Imposing Moment Restrictions from Auxiliary Data by Weighting

Author(s): Judith K. Hellerstein and Guido W. Imbens

Source: *The Review of Economics and Statistics*, Feb., 1999, Vol. 81, No. 1 (Feb., 1999), pp. 1-14

Published by: The MIT Press

Stable URL: https://www.jstor.org/stable/2646780

The Review *of* Economics *and* Statistics

VOL. LXXXI FEBRUARY 1999 NUMBER 1

IMPOSING MOMENT RESTRICTIONS FROM AUXILIARY DATA BY WEIGHTING

Judith K. Hellerstein and Guido W. Imbens*

Abstract—In this paper we analyze the estimation of coefficients in regression models under moment restrictions in which the moment restrictions are derived from auxiliary data. The moment restrictions yield weights for each observation that can subsequently be used in weighted regression analysis. We discuss the interpretation of these weights under two assumptions: that the target population (from which the moments are constructed) and the sampled population (from which the sample is drawn) are the same, and that these populations differ. We present an application based on omitted ability bias in estimation of wage regressions. The National Longitudinal Survey Young Men's Cohort (NLS)—in addition to containing information for each observation on wages, education, and experience—records data on two test scores that may be considered proxies for ability. The NLS is a small dataset, however, with a high attrition rate. We investigate how to mitigate these problems in the NLS by forming moments from the joint distribution of education, experience, and log wages in the 1% sample of the 1980 U.S. Census and using these moments to construct weights for weighted regression analysis of the NLS. We analyze the impacts of our weighted regression techniques on the estimated coefficients and standard errors of returns to education and experience in the NLS controlling for ability, with and without the assumption that the NLS and the Census samples are random samples from the same population.

I. Introduction

Economists seldom make use of weighted estimators.[1] This is due, in part, to the fact that consistency of the estimated coefficients of a correctly specified model is often achieved without weighting. But it is also due to difficulties in determining which weights are appropriate and how to interpret the differences between the results of various weighting schemes. Although sampling weights that accompany longitudinal datasets are initially designed to account for carefully documented stratification schemes, they typically are adjusted in complex ways to mitigate nonresponse and attrition, making it difficult to interpret these weights or to adjust standard errors appropriately. In this paper, we construct weights that are derived from auxiliary data and we propose a weighted estimation methodology that is easy to interpret and implement.

We investigate estimation of wage regressions using data from the 1980 wave of the National Longitudinal Survey Young Men's Cohort (NLS) with weights derived from readily available 1980 Census records. These weights serve two purposes: the weights can increase precision, and in cases in which the primary sample data (in our case the NLS) are not representative of the underlying Census distribution, the weights change the estimand. Rather than estimating population values for the primary sample, the weights shift the sample distribution towards the Census distribution. If the population values corresponding to the Census distribution are of greater interest than those corresponding to the distribution of the primary sample, such a shift will be desirable.

The key idea is to construct weights for the observations from the first dataset to force some moments in the weighted sample to equal the corresponding moments from a second dataset. The weights are constructed optimally in an empirical-likelihood sense to minimize the large sample variance of the estimators of the parameters of interest under the assumption of equality of the two distributions. Given the weights, the functionals of interest are estimated using the same estimating equations that would have been used if the moments from the second dataset were unknown, but with the contribution of each observation in the first dataset multiplied by its corresponding weight. In the case of estimating simple wage regressions, (the case on which we focus), this corresponds to estimating returns to education, experience, and ability, using weighted least squares.

Important links exist between the weighting methods employed in this paper and various strands of the economet-

Received for publication June 2, 1995. Revision accepted for publication February 11, 1998.

* University of Maryland and NBER, and UCLA and NBER, respectively.

A previous version of this paper circulated under the title "Raking and Regression," Harvard Institute of Economic Research working paper, September 1993. We wish to thank Josh Angrist, Gary Chamberlain, Charles Manski, Jim Powell, Geert Ridder, George Tauchen, Robert Valletta, Shlomo Yitzhaki, two referees and Robert Moffitt for comments and suggestions, David Neumark for generously providing the NLS sample, and Kim Bayard for research assistance.

We also wish to thank participants at seminars at Carnegie Mellon University/University of Pittsburgh, Columbia University, Cornell University, Harvard/MIT, Hebrew University, Michigan State, New York University, Northwestern University, Princeton University, Rice University, Tel Aviv University, and Texas A&M University.

