0.246
Jack-knife using $\widehat{\beta}_2$ and $\widehat{\beta}_3$							
β_1	1.000	0.227	0.408	0.466	0.172	0.255	0.247
β_2	1.000	0.072	0.379	0.385	0.033	0.227	0.226

TABLE 1B: Design 1. Sample size=200.

	value	bias	st.dev.	RMSE	m.bias	MAD	MAE
Logit MLE using constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors							
β_1	1.000	0.038	0.253	0.256	0.016	0.161	0.157
β_2	1.000	0.034	0.238	0.241	0.012	0.157	0.158
The estimator based on (6) with $h_n = 0.3 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.050	0.276	0.280	0.013	0.177	0.174
β_2	1.000	0.031	0.264	0.265	0.019	0.176	0.174
The estimator based on (6) with $h_n = 0.9 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.132	0.257	0.288	0.098	0.163	0.166
β_2	1.000	0.008	0.243	0.243	-0.001	0.153	0.154
The estimator based on (6) with $h_n = 2.7 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.324	0.235	0.400	0.290	0.146	0.290
β_2	1.000	-0.041	0.227	0.230	-0.050	0.145	0.153
Jack-knife using $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_2$							
β_1	1.000	0.005	0.296	0.296	-0.021	0.191	0.193
β_2	1.000	0.043	0.285	0.288	0.031	0.186	0.189
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_2$							
β_1	1.000	0.039	0.279	0.282	0.006	0.179	0.180
β_2	1.000	0.033	0.268	0.269	0.024	0.178	0.176
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_3$							
β_1	1.000	0.046	0.276	0.280	0.010	0.177	0.177
β_2	1.000	0.031	0.265	0.266	0.020	0.177	0.175
Jack-knife using $\widehat{\beta}_2$ and $\widehat{\beta}_3$							
β_1	1.000	0.108	0.260	0.282	0.074	0.168	0.161
β_2	1.000	0.014	0.246	0.246	0.004	0.159	0.157

TABLE 1C: Design 1. Sample size=400.

	value	bias	st.dev.	RMSE	m.bias	MAD	MAE
Logit MLE using constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors							
β_1	1.000	0.018	0.164	0.165	0.012	0.115	0.110
β_2	1.000	0.020	0.158	0.159	0.015	0.101	0.102
The estimator based on (6) with $h_n = 0.3 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.030	0.181	0.184	0.024	0.123	0.122
β_2	1.000	0.017	0.170	0.170	0.005	0.109	0.110
The estimator based on (6) with $h_n = 0.9 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.090	0.172	0.194	0.078	0.115	0.126
β_2	1.000	0.000	0.162	0.162	-0.008	0.106	0.107
The estimator based on (6) with $h_n = 2.7 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.279	0.156	0.319	0.265	0.104	0.265
β_2	1.000	-0.041	0.151	0.157	-0.043	0.097	0.105
Jack-knife using $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_2$							
β_1	1.000	0.001	0.190	0.189	-0.006	0.129	0.130
β_2	1.000	0.026	0.176	0.178	0.017	0.112	0.117
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_2$							
β_1	1.000	0.023	0.183	0.184	0.016	0.125	0.124
β_2	1.000	0.019	0.171	0.172	0.006	0.110	0.111
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_3$							
β_1	1.000	0.027	0.182	0.184	0.022	0.123	0.123
β_2	1.000	0.018	0.170	0.171	0.006	0.109	0.110
Jack-knife using $\widehat{\beta}_2$ and $\widehat{\beta}_3$							
β_1	1.000	0.066	0.175	0.187	0.058	0.118	0.120
β_2	1.000	0.006	0.164	0.164	-0.002	0.106	0.107

TABLE 1D: Design 1. Sample size=800.

	value	bias	st.dev.	RMSE	m.bias	MAD	MAE
Logit MLE using constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors							
β_1	1.000	0.010	0.123	0.124	0.004	0.082	0.084
β_2	1.000	0.014	0.114	0.115	0.012	0.075	0.075
The estimator based on (6) with $h_n = 0.3 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.019	0.131	0.132	0.012	0.088	0.088
β_2	1.000	0.012	0.123	0.124	0.006	0.082	0.082
The estimator based on (6) with $h_n = 0.9 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.067	0.128	0.144	0.063	0.089	0.092
β_2	1.000	0.001	0.120	0.120	-0.002	0.079	0.078
The estimator based on (6) with $h_n = 2.7 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.246	0.118	0.273	0.241	0.083	0.241
β_2	1.000	-0.039	0.113	0.120	-0.041	0.076	0.082
Jack-knife using $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_2$							
β_1	1.000	-0.002	0.134	0.134	-0.008	0.088	0.088
β_2	1.000	0.018	0.126	0.127	0.012	0.085	0.085
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_2$							
β_1	1.000	0.014	0.132	0.132	0.007	0.088	0.088
β_2	1.000	0.014	0.124	0.124	0.009	0.082	0.082
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_3$							
β_1	1.000	0.017	0.131	0.132	0.009	0.088	0.088
β_2	1.000	0.013	0.123	0.124	0.007	0.083	0.082
Jack-knife using $\widehat{\beta}_2$ and $\widehat{\beta}_3$							
β_1	1.000	0.045	0.129	0.137	0.040	0.089	0.090
β_2	1.000	0.006	0.121	0.121	0.003	0.080	0.080

