

Measurement Error Models with Auxiliary Data

XIAOHONG CHEN

New York University

HAN HONG

Duke University

and

ELIE TAMER

Northwestern University

First version received April 2002; final version accepted April 2004 (Eds.)

We study the problem of parameter inference in (possibly non-linear and non-smooth) econometric models when the data are measured with error. We allow for *arbitrary* correlation between the true variables and the measurement errors. To solve the identification problem, we require the existence of an *auxiliary* data-set that contains information about the conditional distribution of the true variables given the mismeasured variables. Our main assumption requires that the conditional distribution of the true variables given the mismeasured variables is the same in the primary and auxiliary data. Our methods allow the auxiliary data to be a validation sample, where the primary and validation data are from the same distribution, and more importantly, a stratified sample where the auxiliary data-set is not from the same distribution as the primary data. We also show how to combine the two data-sets to obtain a more efficient estimator of the parameter of interest. We establish the large sample properties of the sieve based estimators under verifiable conditions. In particular, we allow for the mismeasured variables to have unbounded supports without employing the tedious trimming scheme typically used in kernel based methods. We illustrate our methods by estimating a returns to schooling censored quantile regression using the CPS/SSR 1978 exact match files where the dependent variable is measured with error of arbitrary kind.

1. INTRODUCTION

This paper is motivated by concerns in the applied economics literature about the validity of the classical measurement error assumption where the observed variable X is equal to the latent variable of interest X^* plus the measurement error ϵ that is assumed to be uncorrelated with (or even independent of) X^* . For example, in economic data, it is often the case that data-sets rely on individual respondents to provide information. If respondents' reports are off randomly by a small amount, then the classical errors in variables model is attractive. However, it may be hard to tell whether or not respondents are making up their answers, and more crucially whether the measurement error is correlated with some of the variables. For example, Bound and Krueger (1991) compare reported income in the 1977 wave of the CPS to the social security match file for respondents who provide a valid social security number. They find that measurement error is correlated with the true variable, especially in the tails of the income distribution. Bound, Brown, Duncan and Rodgers (1994) used a validation study on one large firm from the PSID and found evidence against the classical measurement error model and especially against the independence assumption between the measurement error and the true variable. Finally, a good

account of the impact of non-classical measurement error on inference in econometric models is given in Bound, Brown and Mathiowetz (2002). In this paper, we show that in the presence of *auxiliary data* obeying a key assumption we can consistently estimate the parameters of interest while allowing for arbitrary correlation between the measurement errors and other variables. We also allow for non-linearity and non-smoothness in the moment conditions and show how one can obtain more efficient estimators using auxiliary data. The auxiliary data may come from a validation study applied to a subset of the primary sample. For example, the auxiliary data we use in the empirical illustration section come from validating the social security income for a subset of the primary sample respondents who reported their social security numbers.

We study the problem of inference about a parameter β_o defined in terms of unconditional moment restrictions when the data are measured with error. For example, consider the moment condition

$$E[m(X^*, \beta_o)] = 0 \quad (1)$$

where $m(\cdot)$ is a known function that is possibly non-linear and non-smooth in β and/or in X^* , and X^* is unobserved. Instead we observe a *primary* data-set of mismeasured variables $\{X_p\}$. The literature often assumes the linear errors in variables model of the classical kind where the observed data are a convolution of the true data and a mean zero random variable that represents the measurement error. The crucial assumption that is maintained in the classical measurement error model is the independence (or uncorrelatedness) between the measurement error and the true (but unobserved) variable. We relax this assumption by allowing *arbitrary correlation* between the measurement error and the true data. To solve the identification problem, we require the presence of an *auxiliary data-set* which we use to estimate β_o . In this paper we use auxiliary data to mean a random sample of mismeasured *and* true variables which may or may not come from the same distribution as the primary data. In the case where the additional data are randomly sampled from the population (X^*, X_p) , the auxiliary data-set becomes a validation data-set. More importantly, the auxiliary data can also be a random sample of variables $\{(X_v^*, X_v)\}$ where the marginal distribution of X_v^* is different from the marginal distribution of the variable X^* in the primary sample. We will refer to this as a case of stratified sampling where a certain segment of the population is oversampled. For example, one might want to oversample higher income individuals if more misreporting is suspected in this subpopulation.

The class of models that we consider is broad and includes non-linear and/or non-smooth moment based models (like Euler equations and censored quantile regressions). For example, in quantile regressions, we allow for measurement error in the dependent variable; this measurement error can be correlated with the regressors. When one suspects the presence of non-classical measurement error, our results can be viewed as providing support for collecting auxiliary data that can be used to obtain consistent estimates. We use a semiparametric sieve based estimator of β_o that combines the information in the auxiliary data and the primary data. Our estimator is formulated using the identifying assumption that the conditional distribution of X_v^* given X_v in the auxiliary sample $\{(X_v^*, X_v)\}$ is the same as that of X^* given X_p in the primary sample $\{(X_p^*, X_p)\}$ where X_p^* is the unobserved counterpart of X^* . The intuition of the estimator is simple. The auxiliary data allows us to obtain the conditional relationship between the true but unobserved variables and the observed and mismeasured variables. This relationship is then used with the primary data to estimate the parameters of interest. Hence, with an auxiliary data-set, one can learn about the relationship between the true variables and their mismeasured counterpart and use this relationship to back out the parameter of interest using the primary data-set. Our estimator is consistent and asymptotically normal under a set of regularity conditions. In the presence of a validation sample, we discuss the efficiency gains achieved by optimally combining the moment conditions implied in the primary data and the validation data.

This paper also makes a theoretical contribution to the asymptotics in a class of semiparametric models. In particular, our sieve based method avoids making the tedious trimming arguments that are typically made with kernel based methods. Using kernel methods to approximate conditional expectations in our setting would require strong tail assumptions on the distribution of the mismeasured variables to guard against small values for the density (similar to the ones obtained in Lavergne and Vuong (1996)). These assumptions are not theoretically attractive and can be avoided when using trimming based on sieve methods (see Section 3 for more details).

Finally, we present an empirical illustration that uses a data-set from the Current Population Survey/social security match data to explore the returns to schooling under general measurement error in earnings. Using this data-set, Bound and Krueger (1991) and Bollinger (1998) documented the presence of non-classical measurement error in earnings. In the 1978 March rotation of the CPS, respondents were asked for their social security number in addition to other questions including earnings. For those who provided a social security number, their income was matched against their employer reported income with the social security administration, which we treat as our auxiliary sample. We implemented a censored quantile regression of log earnings on education and other regressors in which earnings are measured with error.

There is a large literature in econometrics and statistics concerned with inference in linear and non-linear models with measurement errors. A review of the commonly used techniques in statistics can be found in Fuller (1987) and Carroll, Ruppert and Stefanski (1995). Econometric work on the classical independent additive measurement error model dates back to Frisch (1934), who derived bounds on the slope and the constant term in linear regression with measurement error. The method of instrumental variables (IVs) is popular for obtaining consistent estimators of the parameters of interest in linear models with classical independent additive measurement error. In non-linear regression models, Hausman, Ichimura, Newey and Powell (1991) and Hausman, Newey and Powell (1995) generalized this IV method to polynomial functions in the presence of double measurements on the mismeasured variables. Li (2002) and Schennach (2004) presented methods for non-linear regression models with classical measurement error and double measurements. See also Hsiao and Wang (1995) and Newey (2001). Taupin (2001) and Hong and Tamer (2003) use distributional assumptions on the measurement error to obtain a simple estimator in non-linear models when no auxiliary data are present. Chesher (1991) presented useful approximation methods to the true distribution and parameters of interest. Most of these papers impose the classical errors in variables assumption. On the other hand, Horowitz and Manski (1995) used a different model of measurement error, where they assume that the observed sample is contaminated or corrupted (for more on this issue, see Molinari, 2003). Other papers allowing for non-classical measurement errors are the ones assuming the presence of true validation data; see, *e.g.* Carroll and Wand (1991), Sepanski and Carroll (1993) and Lee and Sepanski (1995). Carroll and Wand (1991) use a semiparametric maximum likelihood estimator in a logistic regression model with covariate measurement error. Sepanski and Carroll (1993) use a quasi-likelihood approach to estimate non-linear regression models. In addition to the existence of a validation sample, both papers assumed that the conditional distribution of the response given the mismeasured variable and the true variable is the same as the conditional distribution given the true variable. Lee and Sepanski (1995) proposed an innovative estimator for non-linear regression problems in the presence of validation data. Their method uses a least squares projection onto a fixed finite dimensional collection of functions as a replacement for the conditional expectation of the non-linear function. This avoids the use of a semiparametric estimator for the conditional expectation, but is less efficient.

Our paper differs from the current literature in several important ways. We allow for a general measurement error model that allows arbitrary correlation between the observed and the

true variables while relaxing the requirement of validation data by allowing for stratification. Moreover, our framework handles non-linear models with possibly non-smooth functions. The class of models that Carroll and Wand (1991), Sepanski and Carroll (1993) and Lee and Sepanski (1995) consider is smaller than the one we cover in this paper. They consider non-linear least squares problems where the measurement error is only in the regressors. Our general class of moment based models nests theirs but also contains for example the class of best predictor problems with measurement error in both the regressors and response. This, for example, allows one to study (possibly censored) quantile regression with measurement error in the dependent variable as well as in the regressors, something that the above-mentioned papers cannot handle. In fact, our set-up nests theirs in a similar way that generalized method of moments nests least squares. Moreover, the results in these papers generally fail if the auxiliary data are obtained by stratified sampling.

The remainder of the paper is organized as follows. Section 2 formally introduces the model and our estimator. Section 3 provides the assumptions and the large sample distributional results. It also discusses the efficiency gain of combining the validation data and the primary data, and provides a simple consistent estimator of the asymptotic variance. Section 4 presents the empirical illustration. Section 5 concludes. All technical proofs are collected in the Appendix.

2. THE MODEL AND THE ESTIMATOR

2.1. The main model assumption

We assume that a latent $d^* \times 1$ -vector X^* satisfies the following moment condition:

$$E[m(X^*, \beta_o)] = 0, \quad (2)$$

uniquely at some unknown parameter $\beta_o \in B$, a compact subset of \mathcal{R}^q with $1 \leq q \leq r$, where m is a $r \times 1$ -vector of known moment functions that may be non-linear and non-smooth in X^* and/or β_o , and the distribution of X^* is unspecified. We also assume that we have access to two data-sets: the **primary data-set** $\{X_{pi} : i = 1, \dots, n_p\}$ and the **auxiliary data-set** $\{(X_{vj}^*, X_{vj}) : j = 1, \dots, n_v\}$, where $X_{pi}, X_{vj}^*, X_{vj} \in \mathcal{R}^d$ and $n_v < n_p$. The object of interest is the parameter β_o which is related to the true unknown density $f_{X^*} \equiv f_{X_p^*}$ of the latent variable X^* .

The auxiliary data-set allows us to recover information about arbitrary correlation between X_p^* and X_p . This model nests both the classical error in variables model and the contamination model as special cases. In the following we denote by $f_{X_p}, f_{X_p^*}, f_{X_v}$ and $f_{X_v^*}$ the marginal densities of the proxy variable and the latent variable in the primary and auxiliary data-set, respectively, by $f_{X_p^*|X_p}$ and $f_{X_v^*|X_v}$ the conditional densities of the latent variable given the proxy variable in the primary and auxiliary data-sets; and by $f_{X_p^*, X_p}$ and $f_{X_v^*, X_v}$ the joint densities.¹ We also denote E_p and E_v as expectations with respect to the primary sample and auxiliary sample, respectively. Let

$$g(x, \beta) \equiv E[m(X_p^*, \beta) | X_p = x] = \int m(x^*, \beta) f_{X_p^*|X_p=x}(x^*) dx^*$$

then (2) implies, by the law of iterated expectation, that uniquely at $\beta = \beta_o$

$$E_p[g(X, \beta_o)] = \int g(x, \beta_o) f_{X_p}(x) dx = 0. \quad (3)$$

1. Here all the densities are understood to be Radon–Nykodym derivatives of corresponding probability measures with respect to products of Lebesgue measures and counting measures. That is, we allow for both continuous and discrete random variables.

This gives a set of moment conditions based on the primary data. The next assumption is crucial in allowing us to use the auxiliary data to recover the correlation between the mismeasured variables and the true variables.

Assumption 1 (Main Assumption). $f_{X_v^*|X_v=x} = f_{X_p^*|X_p=x}$ for all x in the support of X_p in \mathcal{R}^d .

This assumption implies that for each fixed β ,

$$g(x, \beta) = E[m(X_v^*, \beta) | X_v = x] = \int m(x^*, \beta) f_{X_v^*|X_v=x}(x^*) dx^*$$

hence information about $g(x, \beta)$ can be recovered from the auxiliary sample and thus one can use the auxiliary sample to consistently estimate β_o by combining this main assumption with the moment condition (3) for the primary sample. Note that in our set-up the random variables X_p, X_v, X_p^*, X_v^* do not need to measure the same thing, and the dimensions of X_p, X_v may differ from those of X_p^*, X_v^* ; see the empirical section for an example.