Imbens also wishes to thank the NSF for financial support through grants 91-22477 and 95-11718 and the Alfred Sloan Foundation for a Research Fellowship.

[1] For example, in a sample of twenty papers that utilize data from the NLS, we found only one reference to the use of the NLS sample weights. Keane et al. (1988) report, "we employ survey weights in all our analyses," and then in a footnote they add, "it turned out that unweighted estimates are almost identical to the weighted ones."

The Review of Economics and Statistics, February 1999, 81(1): 1–14

rics and statistics literature. First, our methods are closely related to recent alternatives to GMM estimation based on empirical likelihood methods (Back & Brown, 1990; Imbens, 1997; Qin & Lawless, 1994; Imbens et al., 1998). The estimators used in this paper are, in fact, a special case of empirical-likelihood estimators for GMM models with the overidentifying moments not depending on unknown parameters. Second, our methods are related to the statistical literature on missing data (Rubin, 1977; Little & Rubin, 1987). A key difference with this literature is that we do not use the unit level Census data, only averages of functions of the Census variables. Third, our methods can be viewed as an extension of the work on estimation of cell probabilities in contingency tables with known marginals (Deming & Stephan, 1942; Ireland & Kullback, 1968; Little & Wu, 1991), where we relax both the multinomial nature of the contingency table problem and the assumption that the marginal distributions are known without sampling error. Finally, our work complements that of Imbens and Lancaster (1994) who analyze estimation of parameters of a conditional distribution under moment restrictions constructed from aggregate data. In contrast to their work, we do not make parametric assumptions. Therefore, although we do not achieve some of the efficiency gains they report from using auxiliary information, we are able to interpret the results when the target and sampled population differ.

II. Background

In this section we discuss two simple examples in order to motivate the dual purposes of weighting. In the first example, we focus on the manner in which incorporating weights into an estimation can increase its precision. In the second example, we show how weighting can shift the estimand.

Example 1. We are interested in the expected value α^* of a random variable Z. We have a random sample of size N of an infinite population. Population averages will be denoted by $E[\cdot]$. Then, with no other information about the shape of the distribution, the efficient estimator for the population average of Z, $\alpha^* = E[Z]$, is

$$\overline{\alpha} = \overline{z} = \frac{1}{N}\sum_{n=1}^{N} z_n, \tag{1}$$

with normalized variance $N \cdot V(\overline{\alpha}) = V(Z)$. Now consider estimation of α^* given prior knowledge of $p^* = Pr(Z > 0)$. While $\overline{\alpha}$ is still a consistent estimator of α^*, it is no longer efficient. The efficient estimator for α^* is a weighted average of the averages in the subpopulations indexed by $Z > 0$ and $Z \leq 0$:

$$\hat{\alpha} = p^* \cdot \overline{z}_1 + (1 - p^*) \cdot \overline{z}_0 \tag{2}$$

where $\overline{z}_1 = (\Sigma\delta(z_n) \cdot z_n)/\Sigma\delta(z_n)$ and $\overline{z}_0 = (\Sigma(1 - \delta(z_n)) \cdot z_n)/\Sigma(1 - \delta(z_n))$. In this notation, $\delta(z)$ is the indicator function for the event $z > 0$. This estimator can also be written as a weighted average of the z_n:

$$\hat{\alpha} = \frac{1}{N}\sum_{n=1}^{N} w_n \cdot z_n \tag{3}$$

with weights

$$w_n = \left(\frac{p^*}{\hat{p}}\right)^{\delta(z_n)} \cdot \left(\frac{1 - p^*}{1 - \hat{p}}\right)^{1-\delta(z_n)}. \tag{4}$$

In this representation, $\hat{p} = \overline{\delta(z)} = \Sigma\delta(z_n)/N$ is the fraction of observations with $z_n > 0$. The normalized variance of the limiting distribution of $\hat{\alpha}$ in large samples is equal to $E[V(Z|\delta(Z))]$. In large samples, the difference between the normalized variances of $\overline{\alpha}$ and $\hat{\alpha}$ is $V(Z) - E[V(Z|\delta(Z))] = V(E[Z|\delta(Z)]) > 0$. It is the last representation of $\hat{\alpha}$ (the weighted average of z_n with the weights depending on the marginal information) that is the focus of this paper. Intuitively, the weighting makes the sample more representative of the population by correcting the relative weights of the positive Z and negative Z subsamples $\hat{p}$ to p^* and $1 - \hat{p}$ to $1 - p^*$, respectively, which in the process leads to a more precise estimator.