TABLE 1E: Design 1. Sample size=1600.

	value	bias	st.dev.	RMSE	m.bias	MAD	MAE
Logit MLE using constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors							
β_1	1.000	0.004	0.078	0.079	0.005	0.053	0.053
β_2	1.000	0.006	0.079	0.079	0.009	0.052	0.053
The estimator based on (6) with $h_n = 0.3 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.010	0.085	0.085	0.009	0.056	0.058
β_2	1.000	0.005	0.083	0.084	0.007	0.054	0.054
The estimator based on (6) with $h_n = 0.9 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.048	0.082	0.095	0.047	0.055	0.067
β_2	1.000	-0.004	0.081	0.081	-0.003	0.053	0.053
The estimator based on (6) with $h_n = 2.7 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.213	0.076	0.226	0.214	0.051	0.214
β_2	1.000	-0.040	0.077	0.087	-0.038	0.051	0.057
Jack-knife using $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_2$							
β_1	1.000	-0.005	0.086	0.086	-0.006	0.057	0.057
β_2	1.000	0.009	0.085	0.085	0.010	0.055	0.056
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_2$							
β_1	1.000	0.006	0.085	0.085	0.005	0.056	0.057
β_2	1.000	0.006	0.084	0.084	0.008	0.053	0.054
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_3$							
β_1	1.000	0.008	0.085	0.085	0.007	0.056	0.057
β_2	1.000	0.006	0.084	0.084	0.007	0.054	0.054
Jack-knife using $\widehat{\beta}_2$ and $\widehat{\beta}_3$							
β_1	1.000	0.027	0.084	0.088	0.027	0.056	0.060
β_2	1.000	0.001	0.082	0.082	0.001	0.053	0.053

TABLE 1F: Design 1. Sample size=3200.

	value	bias	st.dev.	RMSE	m.bias	MAD	MAE
Logit MLE using constant, x_{1i}, x_{2i} and $g(z_i)$ as regressors							
β_1	1.000	0.002	0.059	0.059	0.003	0.040	0.040
β_2	1.000	0.002	0.055	0.055	0.003	0.037	0.038
The estimator based on (6) with $h_n = 0.3 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.006	0.063	0.063	0.006	0.040	0.041
β_2	1.000	-0.000	0.059	0.059	-0.004	0.040	0.041
The estimator based on (6) with $h_n = 0.9 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.035	0.062	0.071	0.035	0.040	0.048
β_2	1.000	-0.007	0.059	0.059	-0.011	0.040	0.040
The estimator based on (6) with $h_n = 2.7 \cdot std(z) \cdot n^{-1/5}$							
β_1	1.000	0.182	0.058	0.191	0.183	0.037	0.183
β_2	1.000	-0.039	0.056	0.068	-0.042	0.039	0.051
Jack-knife using $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_2$							
β_1	1.000	-0.004	0.063	0.063	-0.005	0.041	0.041
β_2	1.000	0.002	0.060	0.060	-0.000	0.040	0.041
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_2$							
β_1	1.000	0.003	0.063	0.063	0.002	0.040	0.040
β_2	1.000	0.001	0.060	0.060	-0.002	0.040	0.041
Jack-knife using $\widehat{\beta}_1$ and $\widehat{\beta}_3$							
β_1	1.000	0.004	0.063	0.063	0.004	0.040	0.041
β_2	1.000	0.000	0.060	0.060	-0.003	0.040	0.041
Jack-knife using $\widehat{\beta}_2$ and $\widehat{\beta}_3$							
β_1	1.000	0.017	0.062	0.064	0.017	0.041	0.041
β_2	1.000	-0.003	0.059	0.059	-0.006	0.040	0.041

TABLE 2: Design 1. Coverage Probabilities.