The motivation of Assumption 1 is statistical in nature and should relate to the nature of the measurement error available in the data. Nevertheless, there is a link between this assumption and the identifying assumption (the so-called “strong ignorability condition” or the “selection on observables” condition) typically made for matching estimators in the program evaluation literature (see Heckman, Ichimura and Todd, 1998). There, conditional on a set of observed covariates, the researcher is interested in identifying the counterfactual untreated distribution of the treated group should they have not been treated. The “strong ignorability condition” assumes that this is equal to the observed outcome distribution of the untreated group. Hence, by using Assumption 1, one can replace the projection of the counterfactual (unobserved) outcome distribution in the treated group by the projection of the (observed) outcome distribution in the untreated group and then average over the regressors. A closely related assumption is used in Wooldridge (1995, 2002). Interestingly, this assumption is also closely related to the notion of “superexogeneity” as discussed in Engle, Hendry and Richard (1983) and White (1994).

It is important to note that Assumption 1 is satisfied in the **stratified sampling** design where a *non-random response based subsample* of the primary data is validated. In a typical example of this stratified sampling design, we first oversample a certain subpopulation of the mismeasured variables X , and then validate the true variables X^* corresponding to this non-random stratified subsample of X . It is very natural and sensible to oversample a subpopulation of the primary data-set where more severe measurement error is suspected to be present. Assumption 1 is valid as long as, in this sampling procedure of the auxiliary data-set, the sampling scheme of X_v is based only on information available in the distribution of the primary data-set $\{X_p\}$. For example, one can choose a subset of the primary data-set $\{X_p\}$ and validate the corresponding $\{X^*\}$, in which case the X_v ’s in the auxiliary data-set are a subset of the primary data X_p . The stratified sampling procedure can be illustrated as follows. Let U_{pi} be i.i.d. $U(0, 1)$ random variables independent of both X_{pi} and X_{pi}^* , and let $T(X_{pi}) \in (0, 1)$ be a measurable function of the primary data. The stratified sample is obtained by validating every observation for which $U_{pi} < T(X_{pi})$. In other words, $T(X_{pi})$ specifies the probability of validating an observation after X_{pi} is observed. This implies, for example, that Assumption 1 is clearly satisfied, although

$$f_{X_v}(x) = f_{X_p}(x)T(x) \Big/ \int f_{X_p}(y)T(y)dy \neq f_{X_p}(x)$$

unless $T(x)$ is constant. On the other hand, in the case of a **validation data-set**, *i.e.* the auxiliary data-set (X_v^*, X_v) is derived from the same sample distribution as the primary data-

set, $f_{X_v^*, X_v} = f_{X_p^*, X_p}$, Assumption 1 is trivially satisfied. This is the common case that is used in the statistics literature where usually a random subset of the primary data is validated. With validation data, we can also use moment condition (2) directly to estimate β_o . In Section 3 we show that efficiency gains can be obtained by optimally combining moment conditions (2) and (3).

2.2. The estimators

In this section, we provide estimators that can be used, under Assumption 1, to consistently estimate the parameter β_o .

2.2.1. General auxiliary data case. Under Assumption 1 and the moment condition (3), we can define a generalized method of moments (GMM) estimator $\hat{\beta}$ of β_o as

$$\hat{\beta} = \arg \min_{\beta} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{g}(X_{pi}, \beta) \right)' \widehat{W} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{g}(X_{pi}, \beta) \right), \quad (4)$$

where \widehat{W} is some random positive definite symmetric weighting matrix, and $\hat{g}(x, \beta)$ is a non-parametric estimator of $g(x, \beta)$ using the auxiliary data-set $\{(X_{vj}^*, X_{vj}) : j = 1, \dots, n_v\}$. We consider a series (sieve) least squares estimator of $g(x, \beta)$:

$$\hat{g}(x, \beta) = \sum_{j=1}^{n_v} m(X_{vj}^*, \beta) p^{k_{nv}}(X_{vj})' (P_v' P_v)^{-1} p^{k_{nv}}(x), \quad (5)$$

where $\{p_l(x), l = 1, 2, \dots\}$ denotes a sequence of known basis functions that can approximate any square-integrable function of x well (to be more precise later), $p^{k_{nv}}(x) = (p_1(x), \dots, p_{k_{nv}}(x))'$ and $P_v = (p^{k_{nv}}(X_{v1}), \dots, p^{k_{nv}}(X_{vn_v}))'$ for some integer k_{nv} , with $k_{nv} \rightarrow \infty$ and $k_{nv}/n_v \rightarrow 0$ as $n_v \rightarrow \infty$.

One can also use a kernel estimator for $g(x, \beta)$:

$$\tilde{g}(x, \beta) = \frac{1}{n_v a^d} \sum_{j=1}^{n_v} \frac{m(X_{vj}^*, \beta) K\left(\frac{X_{vj} - x}{a}\right)}{\frac{1}{n_v a^d} \sum_{j=1}^{n_v} K\left(\frac{X_{vj} - x}{a}\right)},$$

where $K(\cdot) : \mathcal{R}^d \rightarrow \mathcal{R}$ is a kernel function, and a is a bandwidth satisfying $n_v a^d \rightarrow \infty$, $a \rightarrow 0$ as $n_v \rightarrow \infty$. However, this kernel based estimator requires tedious trimming arguments to guard against small values of the denominator. These trimming arguments require strong support conditions on the data and are difficult to interpret.

Although we assume $f_{X_v^*|X_v} = f_{X_p^*|X_p}$, the marginal density f_{X_v} may be different from f_{X_p} . Typically the supports of these two distributions are the same, but the distributions themselves are different. This implies that $f_{X_v^*}$ may be different from $f_{X_p^*}$ but with the same support. That is why the estimator $\hat{\beta}$ given in (4) is based only on the moment condition (3). Since we do not observe X^* in the primary sample, we cannot directly use the moment condition (2) to construct the usual GMM estimator for β_o . However, one can non-parametrically estimate $f_{X_v^*|X_v}$ from the auxiliary sample, and f_{X_p} from the primary sample. Under the assumption $f_{X_v^*|X_v} = f_{X_p^*|X_p}$, we can then non-parametrically estimate $f_{X_p^*}(\cdot) = \int f_{X_v^*|X_v=x}(\cdot) f_{X_p}(x) dx$ by $\widehat{f_{X_p^*}}(\cdot) = \int \widehat{f_{X_v^*|X_v=x}}(\cdot) \widehat{f_{X_p}}(x) dx$. An estimate of β_o can be obtained by either

$$\min_{\beta} \left(\int m(x^*, \beta) \widehat{f_{X_p^*}}(x^*) dx^* \right)' \widehat{W} \left(\int m(x^*, \beta) \widehat{f_{X_p^*}}(x^*) dx^* \right)$$

or

$$\min_{\beta} \left(\frac{1}{S} \sum_{s=1}^S m(X_{ps}^*, \beta) \right)' \widehat{W} \left(\frac{1}{S} \sum_{s=1}^S m(X_{ps}^*, \beta) \right)$$

where $\{X_{ps}^* : s = 1, \dots, S\}$ is a simulated sample from $\widehat{f_{X_p^*}}$. We do not implement these alternative estimators since both are computationally more involved and more difficult to implement than the estimator $\widehat{\beta}$ given in (4).

2.2.2. Estimation with validation data. If in addition to Assumption 1 we have $f_{X_v} = f_{X_p}$, then $f_{X_v^*} = f_{X_p^*}$. Our model (2) implies both (3) and the following additional moment condition:

$$E_v[m(X_v^*, \beta_o)] = 0. \tag{6}$$

In this case the auxiliary data-set is actually a validation data-set, *i.e.* both the validation and the primary data come from the same population.

We can construct a GMM estimator $\widehat{\beta}$ that exploits the information in both (3) and (6), *i.e.*

$$\min_{\beta} \begin{bmatrix} \frac{1}{n_v} \sum_{j=1}^{n_v} m(X_{vj}^*, \beta) \\ \frac{1}{n_p} \sum_{i=1}^{n_p} \widehat{g}(X_{pi}, \beta) \end{bmatrix}' \widehat{W} \begin{bmatrix} \frac{1}{n_v} \sum_{j=1}^{n_v} m(X_{vj}^*, \beta) \\ \frac{1}{n_p} \sum_{i=1}^{n_p} \widehat{g}(X_{pi}, \beta) \end{bmatrix} \tag{7}$$

where \widehat{W} is an appropriate weighing matrix and $\widehat{g}(\cdot, \beta)$ is a non-parametric estimate of $g(\cdot, \beta)$ based on the auxiliary data, such as the sieve estimator given in (5).

In the case of a non-linear regression with validation data, Lee and Sepanski (1995) replace the conditional expectation $g(X, \beta) = E[m(X^*, \beta) | X]$ with a linear projection $h(X, \beta) = \text{Proj}(m(X^*, \beta) | L(X))$, where $L(X)$ is a finite dimensional linear combination of a *fixed* number of functions of X . It is easily seen that as long as $L(X)$ includes a constant term, then in the case where the marginal distribution of X_v is the same as the marginal distribution of X_p we have

$$0 = E_v[m(X^*, \beta_o)] = E_v[\text{Proj}(m(X^*, \beta_o) | L(X))] = E_p[\text{Proj}(m(X^*, \beta_o) | L(X))]. \tag{8}$$

Lee and Sepanski (1995) used the sample projection to estimate the population projection, and replaced $\widehat{g}(X_{pi}, \beta)$ in (7) by $\widehat{h}(X_{pi}, \beta)$, the sample least squares projection of $m(X^*, \beta_o)$ onto $L(x)$, to construct a method of moments estimator of β_o . We show in Section 3 that our estimator based on (7) is more efficient than the one based on this finite dimensional least squares projection. In addition, our estimator based on (4) works even if the two marginal distributions are different.²

3. LARGE SAMPLE RESULTS

In this section, we provide the consistency and asymptotic normality of our semiparametric estimator $\widehat{\beta}$ given in (4) for the case of a general auxiliary sample. We also provide simple consistent estimators of the asymptotic variance of $\widehat{\beta}$.

3.1. Consistency

Assumption 2. *Let the following hold.*

2. In practice, one could non-parametrically test the null hypothesis of $f_{X_p}(\cdot) = f_{X_v}(\cdot)$.

- (1) The primary data-set $\{X_{pi} : i = 1, \dots, n_p\}$ is an i.i.d. sample drawn from f_{X_p} over $\mathcal{X} \subseteq \mathcal{R}^d$; the auxiliary data-set $\{(X_{vj}^*, X_{vj}) : j = 1, \dots, n_v\}$ is an i.i.d. sample drawn from $f_{X_v^*, X_v}$.
- (2) $n_v < n_p, n_v \rightarrow \infty, n_p \rightarrow \infty$, and $\lambda = \lim_{n_v \rightarrow \infty} (n_v/n_p) \in [0, \infty)$.
- (3) $\widehat{W} - W = o_p(1)$, where W is a positive semidefinite matrix.
- (4) $WE[g(X_p, \beta)] = 0$ has a unique solution on B at β_o , with B a compact subset of \mathcal{R}^q .

Assumptions 2(1) and 2(2) require mild restrictions on the relation between the primary data-set and the auxiliary data-set. Assumptions 2(3) and 2(4) are standard regularity conditions for consistency of the GMM estimator of β_o with known functional form $g(\cdot, \beta)$.

Under Assumptions 1 and 2, the estimator $\hat{\beta}$ defined in (4) can be shown to converge to β_o as long as the sieve estimator $\hat{g}(\cdot, \beta)$ defined in (5) converges to $g(\cdot, \beta)$ in some metric. Since the supports of the mismeasured variables could be unbounded, $\mathcal{X} = \mathcal{R}^d$, we use a weighted sup-norm metric defined as $\|h\|_{\infty, \omega} \equiv \sup_{x \in \mathcal{X}, \beta \in B} |h(x, \beta)[1 + |x|^2]^{-\omega/2}|$ for some $\omega > 0$. Assumptions 3(1) and 3(5) below are sufficient to ensure that \hat{g} defined in (5) converges to g under the norm $\|\cdot\|_{\infty, \omega}$.

Assumption 3. Let the following hold:

- (1) for all $\beta \in B$, $g(\cdot, \beta)$ is $H(\gamma, \omega_1)$ -smooth for some $\gamma > 0, \omega_1 \geq 0$;
- (2) $\int (1 + |x|^2)^\omega f_{X_p}(x) dx < \infty, \int (1 + |x|^2)^\omega f_{X_v}(x) dx < \infty$ for some $\omega > \omega_1 \geq 0$;
- (3) for each fixed x , $g(x, \beta)$ is continuous at β for all $\beta \in B$;
- (4) $\text{Var}[\{m(X_v^*, \beta) - g(X_v, \beta)\} | X_v = x]$ is bounded uniformly over x and β ;
- (5) for any $H(\gamma, \omega_1)$ -smooth function $g(\cdot, \beta)$, there is a function $\Pi_{\infty n} g$ in the sieve space $\mathcal{G}_n = \{h(\cdot, \beta) = p^{k_{nv}}(\cdot)' \pi(\beta)\}$ such that $\|g(\cdot, \beta) - \Pi_{\infty n} g(\cdot, \beta)\|_{\infty, \omega} = o(1)$. Also $E_v[p^{k_{nv}}(X)p^{k_{nv}}(X)']$ is non-singular uniformly in k_{nv} .