An alternative interpretation of this example (discussed in Lancaster (1991)) is that, in large samples and conditional on the ancillary statistic $\Sigma\delta(z_n)$ (and hence conditional on $\hat{p}$), $\overline{\alpha}$ and $\hat{\alpha}$ have the same normalized variance $E[V(Z|\delta(Z))]$, but $\hat{\alpha}$ is unbiased while $\overline{\alpha}$ has expectation $E[\overline{\alpha}|\Sigma\delta(z_n)] = \alpha^* + (\hat{p} - p^*) \cdot (E[Z|\delta(Z) = 1) - E[Z|\ \delta(Z) = 0])$, which in general differs from α^*.

Example 2. The second example concerns the Weighted Exogenous Sampling Maximum Likelihood (WESML) estimator for discrete choice models with choice-based sampling proposed by Manski and Lerman (1977) and discussed in Cosslett (1981) and Imbens (1992). Let Y be a binary outcome whose distribution we wish to express in terms of the distribution of some regressor vector X. In the *target* population, the conditional probability of the event $Y = 1$ given $X = x$ is assumed to have a probit form:

$$Pr(Y = 1|X = x) = \Phi(x'\theta) = \int_{-\infty}^{x'\theta} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right] dz,$$

with unknown parameter θ, and the (unknown) marginal probability density function of X is $p(x)$. With a random sample from the target population, the researcher could estimate θ by maximum-likelihood methods, which would

essentially amount to solving the likelihood equation

$$0 = \sum_{n=1}^{N} \frac{\partial \ln f}{\partial \theta}(y_n|x_n; \theta),$$

where the conditional density $f(y|x; \theta)$ equals

$$f(y|x; \theta) = \Phi(x'\theta)^y \cdot (1 - \Phi(x'\theta))^{1-y}.$$

Instead, the researcher is assumed to have a *choice-based* sample, where, with (known) probability r, an observation is drawn randomly from the stratum with $y = 1$, and where, with probability $1 - r$, an observation is drawn from the stratum with $y = 0$. We can view such a sample as a random sample from a different population, which we call the *sampled* population. We are interested in the parameter θ that would solve, in the limit, the likelihood equations in a random sample from the target population, but we have only a random sample from the sampled population. In addition, it is assumed that the probability in the target population of the event $y = 1$, denoted by $q = \int \Phi(x'\theta)\, dP(x)$ is known. Manski and Lerman proposed estimating θ given a choice-based sample by maximizing the weighted likelihood, or, equivalently, by solving the weighted-likelihood equation

$$0 = \sum_{n=1}^{N} w_n \cdot \frac{\partial \ln f}{\partial \theta}(y_n|x_n; \theta),$$

where

$$w_n = \left(\frac{q}{r}\right)^y \cdot \left(\frac{1-q}{1-r}\right)^{1-y}. \tag{5}$$

The WESML estimator is consistent for the parameter of interest, i.e., for the θ that solves the limiting likelihood equations given a random sample from the target population, where solving the unweighted likelihood equations would, in general, not lead to a consistent estimator.

In this paper we present an approach that unifies the roles of weights in affecting precision and in changing the estimand. We also extend the examples by

i. focusing on more general estimands,
ii. allowing the marginal information to be anything that can be represented as the expectation of a known function of the variables in the first dataset,
iii. allowing for more general differences between the target and sampled populations, and
iv. allowing for sampling error in the moments constructed from the second dataset.

III. Linear Regression with Moment Restrictions

We have N independent realizations $\{z_1, z_2, \ldots, z_N\}$ of a random variable $Z = (Y, X)$ with unknown probability density function $f(y, x)$. Y is a scalar random variable, and X is a vector of dimension K. The population quantity of interest, θ^*, is the vector of linear regression coefficients $\theta^* = E[XX']^{-1} E[XY]$. The least-squares estimator is

$$\hat{\theta}_{\text{OLS}} = \left[\sum_{n=1}^{N} x_n x_n'\right]^{-1} \left[\sum_{n=1}^{N} x_n y_n\right].$$

As N gets large, the distribution of $\sqrt{N}(\hat{\theta}_{\text{OLS}} - \theta^*)$ converges to a normal distribution with mean zero and variance:

$$V_{\hat{\theta}_{\text{OLS}}} = E[XX']^{-1} E[(Y - X'\theta^*)^2 XX'] E[XX']^{-1}$$

$$= E[XX']^{-1} E[\epsilon^2 XX'] E[XX']^{-1}.$$

We do not assume that the errors, $\epsilon = Y - X'\theta^*$, are homoskedastic, and therefore the variance is the Eicker (1963) and White (1980) heteroskedasticity-consistent variance.