95% Confidence Interval

	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$	$n = 3200$
$\widehat{\beta}_1$	0.948	0.938	0.962	0.933	0.958	0.940
$\widehat{\beta}_2$	0.905	0.927	0.927	0.917	0.931	0.912
$\widehat{\beta}_3$	0.801	0.726	0.573	0.391	0.202	0.080
$\widehat{\beta}_{123}$	0.968	0.943	0.957	0.931	0.960	0.940
$\widehat{\beta}_{12}$	0.950	0.937	0.965	0.934	0.961	0.940
$\widehat{\beta}_{13}$	0.948	0.938	0.963	0.933	0.961	0.941
$\widehat{\beta}_{23}$	0.910	0.932	0.945	0.929	0.948	0.940

90% Confidence Interval

	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$	$n = 3200$
$\widehat{\beta}_1$	0.894	0.888	0.907	0.885	0.915	0.898
$\widehat{\beta}_2$	0.849	0.850	0.864	0.841	0.866	0.856
$\widehat{\beta}_3$	0.684	0.624	0.465	0.290	0.123	0.044
$\widehat{\beta}_{123}$	0.925	0.897	0.918	0.897	0.909	0.892
$\widehat{\beta}_{12}$	0.903	0.889	0.908	0.885	0.914	0.899
$\widehat{\beta}_{13}$	0.896	0.887	0.906	0.886	0.913	0.901
$\widehat{\beta}_{23}$	0.860	0.866	0.884	0.866	0.903	0.885

80% Confidence Interval

	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$	$n = 3200$
$\widehat{\beta}_1$	0.782	0.786	0.785	0.786	0.817	0.798
$\widehat{\beta}_2$	0.732	0.744	0.751	0.712	0.758	0.740
$\widehat{\beta}_3$	0.526	0.453	0.303	0.174	0.059	0.023
$\widehat{\beta}_{123}$	0.827	0.800	0.814	0.803	0.810	0.808
$\widehat{\beta}_{12}$	0.791	0.789	0.801	0.794	0.810	0.799
$\widehat{\beta}_{13}$	0.784	0.787	0.789	0.789	0.814	0.796
$\widehat{\beta}_{23}$	0.740	0.757	0.773	0.746	0.805	0.794

TABLE 3: Monte Carlo Based on Real Data.

Median Absolute error relative to MLE

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_{123}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{13}$	$\widehat{\beta}_{23}$
Wage rate	1.0229	1.0161	0.9796	1.0531	1.0228	1.0260	1.0206
Kids less than 6	1.0342	0.9860	1.1623	1.0443	1.0273	1.0352	1.0065
Kids between 6 and 18	1.0372	0.9928	1.7117	1.0560	1.0358	1.0416	1.0346
Education	1.0648	1.0303	1.0156	1.0711	1.0625	1.0640	1.0354
Local Unemployment	1.0511	1.0439	1.0386	1.0876	1.0643	1.0518	1.0467
City	1.0454	1.0016	0.9609	1.0371	1.0336	1.0439	0.9909
Other Income	1.0423	1.0181	1.0243	1.0207	1.0395	1.0432	1.0149

Coverage Probability for 80% Confidence Interval

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_{123}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{13}$	$\widehat{\beta}_{23}$
Wage rate	0.786	0.778	0.785	0.792	0.786	0.786	0.777
Kids less than 6	0.808	0.799	0.715	0.817	0.807	0.809	0.802
Kids between 6 and 18	0.795	0.780	0.480	0.801	0.802	0.798	0.790
Education	0.780	0.778	0.770	0.788	0.781	0.779	0.775
Local Unemployment	0.802	0.793	0.800	0.807	0.806	0.802	0.794
City	0.825	0.816	0.817	0.822	0.822	0.822	0.815
Other Income	0.790	0.781	0.778	0.797	0.789	0.790	0.784

Coverage Probability for 90% Confidence Interval

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_{123}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{13}$	$\widehat{\beta}_{23}$
Wage rate	0.889	0.886	0.883	0.893	0.889	0.889	0.887
Kids less than 6	0.900	0.904	0.832	0.909	0.908	0.901	0.900
Kids between 6 and 18	0.885	0.877	0.624	0.897	0.887	0.885	0.885
Education	0.889	0.890	0.883	0.890	0.891	0.890	0.890
Local Unemployment	0.894	0.895	0.889	0.900	0.896	0.894	0.893
City	0.912	0.912	0.909	0.918	0.916	0.913	0.911
Other Income	0.899	0.905	0.899	0.903	0.898	0.898	0.901

Coverage Probability for 95% Confidence Interval

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_{123}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{13}$	$\widehat{\beta}_{23}$
Wage rate	0.947	0.946	0.943	0.946	0.948	0.947	0.946
Kids less than 6	0.960	0.956	0.902	0.959	0.959	0.960	0.959
Kids between 6 and 18	0.942	0.934	0.740	0.950	0.944	0.943	0.944
Education	0.944	0.936	0.935	0.945	0.943	0.944	0.942
Local Unemployment	0.941	0.938	0.943	0.943	0.942	0.941	0.937
City	0.963	0.962	0.954	0.963	0.964	0.964	0.964
Other Income	0.953	0.955	0.958	0.959	0.954	0.952	0.954