To estimate the unknown function $g(\cdot, \beta)$ well by the sieve estimator $\hat{g}(\cdot, \beta)$, we need to assume that $g(x, \beta)$ is smooth in some sense with respect to x_c , the vector of continuous elements of x . (For simplicity, we assume that $x_c = x$ in this paper.) Assumption 3(1) is a standard weighted smoothness condition imposed on the function $g(\cdot; \beta)$; see Appendix A for the definition of $H(\gamma, \omega_1)$ -smooth and other technical details. Assumption 3(2) is a typical condition on the tail behaviour of the marginal densities. In Appendix A we show that under Assumptions 3(1) and 3(5), $\|\hat{g} - g\|_{\infty, \omega} = o_p(1)$. The weight function $[1 + |x|^2]^{-\omega/2}$ employed in our sieve estimation of $g(\cdot, \beta)$ could be regarded as an alternative to the trimming function used in kernel estimation when the support is unbounded, $\mathcal{X} = \mathcal{R}^d$. This kind of smooth weighting to deal with unbounded support³ has been used in Chen, Hansen and Scheinkman (1997) and Ai and Chen (2003). Alternatively one could also impose different weighted function space and different sieve basis such as the Hermite polynomial sieves in Gallant and Nychka (1987).

Assumptions 3(3) and 3(4) are sufficient conditions to ensure consistency of the method of moments estimator of β_o when the functional form of the moment $g(x, \beta) \equiv E[m(X_v^*, \beta) | X_v = x]$ is known. Note that Assumption 3(3) allows for $m(X_v^*, \beta)$ to be non-smooth such as quantile based moment functions.

Theorem 1. Let $\hat{\beta}$ be given in (4). Under Assumptions 1–3, if $\frac{k_{nv}}{n_v} \rightarrow 0, k_{nv} \rightarrow \infty$, then

$$\hat{\beta} - \beta_o = o_p(1).$$

3. The weighting used here will not work in cases where the supports of the data are bounded but the densities go to zero at the boundaries. In those cases, alternative weighting can be used.

3.2. Asymptotic distribution

The following assumptions are used to study the asymptotic behaviour of $\hat{\beta}$.

Assumption 4. Let β_o be an interior point of B , and

- (1) $G'WG$ is finite positive definite where $G = E_p \left[\frac{\partial g(X_{pi}, \beta_o)}{\partial \beta'} \right]$;
- (2) $E_p[g(x, \beta_o)g(x, \beta_o)']$ is finite and positive definite;
- (3) for each fixed x , and for some $\delta > 0$, $\frac{\partial g(x, \beta)}{\partial \beta'}$ is continuous in $\beta \in B$ with $|\beta - \beta_o| \leq \delta$,

$$E_p \left[\sup_{\beta: |\beta - \beta_o| \leq \delta} \left| \frac{\partial g(X_p, \beta)}{\partial \beta'} \right| \right] < \infty;$$

- (4) there exist a constant $\epsilon \in (0, 1]$, a $\delta > 0$ and a measurable function $b(\cdot)$ with $E_p[b(X_p)] < \infty$ such that $\left| \frac{\partial \tilde{g}(x, \beta)}{\partial \beta'} - \frac{\partial g(x, \beta)}{\partial \beta'} \right| \leq b(x)[\|\tilde{g} - g\|_{\infty, \omega}]^\epsilon$ for all $\beta \in B$ with $|\beta - \beta_o| \leq \delta$ and all $H(\gamma, \omega_1)$ -smooth function \tilde{g} with $\|\tilde{g} - g\|_{\infty, \omega} \leq \delta$.

Assumptions 4(1)–4(3) are the usual regularity and dominance conditions to ensure root- n normality of the GMM estimator of β_o when the functional form $g(\cdot, \beta)$ is known. Assumptions 3 and 4(4), together with the following Assumption 5, ensure that the unknown $g(\cdot, \beta)$ can be replaced by the sieve estimator $\hat{g}(\cdot, \beta)$, and that the resulting estimator $\hat{\beta}$ defined in (4) is still root- n consistent and asymptotically normally distributed. In the following, for any square measurable function $h : \mathcal{X} \rightarrow \mathcal{R}$ we define a Hilbert norm $\|h\|_{2,v} \equiv \sqrt{\int h(x)^2 f_{X_v}(x) dx} < \infty$. Also we let $\Pi_{2n}h$ denote the orthogonal projection of h onto the closed linear span of $p^{k_{nv}}(x) = (p_1(x), \dots, p_{k_{nv}}(x))'$ under the norm $\|\cdot\|_{2,v}$.

Assumption 5. Let the following hold:

- (1) $E_v \left[\left(\frac{f_{X_p}(x)}{f_{X_v}(x)} \right)^2 \right] < \infty$;
- (2) Assumption 3(1) is satisfied with $\gamma > d/2$, and Assumption 3(2) is satisfied with $\omega > \omega_1 + \gamma$;
- (3) $k_{nv} = O \left((n_v)^{\frac{d}{2\gamma+d}} \right)$;
- (4) $(n_v)^{-\frac{\gamma}{2\gamma+d}} \times \left\| \frac{f_{X_p}(\cdot)}{f_{X_v}(\cdot)} - \Pi_{2n} \frac{f_{X_p}(\cdot)}{f_{X_v}(\cdot)} \right\|_{2,v} = o(n_v^{-1/2})$.

Assumption 5(1) is satisfied when $[f_{X_p}(x)]^2 \approx f_{X_v}(x)/\sqrt{x'x}^{-1+\epsilon}$ for all large $x'x$ and for some small $\epsilon > 0$. It is also satisfied when f_{X_p} is absolutely continuous with respect to f_{X_v} and $\sup_x \frac{f_{X_p}(x)}{f_{X_v}(x)} < \infty$, which is again satisfied with the validation sample. For the stratification scheme suggested in Section 2.1, Assumption 5(1) is satisfied as long as $E_v[(\frac{1}{T(X)})^2] < \infty$. Assumption 5(2) is a stronger version of Assumptions 3(1) and 3(2). Assumptions 5(2), 5(3) and 3 together imply that $\|\hat{g}(\cdot, \beta_o) - g(\cdot, \beta_o)\|_{2,v} = O_p((n_v)^{-\frac{\gamma}{2\gamma+d}})$, the same as Stone's optimal convergence rate while we allow for unbounded support $\mathcal{X} \subseteq \mathcal{R}^d$. Assumption 5(4) will be satisfied when $\frac{f_{X_p}(\cdot)}{f_{X_v}(\cdot)}$ is a little bit smooth such that $\left\| \frac{f_{X_p}(\cdot)}{f_{X_v}(\cdot)} - \Pi_{2n} \frac{f_{X_p}(\cdot)}{f_{X_v}(\cdot)} \right\|_{2,v} = o \left(n_v^{-\frac{d}{2(2\gamma+d)}} \right)$.

Theorem 2. Let $\hat{\beta}$ be given in (4). Under Assumptions 1–5, we have $\sqrt{n_v}(\hat{\beta} - \beta_o) \Rightarrow \mathcal{N}(0, V)$, with

$$V = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$$

$$\Omega = \text{Avar} \left(\frac{1}{\sqrt{n_v}} \sum_{j=1}^{n_v} \frac{f_{X_p}(X_{vj})}{f_{X_v}(X_{vj})} \{m(X_{vj}^*, \beta_o) - g(X_{vj}, \beta_o)\} + \frac{\sqrt{n_v}}{n_p} \sum_{i=1}^{n_p} g(X_{pi}, \beta_o) \right).$$

Theorem 2 holds regardless of the correlation between the primary and the auxiliary data-set. Note that, as long as there is non-degenerate measurement error distribution in the primary data, the $\sqrt{n_v}$ convergence rate in Theorem 2 applies even when $n_v/n_p \rightarrow 0$ (where $\lambda \rightarrow 0$). However, if the primary data-set is measured without error, then $m(X^*, \beta) \equiv g(X, \beta)$ and $\Omega = \lambda E_p[g(X, \beta_o)g(X, \beta_o)']$. In this case, $\lambda \rightarrow 0$ implies $V \rightarrow 0$ and the correct normalization for a non-degenerate limiting distribution should be $\sqrt{n_p}$.

Notice that the expression of Ω can be simplified under two most important cases: the case where the auxiliary sample is independent of the primary sample, and the case where the auxiliary data-set is a subset of the original primary data. We will call the latter case “*validate in sample*” case, which is relevant if the researcher collects the primary data first, and then decides to validate a subset of the primary data based on the observations of variables X_{pi} . Such is the case for the empirical illustration we consider in Section 4. For both cases, we have

$$\Omega = \Omega_1 + \lambda E_p[g(X, \beta_o)g(X, \beta_o)'], \quad (9)$$

$$\Omega_1 \equiv E_v \left[\left(\frac{f_{X_p}(X)}{f_{X_v}(X)} \right)^2 \{m(X^*, \beta_o) - g(X, \beta_o)\} \{m(X^*, \beta_o) - g(X, \beta_o)\}' \right].$$

For more on the form of the asymptotic variance in the “*validate in sample*” case and the stratified scheme suggested in Section 2.1, see Appendix B below.

3.3. Consistent estimation of the asymptotic variance

Each component of V , the asymptotic variance of $\hat{\beta}$, can be consistently estimated. Let \hat{V} and \hat{G} be such that

$$\hat{V} = (\hat{G}'W\hat{G})^{-1}\hat{G}'W\hat{\Omega}W\hat{G}(\hat{G}'W\hat{G})^{-1}, \quad \hat{G} = \frac{1}{n_p} \sum_{i=1}^{n_p} \frac{\partial \hat{g}(X_{pi}, \hat{\beta})}{\partial \beta'}.$$

We also denote

$$V_o = (G'\Omega^{-1}G)^{-1} \quad \text{and} \quad \hat{V}_o = (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}.$$

For simplicity we consider the two most important cases: (1) the auxiliary data-set is independent of the primary data-set; (2) the auxiliary data-set is a subset of the primary data-set. In both cases, Ω is given in (9), and can be consistently estimated by

$$\hat{\Omega} = \frac{1}{n_v} \sum_{j=1}^{n_v} (\hat{v}_{vj}^* \hat{U}_{vj}) (\hat{v}_{vj}^* \hat{U}_{vj})' + \frac{n_v}{n_p^2} \sum_{i=1}^{n_p} (\hat{g}(X_{pi}, \hat{\beta}) \hat{g}(X_{pi}, \hat{\beta})')$$

$$\hat{U}_{vj} = m(X_{vj}^*, \hat{\beta}) - \hat{g}(X_{vj}, \hat{\beta})$$

$$\hat{v}_{vj}^* = \left[\frac{1}{n_p} \sum_{i=1}^{n_p} p^{k_{nv}}(X_{pi}) \right]' \left(\frac{P_v' P_v}{n_v} \right)^{-1} p^{k_{nv}}(X_{vj}). \quad (10)$$

The idea behind our estimator (10) for $v_{vj}^* \equiv f_{X_p}(X_{vj})/f_{X_v}(X_{vj})$ is as follows. Let $v^*(x) \equiv f_{X_p}(x)/f_{X_v}(x)$, then $\Pi_{2n} v^*(x) = \{(E_v[p^{k_{nv}}(X)p^{k_{nv}}(X)']^{-1} E_v[p^{k_{nv}}(X)v^*(X)]\}' p^{k_{nv}}(x)$ by definition. Now $\frac{1}{n_p} \sum_{i=1}^{n_p} p^{k_{nv}}(X_{pi})$ is a consistent estimator of $\int p^{k_{nv}}(x) f_{X_p}(x) dx =$

$E_v[p^{k_{nv}}(X)v^*(X)]$, and $\frac{P'_v P_v}{n_v}$ is a consistent estimator of $E_v[p^{k_{nv}}(X)p^{k_{nv}}(X)']$. Hence $[\frac{1}{n_p} \sum_{i=1}^{n_p} p^{k_{nv}}(X_{pi})]'(\frac{P'_v P_v}{n_v})^{-1} p^{k_{nv}}(x)$ is a consistent estimator of $\Pi_{2n}v^*(x)$ in some sense, which in turn approximates $v^*(x)$ under Assumption 5(1) as $k_{nv} \rightarrow \infty$.

There are of course many other consistent estimators of v_{vj}^* based on consistent estimators of densities f_{X_p} and f_{X_v} . For example, (a) kernel plug-in estimator $\widehat{v}_{vj}^* = \widehat{f}_{X_p}(X_{vj})/\widehat{f}_{X_v}(X_{vj})$, with

$$\widehat{f}_{X_p}(x) = \frac{1}{n_p a_p^d} \sum_{i=1}^{n_p} K\left(\frac{X_{pi} - x}{a_p}\right), \quad \widehat{f}_{X_v}(x) = \frac{1}{n_v a_v^d} \sum_{j=1}^{n_v} K\left(\frac{X_{vj} - x}{a_v}\right)$$

where $K(\cdot) : \mathcal{R}^d \rightarrow \mathcal{R}$ is a kernel function, and a_p, a_v are bandwidth parameters going to zero slowly. For simplicity we can set $a_p = a_v \rightarrow 0, n_v a_v^d \rightarrow \infty$. (b) Orthogonal series plug-in estimator $\widehat{v}_{vj}^* = \widehat{f}_{X_p}(X_{vj})/\widehat{f}_{X_v}(X_{vj})$, with

$$\begin{aligned} \widehat{f}_{X_p}(x) &= \sum_{\ell=1}^{J_{np}} \widehat{\pi}_{p\ell} A_\ell(x), & \widehat{\pi}_{p\ell} &= \frac{1}{n_p} \sum_{i=1}^{n_p} A_\ell(X_{pi}) \\ \widehat{f}_{X_v}(x) &= \sum_{\ell=1}^{J_{nv}} \widehat{\pi}_{v\ell} A_\ell(x), & \widehat{\pi}_{v\ell} &= \frac{1}{n_v} \sum_{i=1}^{n_v} A_\ell(X_{vi}) \end{aligned}$$

where $\{A_\ell(x) : \ell = 1, \dots, \infty\}$ is an orthonormal basis for the space of square-integrable functions against the Lebesgue measure on \mathcal{R}^d , and J_{np} and J_{nv} are series smoothing parameters going to infinity slowly. For simplicity we can set $J_{np} = J_{nv} \rightarrow \infty, \frac{J_{nv}}{n_v} \rightarrow 0$. Although the orthogonal series estimator of a density might not be non-negative in a finite sample, it does not matter for the purpose of constructing a consistent estimator of $\Omega_1 = E_v[(v_{vj}^* U_{vj})(v_{vj}^* U_{vj})']$.