Now consider estimating θ^* when, in addition to a random sample of Z, we have exact knowledge of the expectation h^*, in the same population of an R dimensional function of Y and X, denoted by $\bar{h}(Y, X)$. Formally, $h^* = E[\bar{h}(Y, X)] = \int \bar{h}(y, x)\, dF(y, x)$. Examples include $\bar{h}(Y, X) = Y$, where the researcher knows the mean of Y, or $\bar{h}(Y, X) = 1[(Y, X) \in C]$, where the researcher knows the probability that (Y, X) is in a particular subset C of the sample space. This implies the moment restriction $E[h(Y, X)] = 0$, where $h(Y, X) = \bar{h}(Y, X) - h^*$. For example, if we know the mean of Y, the corresponding restriction would be $E[h(Y, X)] = 0$ with $h(Y, X) = \bar{h}(Y, X) - h^* = Y - E[Y]$. We propose estimating θ^* in this framework by weighted least squares:

$$\hat{\theta}_{\text{WLS}} = \left[\sum_{n=1}^{N} \hat{w}_n x_n x_n'\right]^{-1} \left[\sum_{n=1}^{N} \hat{w}_n x_n y_n\right], \tag{6}$$

where the scalar weights $\hat{w}_n$ solve

$$\max_{w} \sum_{n=1}^{N} \ln w_n, \quad \text{subject to} \sum_{n=1}^{N} w_n = 1 \quad \text{and} \sum_{n=1}^{N} w_n \cdot h(y_n, x_n) = 0. \tag{7}$$

(If there are no restrictions of the form $\Sigma\, w_n h(y_n, x_n) = 0$, the weights $\hat{w}_n$ equal $1/N$.) Consequently, $\hat{\theta}_{\text{WLS}} = \hat{\theta}_{\text{OLS}}$. The

large sample properties of this estimator are given in the following theorem.

Theorem 1. *Given regularity conditions, the estimator* $\hat{\theta}_{\text{WLS}}$ *for* θ^* *has the following asymptotic properties:*

$$\hat{\theta}_{\text{WLS}} \xrightarrow{p} \theta^*$$

$$\sqrt{N}(\hat{\theta}_{\text{WLS}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, E[XX']^{-1}(E[\epsilon^2 XX'] - E[\epsilon Xh']E[hh']^{-1}E[\epsilon hX'])E[XX']^{-1}).$$

Proof: See appendix.

We could have written the estimation problem in a more standard GMM form as estimating θ^* under the moment restrictions $E[\psi(y, x, \theta^*)] = 0$, where

$$\psi(y, x, \theta) = \begin{pmatrix} x(y - x'\theta) \\ h(y, x) \end{pmatrix}. \tag{8}$$

Given these moment functions, the standard GMM approach (Hansen, 1982; Newey & McFadden, 1994) estimates θ^* by minimizing the quadratic form

$$Q_C(\theta) = \left[\sum_{n=1}^{N} \psi(y_n, x_n, \theta)\right]' \cdot C \cdot \left[\sum_{n=1}^{N} \psi(y_n, x_n, \theta)\right].$$

Let $\hat{\theta}_{GMM}$ be the minimand of $Q_C(\theta)$. The optimal choice for the weight matrix C is $C^* = E[\psi(y, x, \theta^*)\psi(y, x, \theta^*)']^{-1}$, or a consistent estimate thereof. With the optimal weight matrix C^*, the large sample distribution of $\sqrt{N}(\hat{\theta}_{\text{GMM}} - \theta^*)$ is the same as the large sample distribution of $\sqrt{N}(\hat{\theta}_{\text{WLS}} - \theta^*)$ given in Theorem 1. This analogy implies that the same efficiency argument that has been made for conventional GMM estimators (Chamberlain, 1987) can be used to prove efficiency of the estimator proposed in this section. Underlining the link with GMM estimation is the fact that $\hat{\theta}_{\text{WLS}}$ can be viewed as a special case of the empirical-likelihood estimator, which is discussed in the context of GMM problems as an alternative to the conventional two-step estimators by Back and Brown (1990), Qin and Lawless (1994), and Imbens (1997).