Assumption 6. *Let the following hold with $v_v^* \equiv \{f_{X_p}(X_v)/f_{X_v}(X_v)\}$:*

- (1) *There exists a measurable function $b(\cdot)$ such that $\sup_{\beta \in B: |\beta - \beta_o| = o(1)} \left| \frac{\partial g(x, \beta)}{\partial \beta'} \right| \leq b(x)$ and $E_v[\{b(X_v)v_v^*\}^2] < \infty$;*
- (2) *$E_v[(1 + X'_v X_v)^\omega v_v^{*2}] < \infty$ for some $\omega > \omega_1 \geq 0$;*
- (3) *There is a constant $s \in (0, 2]$ such that for all small $\delta > 0$,*

$$E_v \left[v_v^{*2} \sup_{\beta \in B: |\beta - \beta_o| \leq \delta} \left| \{m(X_v^*, \beta) - m(X_v^*, \beta_o)\} \{m(X_v^*, \beta) - m(X_v^*, \beta_o)\}' \right| \right] = O(\delta^s).$$

Assumption 6(3) is trivially satisfied if $m(x, \beta)$ is (pointwise) Lipschitz continuous in (x, β) . Assumption 6(3) also allows for non-smooth moment conditions such as quantile based moments. The next theorem provides the desired result.

Theorem 3. *Under Assumptions 1–4, 5(1) and 6, Ω given in (9), and $\frac{k_{nv}}{n_v} \rightarrow 0, k_{nv} \rightarrow \infty$, we have (i) $\widehat{V} = V + o_p(1)$ and $\widehat{V}_o = V_o + o_p(1)$; (ii) further, if Assumption 5(2)–5(4) and $W = \Omega^{-1} + o_p(1)$ hold, then $\sqrt{n_v}(\widehat{\beta} - \beta_o) \Rightarrow \mathcal{N}(0, V_o)$.*

In the case in which the auxiliary data are a subset of the primary data, we can also estimate Ω by

$$\begin{aligned} \bar{\Omega} &= \frac{1}{n_v} \sum_{j=1}^{n_v} (\widehat{v}_{vj}^* \widehat{U}_{vj})(\widehat{v}_{vj}^* \widehat{U}_{vj})' + \frac{n_v}{n_p^2} \sum_{i=1}^{n_p} (\widehat{g}(X_{pi}, \widehat{\beta}) \widehat{g}(X_{pi}, \widehat{\beta})') \\ &\quad + \frac{2}{n_p} \sum_{j=1}^{n_v} \widehat{v}_{vj}^* \widehat{U}_{vj} \widehat{g}(X_{vj}, \widehat{\beta})'. \end{aligned}$$

Under our assumption that $f_{X_v^*|X_v} = f_{X_p^*|X_p}$, $\widehat{\Omega} - \bar{\Omega} = o_p(1)$. The difference between $\widehat{\Omega}$ and $\bar{\Omega}$ can also be used as a specification check on the plausibility of Assumption 1. As an alternative to consistent estimation of the limiting variance covariance matrix, resampling methods including bootstrap and subsampling methods could be used for inference. In both special cases that we considered, the observations with both the primary data-set and the auxiliary data-set are assumed to be independent from each other and identically distributed. If serial correlations, heteroscedasticity, cluster structure or panel data structure are present in either or both data-sets, we will need to take into account these correlation structures in estimating the limiting variance of Theorem 2 or in the resampling procedures.

3.4. Efficiency gains with a validation sample

When the primary is independent of the auxiliary data and when they both come from the same distribution, we can use the two sets of moment conditions (3) and (6) above. These explore the information in both the auxiliary sample (the validation data in this case) and the primary sample, and should produce a more efficient estimate of β . Let $\hat{\beta}$ be given in (7). Then, under mild conditions and by a similar proof as that for Theorem 2, we obtain $\sqrt{n_v}(\hat{\beta} - \beta_o) \Rightarrow \mathcal{N}(0, V)$, with

$$V = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$$

$$G = \left\{ \frac{\partial}{\partial \beta} E[m(X_{pi}, \beta), g(X_{pi}, \beta)]' \right\} \Big|_{\beta=\beta_o} = [I, I]'A \quad \text{with} \quad A = \frac{\partial E[m(X_{pi}, \beta)]}{\partial \beta'} \Big|_{\beta=\beta_o}$$

$$\Omega = \begin{bmatrix} V_2 & V_2 - V_1 \\ V_2 - V_1 & \lambda V_1 + (V_2 - V_1) \end{bmatrix}$$

where $V_1 = E[g(X_p, \beta_o)g(X_p, \beta_o)']$ and $V_2 = E[m(X_v, \beta_o)m(X_v, \beta_o)']$. Further if $W = \Omega^{-1}$, then

$$\sqrt{n_v}(\hat{\beta} - \beta_o) \Rightarrow \mathcal{N}(0, V_o), \quad \text{with}$$

$$V_o = (G'\Omega^{-1}G)^{-1} = (A(V_2^{-1} + V_2^{-1}[(\lambda + 1)V_1^{-1} - V_2^{-1}]^{-1}V_2^{-1})A')^{-1}.$$

Let $\tilde{\beta}$ be the optimally weighted GMM estimator of β_o using only the moment condition (6) and the validation data-set, $\frac{1}{n_v} \sum_{j=1}^{n_v} m(X_{vj}^*, \beta)$. Then under mild conditions,

$$\sqrt{n_v}(\tilde{\beta} - \beta_o) \Rightarrow \mathcal{N}(0, (AV_2^{-1}A')^{-1}).$$

The efficiency gain of using both moment conditions (3) and (6) is confirmed by observing that

$$V_o^{-1} = AV_2^{-1}(I + [(\lambda + 1)V_1^{-1}V_2 - I]^{-1})A' \geq AV_2^{-1}A',$$

since $V_2 \geq V_1$ and $\lambda \geq 0$. Moreover, the efficiency gain is monotonically decreasing in λ . In particular, the efficiency gain disappears if $\lambda = \lim_{n_v \rightarrow \infty} (n_v/n_p) = \infty$, in which case one can ignore the primary sample. And the gain in efficiency is maximized when $\lambda = 0$. If $\lambda = 0$ and the primary data-set is measured without error, then $V_1 = V_2$, $V_0 = 0$ and $V_0^{-1} = \infty$, the efficiency gain from combining both moment conditions is infinitely large.

With true validation data, it is also easy to show that if we replace our sieve based estimate of the conditional expectation $\hat{g}(X_{pi}, \beta)$ in (7) by the fixed finite dimensional linear projection $\hat{h}(X_{pi}, \beta)$ of Lee and Sepanski (1995), the resulting optimally weighted GMM estimator $\bar{\beta}$ of β_o will satisfy

$$\sqrt{n_v}(\bar{\beta} - \beta_o) \Rightarrow \mathcal{N}(0, \bar{V}), \quad \text{with}$$

$$\bar{V} = (A(V_2^{-1} + V_2^{-1}[(\lambda + 1)\bar{V}_1^{-1} - V_2^{-1}]^{-1}V_2^{-1})A')^{-1}, \quad \bar{V}_1 = Eh(X_p, \beta_o)h(X_p, \beta_o)'$$

TABLE 1
Descriptive statistics for match and unmatched respondents

Variable	SSR unmatched (st. dev.)	Matched (st. dev.)
Earnings	23,569 (12,470)	24,528 (11,854)
Age	44.9 (3.19)	45.04 (3.22)
Education	13.4 (3.22)	13.5 (3.07)
Race	0.89 (0.303)	0.92 (0.269)

Therefore, Lee and Sepanski's (1995) fixed projection estimator is more efficient than using the validation sample alone since $\bar{V}_1 \leq V_2$ and $\bar{V}^{-1} - (AV_2^{-1}A') \geq 0$; but it is less efficient than our sieve based estimator since $V_1 \geq \bar{V}_1$ and $V_o^{-1} - \bar{V}^{-1} \geq 0$. This is natural because the conditional expectation $g(X_p, \beta_o)$ explains more variation in $m(\cdot, \beta_o)$ than any fixed linear projection $h(X_p, \beta_o)$ does.

4. EMPIRICAL ILLUSTRATION: RETURNS TO SCHOOLING FROM THE CPS/SSR 1978 MATCH FILES

To illustrate our methods empirically, we use a data-set that matches the Current Population Survey (CPS) to employer-reported social security earnings (SSR) from 1978 (the CPS/SSR Exact Match File). This data-set has been used by Bound and Krueger (1991) and Bollinger (1998) to study the extent of measurement error in earnings. They both document correlation between measurement error and the true variable. Individuals in the CPS were asked about their social security numbers. For those who provided the number, their earnings were matched against the social security records. Applied researchers are often interested in using the CPS data-set to implement a standard Mincer regression to compute returns to education and experience. However, income in the CPS data-set is mismeasured. In this section, we show how one can use our results to make inferences on parameters in such a returns to education regression. Another complication arises due to the top-coding (censoring) of the social security earnings. The estimator we present below will not only correct for the measurement error in income in a returns to schooling regression, but will also deal with the top coding issue by using a censoring robust estimator.

Our maintained assumption is that social security earnings data are more accurate than the household reported earnings data (CPS). Hence earnings from the social security records (SSR) are treated as the auxiliary data. We eliminate observations where the respondent has been working for less than a year or is working in occupations that are likely to have unreported tips like bartenders, waiters or taxicab drivers. An issue that arises is that individuals that provided their social security number might be a selected group. We present a table similar to Table 2 of Bound and Krueger (1991) that compares the observable characteristic of the matched and the non-matched respondents. As in Bound and Krueger, we see in Table 1 that the observable characteristics of both respondents are similar (Bound and Krueger used more variables and found that these characteristics are similar also).

Another shortcoming of the SSR data is that earnings are capped at the social security maximum of \$16,500. This maps into a censoring level of 40% for males and about 4% for

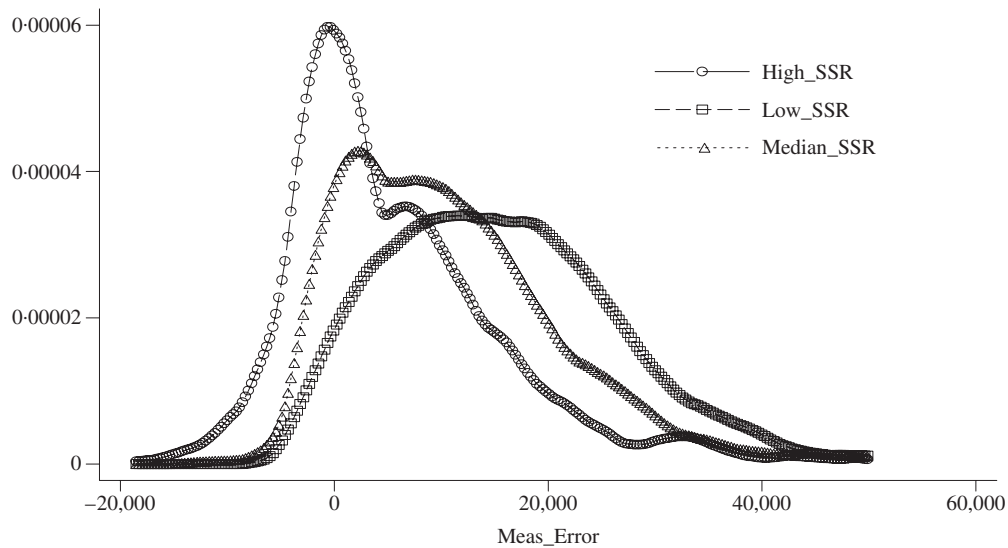


FIGURE 1
Measurement error density by income quantile

females. To reduce the effect of sample selection that would arise using the female sample,⁴ we use the male sample and restrict it to individuals between the ages of 40 and 50 (for computational simplicity) and use censored least absolute deviation estimators to deal with censoring. The number of observations in the primary sample is $n_p = 7362$ and in the validation sample is $n_v = 4809$.

We provide further evidence against the classical measurement error model in Figure 1. There, we divide the data into the lowest quartile (based on SSR) and call that Low_SSR in the figure, the 25-th–75-th range (Median_SSR in the figure) and the top quartile (High_SSR) and graph the density of the measurement error (CPS–SSR) for matched individuals whose SSR income is *below* the topcode of \$16,500. The median SSR income in the Low_SSR group is \$4578, the median SSR income in the Median_SSR group is \$10,012 and the median in the High_SSR group is \$15,316. As one can see, the error densities are different for different income levels.⁵ In particular, low income individuals (based on their SSR income) tend to overreport their income.

A Mincer regression studies the log income level $\log Y_p^*$ as a function of Z_p , a set of regressors that include education and experience. We use a quantile based censored least absolute deviation model (CLAD) where we correct for measurement error of unknown form using methods developed in this paper. Moreover, by restricting ourselves to older males, we minimize the sample selection problem. To map the problem back into our framework which is based on method of moments (rather than M-estimation) one can take the first order conditions of the CLAD objective function. However, it is well known that the moment conditions based on the first order conditions are set equal to zero not only by the true parameter value, but also by other parameters (like zero) which creates problems of local minima when using method of moments as a basis for inference. To remedy this, we base our inference on the objective function and

4. For a quantile approach to studying female wages that also handles selection, see Buchinsky (1998).

5. The measurement error (SSR–CPS) can be as small as $-\$45,000$. The reason for this is that, for individuals whose SSR income is less than the topcode, their CPS income (or reported income) can be high. The median SSR income below the topcode is \$10,034 while the median CPS income for individuals whose SSR is below the topcode is \$19,000.

adapt our approach above to it. This objective function, if we can observe the true income level in the primary data-set, is given by

$$\min_{\beta} E_p |\log Y_p^* - \min(Z'_p \beta, c)|,$$

where $c = \log(16,500)$ is the fictitious level of censoring for the primary data $\log Y_p^*$ (which is not observed). Given that we only observe $\log Y^*$ in the validation data, and assuming that the regressors ($Z =$ education, experience, experience², and race) are measured without error, we use the modified objective function that we obtain by projecting the unobserved objective function first onto the primary data $\mathbf{X}_p = (\log Y_p, \text{edu}, \text{exp}, \text{race})$

$$\min_{\beta} E_p E_p(|\log Y_p^* - \min(Z'_p \beta, c)| | \mathbf{X}_p) = \min_{\beta} E_p E_v(|\log Y_v^* - \min(Z'_v \beta, c)| | \mathbf{X}_v = \mathbf{X}_p)$$

where the equality follows from our main Assumption 1 above. We then minimize the following sample analogue of the above objective function:

$$\begin{aligned} \min_{\beta} \frac{1}{n_p} \sum_{i=1}^{n_p} \hat{E}_v(|\log Y_v^* - \min(Z'_v \beta, c)| | X_v = X_{pi}) \\ = \frac{1}{n_p} \sum_{i=1}^{n_p} \left(\sum_{j=1}^{n_v} |\log Y_{vj}^* - \min(Z'_{vj} \beta, c)| p^{k_{nv}}(X_{vj})' (P'_v P_v)^{-1} p^{k_{nv}}(X_{pi}) \right) \end{aligned} \quad (11)$$

where $p^{k_{nv}}(X_{vj})$ is a tensor product polynomial spline sieve in the empirical application. In particular, we used second order polynomial splines with K knots as the sieve basis $\{(X_l)^j, j = 0, 1, 2, \max(0, X_l - \tau_{l,k})^2, k = 1, \dots, K\}$ to approximate square integrable functions of X_l for $l = \log(Y_p), \text{educ}, \text{exp}$, where $\tau_{l,k}, k = 1, \dots, K$ are the equal range quantiles of the empirical distribution of X_l for $l = \log(Y_p), \text{educ}, \text{exp}$. We interact the race dummy with all the polynomial spline base functions. In the application we have tried with a number of knots $K = 3, 4, 5$, and the resulting estimates of β_o do not change much. The estimation result reported in Table 2 corresponds to $K = 4$. Although there are theoretical results on generalized cross-validation (GCV) procedure when deciding on the number of sieve terms for purely non-parametric estimation of a conditional mean function, in our application the parameter of interest is the finite dimensional parameter β_o , not the conditional mean function $g(\cdot, \beta) = E_p(|\log Y_p^* - \min(Z'_p \beta, c)| | X_p = \cdot)$. Theoretical results on GCV in semiparametric models is a current area of research. Nevertheless, there is much empirical and Monte Carlo evidence showing that for inference about a finite dimensional parameter β neither the choice of basis functions (splines, wavelets, Fourier series) nor the choice of the number of sieve terms are important;⁶ see for example Newey (1994), Ai and Chen (2003), and Blundell, Chen and Kristensen (2003).

The results of our regression are summarized in Table 2. The first column provides estimates using the censored LAD model on the auxiliary data. The estimator here is the median censored regression estimator of Powell (1984) which is consistent in the absence of stratification. The second column provides LAD estimates from the primary sample. Here the dependent variable is the log of *reported* income and is in general inconsistent in the presence of measurement error. The third column reports estimates that are obtained using our estimator in (11). The standard errors for the third column estimator were obtained using formulas derived in Appendix C. As we can see, the returns to schooling coefficient is stable around 5.5% for the first and third estimator with (11) having the highest standard error which is to be expected. The returns to schooling using LAD (second column) on the primary data is higher (at almost 7%). It is interesting to note that the returns to experience is negative and insignificant for both the auxiliary estimator

6. One safe rule is that spline sieve is always better than power series. Also, in the semiparametric sieve set-up one can still use the same number of sieve terms as suggested by GCV in the purely non-parametric regressions.

TABLE 2
CLAD estimates of β

	CLAD		
	Auxiliary Estimate St. error	Primary Estimate St. error	Ours Estimate St. error
Education	0.0551 (4.84e-03)	0.0691 (3.02e-03)	0.055 (5.9e-03)
Experience	-0.0156 (0.0178)	0.0507 (0.011)	-0.019 (0.025)
Experience ²	3.8e-04 (3.78e-04)	-7.89e-04 (2.26e-04)	4.27e-04 (4.76e-04)
Race	0.17 (0.0299)	0.136 (0.023)	0.18 (0.0368)
Constant	8.75 (0.153)	8.18 (0.16)	8.81 (0.321)

and our estimator, while the estimator using the primary data is positive and significant. Under our assumptions, this estimator is inconsistent. We suspect that the reason for the insignificance of the experience coefficient is that the effect of experience in the population of males above 40 years of age might not be as important.

5. CONCLUSION

We study the problem of parameter estimation in models defined in terms of moment conditions when the data are measured with error. Our paper is partially motivated by the distrust of the classical measurement error assumption that requires independence of the measurement error and the true variable. To allow for arbitrary correlation between the true variable and the measurement error, we use auxiliary data that identify the conditional distribution of the true variables given the mismeasured variables. This assumption is satisfied if the auxiliary data-set is a validated stratified subsample of the primary data-set, or if it is a validation data-set that is drawn from the population. Under the identification conditions, we provide a semiparametric sieve based estimator that exploits the information in both data-sets and show that this estimator is consistent and asymptotically normal.

In survey data where the presence of measurement error can bias estimates of parameters of interest, it is sometimes questionable to assume the classical errors in variables model. For a general and possibly non-linear model, this paper shows that collecting auxiliary data remedies the identification problem while allowing for arbitrary correlation between the measurement error and the true variables. The auxiliary data need to satisfy our main Assumption 1. One might obtain validation data by randomly validating a subsample of the primary data, or one can also collect a stratified sample based on the primary data, and validate the corresponding variables. We illustrate our estimator using data from the CPS/SSR exact match data where we show how our estimator can be used in the presence of censoring.

APPENDIX A

Preliminaries

We first provide a few definitions. A typical smoothness assumption is that a function belongs to a Hölder space. For any $1 \times d$ vector $\mathbf{a} = (a_1, \dots, a_d)$ of non-negative integers, we write $|\mathbf{a}| = \sum_{k=1}^d a_k$, and for any $x = (x_1, \dots, x_d)' \in \mathcal{X} \subseteq$

\mathcal{R}^d , we denote the $|\mathbf{a}|$ -th derivative of a function $h : \mathcal{X} \rightarrow \mathcal{R}$ as

$$\nabla^{\mathbf{a}} h(x) = \frac{\partial^{|\mathbf{a}|}}{\partial x_1^{a_1} \dots \partial x_d^{a_d}} h(x).$$

For some $\gamma > 0$, let $\underline{\gamma}$ be the largest integer smaller than γ , and let $\Lambda^\gamma(\mathcal{X})$ denote a Hölder space with smoothness γ , i.e. a space of functions $h : \mathcal{X} \rightarrow \mathcal{R}$ which have up to $\underline{\gamma}$ -th continuous derivatives, and the highest ($\underline{\gamma}$ -th) derivatives are Hölder continuous with the Hölder exponent $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$\|h\|_{\Lambda^\gamma} = \sup_{\mathcal{X}} |h(x)| + \max_{|\mathbf{a}|=\underline{\gamma}} \sup_{x \neq \bar{x}} \frac{|\nabla^{\mathbf{a}} h(x) - \nabla^{\mathbf{a}} h(\bar{x})|}{\sqrt{(x - \bar{x})'(x - \bar{x})}^{\gamma - \underline{\gamma}}} < \infty.$$

Let $\Lambda^\gamma(\mathcal{X}, \omega_1)$ denote a weighted Hölder space of functions $h : \mathcal{X} \rightarrow \mathcal{R}$ such that $h(\cdot)[1 + |\cdot|^2]^{-\omega_1/2}$ is in $\Lambda^\gamma(\mathcal{X})$. We call $\Lambda_c^\gamma(\mathcal{X}, \omega_1) \equiv \{h \in \Lambda^\gamma(\mathcal{X}, \omega_1) : \|h(\cdot)[1 + |\cdot|^2]^{-\omega_1/2}\|_{\Lambda^\gamma} \leq c < \infty\}$ a weighted Hölder ball (with radius c).

Definition 1. For all β , a function $g(\cdot; \beta)$ is $H(\gamma, \omega_1)$ -smooth if it belongs to a weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ for some $\gamma > 0$ and $\omega_1 \geq 0$.

The weighted Hölder ball with $\omega_1 = 0$ reduces to the standard Hölder ball $\Lambda_c^\gamma(\mathcal{X})$ condition, which is a typical sufficient condition especially when the support \mathcal{X} is a bounded subset of \mathcal{R}^d . However, when $\mathcal{X} = \mathcal{R}^d$, the standard Hölder ball $\Lambda_c^\gamma(\mathcal{X})$ may exclude some moment functions such as $g(x, \beta) = x'(\beta_o - \beta)$. It is clear that the weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ with $\omega_1 > 0$ is a strictly larger space and $x'(\beta_o - \beta) \in \Lambda_c^\gamma(\mathcal{R}^d, \omega_1)$ with $\omega_1 = 1$.

Denote $\mathcal{L}_{2,v}(\mathcal{X}) = \{h : \mathcal{X} \rightarrow \mathcal{R} : \|h\|_{2,v} = \sqrt{\int h(x)^2 f_{X_v}(x) dx} < \infty\}$ and $\mathcal{L}_{2,p}(\mathcal{X}) = \{h : \mathcal{X} \rightarrow \mathcal{R} : \|h\|_{2,p} = \sqrt{\int h(x)^2 f_{X_p}(x) dx} < \infty\}$ as the two Hilbert spaces. Recall that

$$\|h\|_{\infty,\omega} = \sup_{x \in \mathcal{X}, \beta \in B} |h(x, \beta)[1 + |x|^2]^{-\omega/2}| = o_p(1).$$

Proposition A1. Under Assumptions 1–3, and $\frac{k_{nv}}{n_v} \rightarrow 0, k_{nv} \rightarrow \infty, n_p \rightarrow \infty$, we have (i)

$$\begin{aligned} \|\hat{g}(\bullet, \bullet) - g(\bullet, \bullet)\|_{\infty,\omega} &= o_p(1); \\ \sup_{\beta \in B} \|\hat{g}(\bullet, \beta) - g(\bullet, \beta)\|_{2,p} &= o_p(1), \quad \sup_{\beta \in B} \|\hat{g}(\bullet, \beta) - g(\bullet, \beta)\|_{2,v} = o_p(1); \end{aligned}$$

(ii) in addition, if Assumption 5(2) holds, then

$$\|\hat{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\|_{2,v} = O_p\left(\sqrt{\frac{k_{nv}}{n_v}} + (k_{nv})^{-\gamma/d}\right).$$

Proof (Proposition A1). (i) Recall that $\hat{g}(x, \beta)$ is the sieve least squares estimator of $g(x, \beta) = E[m(X^*, \beta) | X = x] \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$ based on the auxiliary sample. That is, $\hat{g}(x, \beta)$ solves

$$\inf_{\tilde{g} \in \mathcal{G}_n} \frac{1}{n_v} \sum_{j=1}^{n_v} [m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2$$

where \mathcal{G}_n increases with auxiliary sample size n_v , and is dense in $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ as $k_{nv} \rightarrow \infty$. Moreover, by Assumptions 2(1) and 3 we have the following results: (1) the parameter space is compact under the norm $\|\cdot\|_{\infty,\omega}$ for $\omega > \omega_1 \geq 0$, see Chen *et al.* (1997) or Ai and Chen (2003); (2) $E_v\{[m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2\}$ is uniquely minimized at $g(x, \beta) = E[m(X^*, \beta) | X = x] \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$; (3) $E_v\{[m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2\}$ is continuous in \tilde{g} under the metric $\|\cdot\|_{\infty,\omega}$. This is due to the fact that because

$$E_v\{[m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2\} = E_v\{[m(X_{vj}^*, \beta) - g(X_{vj}, \beta)]^2\} + E_v\{[g(X_{vj}, \beta) - \tilde{g}(X_{vj}, \beta)]^2\},$$

we can write, for any $\tilde{g}, \bar{g} \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$, and for $\omega > \omega_1 \geq 0$,

$$\begin{aligned} & E_v\{[m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2\} - E_v\{[m(X_{vj}^*, \beta) - \bar{g}(X_{vj}, \beta)]^2\} \\ &= E_v\{[g(X_{vj}, \beta) - \tilde{g}(X_{vj}, \beta)]^2\} - E_v\{[g(X_{vj}, \beta) - \bar{g}(X_{vj}, \beta)]^2\} \\ &= E_v\{[\tilde{g}(X_{vj}, \beta) - \bar{g}(X_{vj}, \beta)][2g(X_{vj}, \beta) - (\tilde{g}(X_{vj}, \beta) + \bar{g}(X_{vj}, \beta))]\} \\ &\leq \text{const.} \times E_v\{(1 + X'_{vj}X_{vj})^{(\omega_1 + \omega)/2} [|\tilde{g}(X_{vj}, \beta) - \bar{g}(X_{vj}, \beta)| (1 + X'_{vj}X_{vj})^{-\omega/2}]\} \\ &\leq \text{const.} \sqrt{\int (1 + x'x)^\omega f_{X_v}(x) dx} \times \|\tilde{g} - \bar{g}\|_{\infty, \omega}, \end{aligned}$$

and (4)

$$\sup_{\tilde{g}} \left| \frac{1}{n_v} \sum_{j=1}^{n_v} [m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2 - E_v\{[m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2\} \right| = o_p(1),$$

which is because, for any $\tilde{g}, \bar{g} \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$,

$$\begin{aligned} & |[m(X_{vj}^*, \beta) - \tilde{g}(X_{vj}, \beta)]^2 - [m(X_{vj}^*, \beta) - \bar{g}(X_{vj}, \beta)]^2| \\ &\leq |\tilde{g}(X_{vj}, \beta) - \bar{g}(X_{vj}, \beta)| \times |2m(X_{vj}^*, \beta) - (\tilde{g}(X_{vj}, \beta) + \bar{g}(X_{vj}, \beta))| \\ &\leq |\tilde{g}(X_{vj}, \beta) - \bar{g}(X_{vj}, \beta)| \times \{|2g(X_{vj}, \beta) - (\tilde{g}(X_{vj}, \beta) + \bar{g}(X_{vj}, \beta))| + 2|m(X_{vj}^*, \beta) - g(X_{vj}, \beta)|\} \\ &\leq \|\tilde{g} - \bar{g}\|_{\infty, \omega} \times M(X_{vj}^*, X_{vj}), \end{aligned}$$

with $E[M(X_{vj}^*, X_{vj})] < \infty$ where

$$M(X_{vj}^*, X_{vj}) \equiv (1 + X'_{vj}X_{vj})^{\omega/2} \times \{\text{const.}(1 + X'_{vj}X_{vj})^{\omega_1/2} + 2 \sup_{\beta} |m(X_{vj}^*, \beta) - g(X_{vj}, \beta)|\}.$$

Hence by Lemma 2.9 and Theorem 2.1 in Newey and McFadden (1994), $\|\hat{g}(\bullet, \bullet) - g(\bullet, \bullet)\|_{\infty, \omega} = o_p(1)$. Now

$$\begin{aligned} \sup_{\beta \in B} \|\hat{g}(\bullet, \beta) - g(\bullet, \beta)\|_{2, p} &= \sup_{\beta \in B} \sqrt{\int [\hat{g}(x, \beta) - g(x, \beta)]^2 f_{X_p}(x) dx} \\ &\leq \sqrt{(\|\hat{g}(\bullet, \bullet) - g(\bullet, \bullet)\|_{\infty, \omega})^2 \int (1 + x'x)^\omega f_{X_p}(x) dx} \\ &\rightarrow 0 \text{ as } n_v \rightarrow \infty \quad (\text{by Assumption 3(2)}), \end{aligned}$$

and similarly

$$\sup_{\beta \in B} \|\hat{g}(\bullet, \beta) - g(\bullet, \beta)\|_{2, v} = o_p(1).$$

(ii) We can obtain the convergence rate of $\|\hat{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\|_{2, v}$ by simple application of Theorem 1 in Chen and Shen (1998). All the assumptions of the Chen–Shen Theorem 1 are satisfied given our Assumptions 2(1) and 3. We obtain

$$\|\hat{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\|_{2, v} = O_p \left(\max \left\{ \sqrt{\frac{k_{nv}}{n_v}}, \|g(\bullet, \beta_o) - \Pi_{2n} g(\bullet, \beta_o)\|_{2, v} \right\} \right).$$

Under Assumption 3(1), for $g(\bullet, \beta_o) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$, there exists $\Pi_{\infty n} g(\bullet, \beta_o) = \pi(\beta_o)' p^{k_{nv}}(\bullet) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$, such that for any fixed $\omega > \gamma + \omega_1$

$$\|g(\bullet, \beta_o) - \Pi_{\infty n} g(\bullet, \beta_o)\|_{\infty, \omega} = \sup_x |g(x, \beta_o) - \Pi_{\infty n} g(x, \beta_o)| (1 + |x|^2)^{-\omega/2} \leq c(k_{nv})^{-\gamma/d},$$

see Chen *et al.* (1997) or Ai and Chen (2003). Hence by Assumption 5(2) with $\omega = \gamma + \omega_1 + \epsilon$ for a small $\epsilon > 0$,

$$\begin{aligned} \|g(\bullet, \beta_o) - \Pi_{2n} g(\bullet, \beta_o)\|_{2, v} &\leq \|g(\bullet, \beta_o) - \Pi_{\infty n} g(\bullet, \beta_o)\|_{2, v} = \sqrt{\int [g(x, \beta_o) - \Pi_{\infty n} g(x, \beta_o)]^2 f_{X_v}(x) dx} \\ &\leq \sqrt{(\|g(\bullet, \beta_o) - \Pi_{\infty n} g(\bullet, \beta_o)\|_{\infty, \omega})^2 \int (1 + x'x)^\omega f_{X_v}(x) dx} \leq c'(k_{nv})^{-\gamma/d}. \end{aligned}$$

Then

$$\|\hat{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\|_{2, v} = O_p \left(\sqrt{\frac{k_{nv}}{n_v}} + (k_{nv})^{-\gamma/d} \right) = o_p(1). \quad \parallel$$

Proposition A1 establishes that under Assumptions 3(1) and 3(5)

$$\|\hat{g} - g\|_{\infty, \omega} = \sup_{x \in \mathcal{X}, \beta \in B} |\{\hat{g}(x, \beta) - g(x, \beta)\}[1 + |x|^2]^{-\omega/2}| = o_p(1).$$

This and Assumption 3(2) imply that $\sup_{\beta \in B} \int \{\hat{g}(x, \beta) - g(x, \beta)\}^2 f_{X_v}(x) dx = o_p(1)$. We note that the consistency in the weighted sup-norm metric $\|\hat{g} - g\|_{\infty, \omega} = o_p(1)$ is a strictly weaker notion of consistency than the standard sup-norm metric $\|\hat{g} - g\|_{\infty} = \sup_{x \in \mathcal{X}, \beta \in B} |\hat{g}(x, \beta) - g(x, \beta)| = o_p(1)$.

Proof (Theorem 1). Given $\|\hat{g}(\bullet, \bullet) - g(\bullet, \bullet)\|_{\infty, \omega} = o_p(1)$ (Proposition A1(i)), Assumptions 2 and 3(1)–3(3), all conditions of Lemma 5.2 in Newey (1994) are satisfied, hence $\hat{\beta} - \beta_o = o_p(1)$. \square

Proof (Theorem 2). By first order condition and mean-value expansion, we obtain

$$\begin{aligned} & \left\{ \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \frac{\partial \hat{g}(X_{pi}, \hat{\beta})}{\partial \beta'} \right)' \widehat{W} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \frac{\partial \hat{g}(X_{pi}, \bar{\beta})}{\partial \beta'} \right) \right\} (\hat{\beta} - \beta_o) \\ &= - \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \frac{\partial \hat{g}(X_{pi}, \hat{\beta})}{\partial \beta'} \right)' \widehat{W} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{g}(X_{pi}, \beta_o) \right) \end{aligned}$$

where, for $l = 1, \dots, q$, $\bar{\beta}_l$ is a convex combination of $\hat{\beta}_l$ and β_{ol} . In the following we shall establish

$$\sup_{|\beta - \beta_o| = o(1)} \left| \frac{1}{n_p} \sum_{i=1}^{n_p} \frac{\partial \hat{g}(X_{pi}, \beta)}{\partial \beta'} - E_p \left[\frac{\partial g(X_{pi}, \beta_o)}{\partial \beta'} \right] \right| = o_p(1) \quad (\text{A.1})$$

$$\begin{aligned} & \sqrt{n_v} \frac{1}{n_p} \sum_{i=1}^{n_p} \hat{g}(X_{pi}, \beta_o) \\ &= \frac{1}{\sqrt{n_v}} \sum_{j=1}^{n_v} \frac{f_{X_p}(X_{vj})}{f_{X_v}(X_{vj})} \{m(X_{vj}^*, \beta_o) - g(X_{vj}, \beta_o)\} + \frac{\sqrt{n_v}}{n_p} \sum_{i=1}^{n_p} g(X_{pi}, \beta_o) + o_p(1). \end{aligned} \quad (\text{A.2})$$

For (A.1), it suffices to show

$$\sup_{|\beta - \beta_o| = o(1), \|\tilde{g} - g\|_{\infty, \omega} = o(1)} \left| \frac{1}{n_p} \sum_{i=1}^{n_p} \frac{\partial \tilde{g}(X_{pi}, \beta)}{\partial \beta'} - E_p \left[\frac{\partial g(X_{pi}, \beta)}{\partial \beta'} \right] \right| = o_p(1) \quad (\text{A.1.1})$$

and

$$\sup_{|\beta - \beta_o| = o(1)} \left| E_p \left[\frac{\partial g(X_{pi}, \beta)}{\partial \beta'} \right] - E_p \left[\frac{\partial g(X_{pi}, \beta_o)}{\partial \beta'} \right] \right| = o(1). \quad (\text{A.1.2})$$

Equation (A.1.1) is implied by Assumptions 2(1), 3(2) and 4(4), and (A.1.2) is implied by Assumption 4(3).

For (A.2), we notice

$$\begin{aligned} \frac{1}{n_p} \sum_{i=1}^{n_p} \hat{g}(X_{pi}, \beta_o) - E_p[g(X_{pi}, \beta_o)] &= \int [\hat{g}(x, \beta_o) - g(x, \beta_o)] f_{X_p}(x) dx + \int g(x, \beta_o) d[\widehat{F}_{X_p}(x) - F_{X_p}(x)] \\ &\quad + \int [\hat{g}(x, \beta_o) - g(x, \beta_o)] d[\widehat{F}_{X_p}(x) - F_{X_p}(x)]. \end{aligned}$$

By Assumption 2(1) and Assumption 4(2),

$$\sqrt{n_p} \int g(x, \beta_o) d[\widehat{F}_{X_p}(x) - F_{X_p}(x)] = \frac{1}{\sqrt{n_p}} \sum_{j=1}^{n_p} g(X_{pi}, \beta_o) \Rightarrow \mathcal{N}(0, E_p[g(x, \beta_o)g(x, \beta_o)']).$$

In the following we let $N_{[\square]}(\varepsilon, \Lambda_c^\gamma(\mathcal{X}, \omega_1), \|\cdot\|_{2,p})$ denote the $\|\cdot\|_{2,p}$ -covering number of $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ with bracketing (i.e. the minimal number of N for which there exist ε -brackets $\{[l_j, u_j] : \|l_j - u_j\|_{2,p} \leq \varepsilon, \|l_j\|_{2,p}, \|u_j\|_{2,p} < \infty, j = 1, \dots, N\}$ to cover $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$). We also let $N(\varepsilon, \Lambda_c^\gamma(\mathcal{X}, \omega_1), \|\cdot\|_{\infty, \omega})$ denote the $\|\cdot\|_{\infty, \omega}$ -covering number of $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ (i.e. the minimal number of N for which there exist ε -balls $\{h : \|h - u_j\|_{\infty, \omega} \leq \varepsilon, j = 1, \dots, N$ to cover $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$). By Assumptions 2(1) and 3(2) with $\omega > \omega_1 + \gamma$, we have

$$\log N_{[\square]}(\delta, \Lambda_c^\gamma(\mathcal{X}, \omega_1), \|\cdot\|_{2,p}) \leq \log N(\delta, \Lambda_c^\gamma(\mathcal{X}, \omega_1), \|\cdot\|_{\infty, \omega}) \leq \text{const.} \left(\frac{c}{\delta}\right)^{d/\gamma}, \quad (12)$$

see Chen *et al.* (1997) and Blundell *et al.* (2003). Thus under Assumption 5(2) ($\gamma > d/2$),

$$\int_0^1 \sqrt{\log N_{[\square]}(\delta, \Lambda_c^\gamma(\mathcal{X}, \omega_1), \|\cdot\|_{2,p})} d\delta < \infty,$$

and the class $\{\tilde{g}(\bullet, \beta_o) : \tilde{g}(\bullet, \beta_o) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)\}$ is a F_{X_p} -Donsker class. This and the fact that $\hat{g}(x, \beta_o) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$, $\gamma > d/2$ with probability approaching one as $n_v \rightarrow \infty$ imply that

$$\sup_{\tilde{g}(\bullet, \beta_o) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1) : \|\tilde{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\|_{2,p} = o(1)} \left| \int [\tilde{g}(x, \beta_o) - g(x, \beta_o)] d[\widehat{F}_{X_p}(x) - F_{X_p}(x)] \right| = o_p\left(\frac{1}{\sqrt{n_p}}\right),$$

this together with Assumption 2(2) implies

$$\int [\hat{g}(x, \beta_o) - g(x, \beta_o)] d[\widehat{F}_{X_p}(x) - F_{X_p}(x)] = o_p\left(\frac{\sqrt{\lambda}}{\sqrt{n_v}}\right) = o_p\left(\frac{1}{\sqrt{n_v}}\right).$$

Therefore (A.2) will be established after we obtain (A.2.0):

$$\begin{aligned} \sqrt{n_v} \int [\hat{g}(x, \beta_o) - g(x, \beta_o)] f_{X_p}(x) dx &= \sqrt{n_v} \int [\hat{g}(x, \beta_o) - g(x, \beta_o)] v^*(x) f_{X_v}(x) dx \\ &= \frac{1}{\sqrt{n_v}} \sum_{j=1}^{n_v} v^*(X_{vj}) [m(X_{vj}^*, \beta_o) - g(X_{vj}, \beta_o)] + o_p(1) \Rightarrow \mathcal{N}(0, \Omega_1) \end{aligned} \quad (\text{A.2.0})$$

with

$$v^*(\bullet) \equiv \frac{f_{X_p}(\bullet)}{f_{X_v}(\bullet)} \in \mathcal{L}_{2,v}(\mathcal{X}) \quad \text{by Assumption 5(1).}$$

We now apply the approach taken in Chen and Shen (1998) and Ai and Chen (2003) to show (A.2.0). Recall that $\hat{g}(x, \beta_o)$ is the sieve LS estimator of $g(x, \beta_o) = E[m(X^*, \beta_o) | X] \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$ based on the validation sample, and that $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ is dense in the Hilbert space $\mathcal{L}_{2,v}(\mathcal{X})$. For any $\tilde{g}(\bullet, \beta_o) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)$, we have

$$\int [\tilde{g}(x, \beta_o) - g(x, \beta_o)] f_{X_p}(x) dx = \int [\tilde{g}(x, \beta_o) - g(x, \beta_o)] v^*(x) f_{X_v}(x) dx.$$

In the following we denote $Z_j = (X_{vj}^{*t}, X_{vj}^{*l})'$ and $U_{vj} = m(X_{vj}^*, \beta_o) - g(X_{vj}, \beta_o)$. By definition $E[U_{vj} | X_{vj}] = 0$. Let $L_n(\tilde{g}) = \frac{1}{n_v} \sum_{j=1}^{n_v} \ell(Z_j, \tilde{g})$ with $\ell(Z_j, \tilde{g}) = -\frac{1}{2} [m(X_{vj}^*, \beta_o) - \tilde{g}(X_{vj}, \beta_o)]^2$. We also denote $\mu_n(h) = \frac{1}{n_v} \sum_{j=1}^{n_v} [h(Z_j) - E_v(h(Z_j))]$ as the empirical process indexed by h , and let ε_n be any positive sequence with $\varepsilon_n = o(\frac{1}{\sqrt{n_v}})$. Then by definition,

$$\begin{aligned} 0 &\leq L_n(\hat{g}) - L_n(\hat{g} \pm \varepsilon_n \Pi_{2n} v^*) \\ &= \mu_n(\ell(Z_j, \hat{g}) - \ell(Z_j, \hat{g} \pm \varepsilon_n \Pi_{2n} v^*)) + E_v(\ell(Z_j, \hat{g}) - \ell(Z_j, \hat{g} \pm \varepsilon_n \Pi_{2n} v^*)). \end{aligned}$$

Simple calculation yields

$$\begin{aligned} E_v(\ell(Z_j, \hat{g}) - \ell(Z_j, \hat{g} \pm \varepsilon_n \Pi_{2n} v^*)) &= \pm \varepsilon_n E_v[\Pi_{2n} v^*(X_{vj}) \{\hat{g}(X_{vj}, \beta_o) - g(X_{vj}, \beta_o)\}] \\ &\quad + \frac{1}{2} \varepsilon_n^2 E_v[\{\Pi_{2n} v^*(X_{vj})\}^2] \\ \mu_n(\ell(Z_j, \hat{g}) - \ell(Z_j, \hat{g} \pm \varepsilon_n \Pi_{2n} v^*)) &= \mp \varepsilon_n \times \mu_n(\Pi_{2n} v^* U_{vj}) \pm \varepsilon_n \\ &\quad \times \mu_n\left(\Pi_{2n} v^* \frac{2\{\hat{g}(\cdot, \beta_o) - g(\cdot, \beta_o)\} \pm \varepsilon_n \Pi_{2n} v^*}{2}\right) \end{aligned}$$

hence

$$\begin{aligned} 0 &\leq \mp \mu_n(\Pi_{2n} v^*(X_{vj}) U_{vj}) \pm E_v[\Pi_{2n} v^*(X_{vj}) \{\hat{g}(X_{vj}, \beta_o) - g(X_{vj}, \beta_o)\}] \\ &\quad \pm \mu_n(\Pi_{2n} v^*(X_{vj}) \{\hat{g}(X_{vj}, \beta_o) - g(X_{vj}, \beta_o)\}) + \frac{\varepsilon_n}{2n_v} \sum_{j=1}^{n_v} \{\Pi_{2n} v^*(X_{vj})\}^2 \\ &= \mp \mu_n([\Pi_{2n} v^* - v^*] U_{vj}) \pm \mu_n(v^* U_{vj}) \pm E_v[[\Pi_{2n} v^* - v^*] \{\hat{g} - g\}] \mp E_v[v^* \{\hat{g} - g\}] \\ &\quad \pm \mu_n(\Pi_{2n} v^*(X_{vj}) \{\hat{g}(X_{vj}, \beta_o) - g(X_{vj}, \beta_o)\}) + \frac{\varepsilon_n}{2n_v} \sum_{j=1}^{n_v} \{\Pi_{2n} v^*(X_{vj})\}^2. \end{aligned}$$

Then we obtain (A.2.0) after we establish the following (A.2.1)–(A.2.4):

$$\mu_n([\Pi_{2n}v^*(X_{vj}) - v^*(X_{vj})]U_{vj}) = o_p\left(\frac{1}{\sqrt{n_v}}\right) \quad (\text{A.2.1})$$

$$E_v([\Pi_{2n}v^*(X_{vj}) - v^*(X_{vj})]\{\hat{g}(X_{vj}, \beta_o) - g(X_{vj}, \beta_o)\}) = o_p\left(\frac{1}{\sqrt{n_v}}\right) \quad (\text{A.2.2})$$

$$\mu_n(\Pi_{2n}v^*\{\hat{g}(X_{vj}, \beta_o) - g(X_{vj}, \beta_o)\}) = o_p\left(\frac{1}{\sqrt{n_v}}\right) \quad (\text{A.2.3})$$

$$\frac{1}{n_v} \sum_{j=1}^{n_v} \{\Pi_{2n}v^*(X_{vj})\}^2 = O_p(1). \quad (\text{A.2.4})$$

Now (A.2.1) is implied by Chebychev inequality, Assumptions 2(1), 3(4) and 5(4). (A.2.2) is implied by Assumption 5(4) and $\|\hat{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\|_{2,v} = O_p\left((n_v)^{-\frac{\gamma}{2\gamma+d}}\right)$ from Proposition A1(ii). (A.2.4) is implied by Markov inequality, Assumptions 2(1), 5(1) and 5(4). Finally, for (A.2.3), let $\mathcal{F}_n = \{\Pi_{2n}v^*(\bullet)\tilde{g}(\bullet, \beta_o) : \tilde{g}(\bullet, \beta_o) \in \Lambda_c^\gamma(\mathcal{X}, \omega_1)\}$, then by Assumptions 2(1) and 5, $\log N_{[]}(\delta, \mathcal{F}_n, \|\cdot\|_{2,v}) \leq \text{const} \cdot (\frac{\delta}{\epsilon})^{d/\gamma}$ for any $\delta > 0$. Applying Theorem 3 in Chen and Shen (1998) with $\delta_n = (n_v)^{-\gamma/(2\gamma+d)}$, we have

$$\sup_{\tilde{g} \in \mathcal{F}_n: \|\tilde{g} - g(\bullet, \beta_o)\|_{2,v} \leq \delta_n} |\sqrt{n_v} \mu_n(\Pi_{2n}v^*\{\tilde{g}(\bullet, \beta_o) - g(\bullet, \beta_o)\})| = O_p\left((n_v)^{-\frac{2\gamma-d}{2(2\gamma+d)}}\right) = o_p(1).$$

Hence we obtain (A.2.3) and hence (A.2.0). \parallel

Proof (Theorem 3). We have already established $\widehat{G} = G + o_p(1)$ in the proof of Theorem 2; see (A.1.1) above. It suffices to show that $\widehat{\Omega} = \Omega + o_p(1)$ for Ω expressed in (9). Given Proposition A1(i), Theorem 1, Assumptions 2(1), 3(1)–3(3) and 4(1), 4(2), we have

$$\frac{n_v}{n_p^2} \sum_{i=1}^{n_p} \hat{g}(X_{pi}, \hat{\beta}) \hat{g}(X_{pi}, \hat{\beta})' = \lambda E_p[g(X, \beta_o)g(X, \beta_o)'] + o_p(1).$$

It remains to show that

$$\frac{1}{n_v} \sum_{j=1}^{n_v} (\widehat{v}_{vj}^* \widehat{U}_{vj}) (\widehat{v}_{vj}^* \widehat{U}_{vj})' = E_v[(v_{vj}^* U_{vj})(v_{vj}^* U_{vj})'] + o_p(1).$$

Which, given Assumptions 2(1), 3(4), 5(1) and 5(4), will be implied by (A.3.1)–(A.3.3):

$$\frac{1}{n_v} \sum_{j=1}^{n_v} (\widehat{v}_{vj}^* - v_{vj}^*)^2 U_{vj} U_{vj}' = o_p(1) \quad (\text{A.3.1})$$

$$\frac{1}{n_v} \sum_{j=1}^{n_v} v_{vj}^{*2} (\widehat{U}_{vj} - U_{vj}) (\widehat{U}_{vj} - U_{vj})' = o_p(1) \quad (\text{A.3.2})$$

$$\frac{1}{n_v} \sum_{j=1}^{n_v} (v_{vj}^* U_{vj})(v_{vj}^* U_{vj})' - E_v[(v_{vj}^* U_{vj})(v_{vj}^* U_{vj})'] = o_p(1). \quad (\text{A.3.3})$$

For (A.3.1), given Assumptions 2(1), 3(4), 5(1) and 5(4), we have

$$\begin{aligned} E_v[(\widehat{v}_{vj}^* - v_{vj}^*)^2 U_{vj} U_{vj}'] &\leq E_v[(\widehat{v}_{vj}^* - v_{vj}^*)^2 E[U_{vj} U_{vj}' | X_{vj}]] \\ &\leq \text{const} \cdot E_v[(\widehat{v}_{vj}^* - v_{vj}^*)^2] = o_p(1). \end{aligned}$$

Hence (A.3.1) holds by Markov inequality. (A.3.3) is obviously true given Assumptions 2(1), 3(4) and 5(1). For (A.3.2), we note that

$$\widehat{U}_{vj} - U_{vj} = \{m(X_{vj}^*, \hat{\beta}) - m(X_{vj}^*, \beta_o)\} - \{\hat{g}(X_{vj}, \hat{\beta}) - g(X_{vj}, \hat{\beta})\} - \{g(X_{vj}, \hat{\beta}) - g(X_{vj}, \beta_o)\}.$$

By Assumptions 6(1) and 6(2), Proposition A1(i) and Theorem 1, we have

$$|v_{vj}^* \{\hat{g}(X_{vj}, \hat{\beta}) - g(X_{vj}, \beta_o)\}| \leq |v_{vj}^*| \{(1 + X_{vj}' X_{vj})^{\frac{\alpha}{2}} \|\hat{g} - g\|_{\infty, \omega} + b(X_{vj})|\hat{\beta} - \beta_o|\}$$

and

$$\frac{1}{n_v} \sum_{j=1}^{n_v} v_{vj}^{*2} (\hat{g}(X_{vj}, \hat{\beta}) - g(X_{vj}, \beta_o)) (\hat{g}(X_{vj}, \hat{\beta}) - g(X_{vj}, \beta_o))' = o_p(1).$$

Finally by Assumptions 2(1) and 6(3) and Theorem 1, we have

$$\frac{1}{n_v} \sum_{j=1}^{n_v} v_{vj}^{*2} (m(X_{vj}^*, \hat{\beta}) - m(X_{vj}^*, \beta_o)) (m(X_{vj}^*, \hat{\beta}) - m(X_{vj}^*, \beta_o))' = o_p(1),$$

thus (A.3.2) is satisfied. \parallel

APPENDIX B. VARIANCE FORMULAS FROM SECTION 3.2 ABOVE

It is obvious that Ω takes the form (9) when the auxiliary sample is independent of the primary. It turns out that (9) is also valid for the "validate in sample" case. Define $v_{vj}^* \equiv f_{X_p}(X_{vj})/f_{X_v}(X_{vj})$, $U_{vj} \equiv m(X_{vj}^*, \beta_o) - g(X_{vj}, \beta_o)$. Then for the "validate in sample" case

$$\begin{aligned} \Omega &= \Omega_1 + \lambda E_p [g(X_{pi}, \beta_o) g(X_{pi}, \beta_o)'] + 2\lambda E_v [v_{vj}^* U_{vj} g(X_{vj}, \beta_o)'] \\ &= \Omega_1 + \lambda E_p [g(X_{pi}, \beta_o) g(X_{pi}, \beta_o)'] \quad (\text{since } E_v [U_{vj} | X_{vj}] = 0). \end{aligned}$$

Let $g(X_{pi}) = g(X_{pi}, \beta_o)$ and $m(X_{pi}^*) = m(X_{pi}^*, \beta_o)$. For the stratification scheme suggested in Section 2.1, we have

$$\lambda = p \lim \frac{1}{n_p} \sum_{i=1}^{n_p} 1(U_{pi} \leq T(X_{pi})) = \int f_{X_p}(x) T(x) dx.$$

Then,

$$\begin{aligned} \Omega &= \text{Avar} \left(\frac{\sqrt{n_v}}{n_p} \sum_{i=1}^{n_p} g(X_{pi}) + \frac{1}{\sqrt{n_v}} \sum_{i=1}^{n_p} \frac{f_{X_p}(X_{pi})}{f_{X_v}(X_{pi})} (m(X_{pi}^*) - g(X_{pi})) 1(U_{pi} \leq T(X_{pi})) \right) \\ &= \text{Avar} \left(\frac{1}{\sqrt{n_p}} \sum_{i=1}^{n_p} \left\{ \sqrt{\lambda} g(X_{pi}) + \frac{1}{\sqrt{\lambda}} \frac{f_{X_p}(X_{pi})}{f_{X_v}(X_{pi})} (m(X_{pi}^*) - g(X_{pi})) 1(U_{pi} \leq T(X_{pi})) \right\} \right). \end{aligned}$$

We can write

$$\begin{aligned} \Omega &= \lambda E g(X_{pi})^2 + \frac{1}{\lambda} E \frac{f_{X_p}(X_{pi})^2}{f_{X_v}(X_{pi})^2} [m(X_{pi}^*) - g(X_{pi})]^2 1(U_{pi} \leq T(X_{pi})) \\ &\quad + 2E g(X_{pi}) \frac{f_{X_p}(X_{pi})}{f_{X_v}(X_{pi})} [m(X_{pi}^*) - g(X_{pi})] 1(U_{pi} \leq T(X_{pi})). \end{aligned}$$

The third term drops out because of the definition $g(X_{pi}) = E(m(X_{pi}^*) | X_{pi})$, and the second term can be written as

$$2 \frac{1}{\lambda} E \frac{f_{X_p}(X_{pi})^2}{f_{X_v}(X_{pi})^2} [m(X_{pi}^*) - g(X_{pi})]^2 T(X_{pi}) = 2E \frac{\lambda}{T(X_{pi})} [m(X_{pi}^*) - g(X_{pi})]^2.$$

This is consistent with the earlier definition of Ω_1 since Ω_1 can be written as

$$\begin{aligned} E_v \left[\left(\frac{\lambda}{T(X)} \right)^2 \{m(X^*, \beta_o) - g(X, \beta_o)\}^2 \right] &= \int \left[\left(\frac{\lambda}{T(X)} \right)^2 \{m(X^*, \beta_o) - g(X, \beta_o)\}^2 \right] f_{X_p}(X) \frac{T(x)}{\lambda} dX \\ &= \int \left(\frac{\lambda}{T(X)} \right) \{m(X^*, \beta_o) - g(X, \beta_o)\}^2 f_{X_p}(X) dX. \end{aligned}$$

APPENDIX C. STANDARD ERRORS FOR THE CLAD ESTIMATOR

First, to simplify notation, we let $W = p^{k_{nv}}(X)$. Then $\hat{\beta}$ minimizes⁷

$$\begin{aligned} &\frac{1}{n_p} \sum_{i=1}^{n_p} \left(\sum_{j=1}^{n_v} \left| \log Y_{vj}^* - \min(Z'_{vj} \beta, c) \right| W'_{vj} (W'_v W_v)^{-1} W_{pi} \right) \\ &= \left(\sum_{j=1}^{n_v} \left| \log Y_{vj}^* - \min(Z'_{vj} \beta, c) \right| W'_{vj} \right) (W'_v W_v)^{-1} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} W_{pi} \right). \end{aligned}$$

7. Here, we just provide a sketch of the asymptotic variance calculations without providing the regularity conditions. For similar conditions, see Powell (1984).

The approximate first order condition for $\hat{\beta}$ is

$$\left(\frac{1}{n_v} \sum_{j=1}^{n_v} 1(Z'_{vj}\hat{\beta} < c) \left[1(\log Y_{vj}^* < Z'_{vj}\hat{\beta}) - \frac{1}{2}\right] Z_{vj} W'_{vj}\right) \left(\frac{W'_v W_v}{n_v}\right)^{-1} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} W_{pi}\right) = o_p(n_v^{-1/2}).$$

With the usual Taylor expansion, we can write

$$\begin{aligned} \sqrt{n_v}(\hat{\beta} - \beta_o) &= -H^{-1} \left(\frac{1}{\sqrt{n_v}} \sum_{j=1}^{n_v} 1(Z'_{vj}\beta_o < c) \left[1(\log Y_{vj}^* < Z'_{vj}\beta_o) - \frac{1}{2}\right] Z_{vj} W'_{vj}\right) \left(\frac{W'_v W_v}{n_v}\right)^{-1} \\ &\quad \times \left(\frac{1}{n_p} \sum_{i=1}^{n_p} W_{pi}\right) + o_p(1), \end{aligned}$$

where

$$H_{kl} = E_v[1(Z'_{vj}\beta_o < c) f_{\log Y_{vj}^* - Z'_{vj}\beta_o}(0 | Z_{vj}) Z_{vj}^k Z_{vj}^l W'_{vj} W_{vj}] (E_v W_{vj} W'_{vj})^{-1} (E_p W_{pi}).$$

Assuming that the bias term is sufficiently small

$$E_v[1(Z'_{vj}\beta_o < c) [1(\log Y_{vj}^* < Z'_{vj}\beta_o) - \frac{1}{2}] Z_{vj} W'_{vj}] = o_p(1/\sqrt{n_v}),$$

we can write approximately

$$\sqrt{n_v}(\hat{\beta} - \beta_o) \sim N(0, H^{-1} \Lambda H^{-1}),$$

where

$$\Lambda_{kl} = \frac{1}{4} (E_p W_{pi})' (E_v W_{vj} W'_{vj})^{-1} E_v[1(Z'_{vj}\beta_o < c) Z_{vj}^k Z_{vj}^l W'_{vj} W_{vj}] (E_v W_{vj} W'_{vj})^{-1} (E_p W_{pi}).$$

Consistent estimators for H_{kl} and Λ_{kl} are given by

$$\hat{H}_{kl} = \left[\frac{1}{n_v} \sum_{j=1}^{n_v} 1(Z'_{vj}\hat{\beta} < c) \sum_i \frac{1}{a} k\left(\frac{\log Y_{vi}^* - Z'_{vi}\hat{\beta}}{a}\right) Z_{vj}^k Z_{vj}^l W'_{vj} W_{vj} \right] \left(\frac{W'_v W_v}{n_v}\right)^{-1} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} W_{pi}\right)$$

and

$$\hat{\Lambda}_{kl} = \frac{1}{4} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} W_{pi}\right)' \left(\frac{W'_v W_v}{n_v}\right)^{-1} \frac{1}{n_v} \sum_{j=1}^{n_v} 1(Z'_{vj}\hat{\beta} < c) Z_{vj}^k Z_{vj}^l W'_{vj} W_{vj} \left(\frac{W'_v W_v}{n_v}\right)^{-1} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} W_{pi}\right)$$

where $k(\cdot)$ is a kernel function and $a > 0$ goes to zero as sample size increases.

Acknowledgements. We thank M. Buchinsky, S. Cosslett, J. Ham, B. Honoré, J. Horowitz, A. Krueger, L. Lee, C. Manski, I. Prucha, H. White and seminar participants at Maryland, Ohio State, Princeton and the 2003 Summer meetings of the Econometric Society in Evanston, Illinois, for helpful comments. We also thank two anonymous reviewers and especially Hide Ichimura and Bernard Salanié for comments that greatly improved the paper. All remaining errors are our responsibility. This research was completed while Chen was visiting the Department of Economics at Princeton University from September 2001 to June 2002. She thanks the great hospitality provided by Princeton University, especially the Chow Econometrics Research Program. Chen acknowledges the financial support from ESRC/UK grant R00023952. Hong acknowledges the partial support from the NSF and the Chow Econometrics Research Program at Princeton. Tamer acknowledges the support from the National Science Foundation (SES-0112311).

REFERENCES

- AI, C. and CHEN, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica*, **71** (6), 1795–1844.
- BLUNDELL, R., CHEN, X. and KRISTENSEN, D. (2003), "Semi-nonparametric Estimation of Shape Invariant Engel Curves with Endogenous Expenditure" (Working Paper, New York University).
- BOLLINGER, C. (1998), "Measurement Error in the Current Population Survey: A Nonparametric Look", *Journal of Labor Economics*, **16**, 576–594.
- BOUND, J., BROWN, C., DUNCAN, G. and RODGERS, W. (1994), "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data", *Journal of Labor Economics*, **12**, 345–368.
- BOUND, J., BROWN, C. and MATHIOWETZ, N. (2002), "Measurement Error in Survey Data", in J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, Vol. 5 (Amsterdam: North-Holland).
- BOUND, J. and KRUEGER, A. (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?", *Journal of Labor Economics*, 1–24.
- BUCHINSKY, M. (1998), "The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach", *Journal of Applied Econometrics*, **13** (1), 1–30.

- CARROLL, R., RUPPERT, D. and STEFANSKI, L. A. (1995) *Measurement Error in Nonlinear Models* (Chapman and Hall).
- CARROLL, R. and WAND, M. (1991), "Semiparametric Estimation in Logistic Measurement Error Models", *Journal of the Royal Statistical Society*, **53**, 573–585.
- CHEN, X., HANSEN, L. and SCHEINKMAN, J. (1997), "Shape-Preserving Estimation of Diffusions" (Working Paper, Department of Economics, University of Chicago).
- CHEN, X. and SHEN, X. (1998), "Sieve Extremum Estimates for Weakly Dependent Data", *Econometrica*, 289–314.
- CHESHER, A. (1991), "The Effect of Measurement Error", *Biometrika*, **78**, 451–462.
- ENGLE, R., HENDRY, D. and RICHARD, J. (1983), "Exogeneity", *Econometrica*, **51** (2), 277–304.
- FRISCH, R. (1934) *Statistical Confluence Analysis*, Publication No. 5 (Oslo: University Institute of Economics).
- FULLER, W. (1987) *Measurement Error Models* (Wiley).
- GALLANT, R. A. and NYCHKA, D. W. (1987), "Semiparametric Maximum Likelihood Estimation", *Econometrica*, **55**, 363–390.
- HAUSMAN, J., ICHIMURA, H., NEWEY, W. and POWELL, J. (1991), "Measurement Errors in Polynomial Regression Models", *Journal of Econometrics*, **50**, 271–295.
- HAUSMAN, J., NEWEY, W. and POWELL, J. (1995), "Nonlinear Errors in Variables: Estimation of Some Engel Curves", *Journal of Econometrics*, **65**, 205–233.
- HECKMAN, J., ICHIMURA, H. and TODD, P. (1998), "Matching as an Econometric Evaluation Estimator", *The Review of Economic Studies*, **65** (2).
- HONG, H. and TAMER, E. (2003), "A Simple Estimator for Nonlinear Error in Variable Models", *Journal of Econometrics*, **117** (1), 1–19.
- HOROWITZ, J. and MANSKI, C. (1995), "Identification and Robustness with Contaminated and Corrupted Data", *Econometrica*, **63**, 281–302.
- HSIAO, C. and WANG, L. (1995), "A Simulation Based Semi-parametric Estimation of Nonlinear Errors-in-Variables Models" (Working Paper, University of Southern California).
- LAVERGNE, P. and VUONG, Q. (1996), "Nonparametric Selection of Regressors: The Nonnested Case", *Econometrica*, 207–219.
- LEE, L. F. and SEPANSKI, J. H. (1995), "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data", *Journal of the American Statistical Association*, **90** (429), 130–140.
- LI, T. (2002), "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models", *Journal of Econometrics*, **110** (1), 1–26.
- MOLINARI, F. (2003), "Partial Identification of Probability Distributions with Misclassified Data" (Cornell University Working Paper).
- NEWNEY, W. (2001), "Flexible Simulated Moment Estimation of Nonlinear Errors in Variables Models", *Review of Economics and Statistics*, **83** (4), 616–627.
- NEWNEY, W. and MCFADDEN, D. (1994), "Large Sample Estimation and Hypothesis Testing", in R. Engle and D. McFadden (eds.) *Handbook of Econometrics*, Vol. 4 (Amsterdam: North-Holland).
- NEWNEY, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, **62**, 1349–1382.
- POWELL, J. (1984), "Least Absolute Deviations Estimation for the Censored Regression Model", *Journal of Econometrics*, 303–325.
- SCHENNACH, S. M. (2004), "Estimation of Nonlinear Models with Measurement Error", *Econometrica*, **72**, 33–76.
- SEPANSKI, J. and CARROLL, R. (1993), "Semiparametric Quasi-Likelihood and Variance Estimation in Measurement Error Models", *Journal of Econometrics*, **58**, 223–256.
- TAUPIN, M. L. (2001), "Semiparametric Estimation in the Nonlinear Structural Errors-in-Variables Model", *Annals of Statistics*, **29**, 66–93.
- WHITE, H. (1994), "Estimation, Inference and Specification Analysis", in *Econometric Society Monographs*, Vol. 22 (Cambridge University Press).
- WOOLDRIDGE, J. (1995), "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions", *Journal of Econometrics*, **68**, 115–132.
- WOOLDRIDGE, J. (2002), "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification", *Portuguese Economic Journal*, **1**, 117–139.