We have not made any assumptions on the distribution of $\epsilon = Y - X'\theta^*$. By definition of θ^*, ϵ is uncorrelated with X, but it need not be independent of X, nor does it have to have a normal distribution. If, however, ϵ is known to have a normal distribution, one can improve considerably upon the estimators discussed here. This may be surprising because, in the absence of auxiliary information, knowledge of normality of ϵ does not affect inference nor increase precision. Combined with auxiliary information, however, knowledge of the parametric form of the density of ϵ does affect inference, and efficient estimators no longer have the simple form described above. This case has been analyzed by Imbens and Lancaster (1994).

IV. Estimation when the Target and Sampled Population Differ

In the preceding section we analyzed the proposed estimator $\hat{\theta}_{WLS}$ under the assumption that the moment restrictions are correctly specified, i.e., under the assumption that $E[h(y, x)] = \int h(y, x)\, dF(y, x) = 0$. This need not be the case, and, as in Example 2, weighted estimation is often motivated by the presumption that the population from which the sample was drawn differs from the population of interest. The weights provided in the NLS, for example, are explicitly motivated by the original sampling scheme and by subsequent changes in the sample (due to attrition) over time, and are intended to make the weighted sample representative of the corresponding age cohort of the entire US population.

The case in which the target and sampled populations differ requires additional notation. Let $(y_n, x_n)_{n=1}^N$ be a random sample from a population which we label the *sampled population,* with common density function $f_s(y, x)$. Let $f_t(y, x)$ be the probability density function of the *target population,* borrowing the terminology from Little and Wu (1991). We do not actually have a random sample from this target population, but we know the expectation of a vector-valued function $\bar{h}(Y, X)$ of Y and X over its distribution:

$$E_t[\bar{h}(Y, X)] = \int \bar{h}(y, x)\, dF_t(y, x) = h^*.$$

The subscript t of the expectations operator indicates the distribution over which the expectation is taken. We can also capture this information as knowledge of a function $h(Y, X) = \bar{h}(Y, X) - h^*$, which is known to have expectation zero in the target population ($E_t[h(Y, X)] = 0$). The function $h(Y, X)$ need not have expectation zero when the expectation is taken over the sampled population. In this case, we have to take extra care in defining the parameters of interest. Let θ_g^* be the population value corresponding to the solution to the estimating equations $\Sigma_n x_n(y_n - x_n'\theta) = 0$ given a sample drawn randomly from the population with probability density function $f_g(y, x)$. The following theorem gives the large sample results for this case.

Theorem 2. *If the target distribution* $f_t(y, x)$ *and the sampled distribution* $f_s(y, x)$ *differ, then, under regularity conditions,*

$$\hat{\theta}_{\text{WLS}} \xrightarrow{p} \theta_{st}^*,$$

where

$$f_{st}(y, x) = \frac{f_s(y, x)}{1 + \lambda_{st}^{*\prime} h(y, x)},$$

with λ_{st}^{*} *the solution to*

$$\max_{\lambda} E_s\,[\ln\,(1 + \lambda' h(Y, X))].$$

In addition,

$$\sqrt{N}(\hat{\theta}_{\mathrm{WLS}} - \theta_{st}^{*}) \xrightarrow{d} \mathcal{N}(0, E_s[\tilde{X}\tilde{X}']^{-1}(E_s[\tilde{\epsilon}^2\tilde{X}\tilde{X}'] - E_s[\tilde{\epsilon}\tilde{X}\tilde{h}']E_s[\tilde{h}\tilde{h}']^{-1}E_s[\tilde{\epsilon}\tilde{h}\tilde{X}'])E_s[\tilde{X}\tilde{X}']^{-1}),$$

where

$$\tilde{X} = X/\sqrt{1 + \lambda_{st}^{*\prime} h(Y, X)},$$

$$\tilde{\epsilon} = (Y - X'\theta_{st}^{*})/\sqrt{1 + \lambda_{st}^{*\prime} h(Y, X)},$$

and

$$\tilde{h}(Y, X) = h(Y, X)/(1 + \lambda_{st}^{*\prime} h(Y, X)).$$

Proof: See appendix.

Theorem 2 shows that, in the general case where $E_s[h(Y, X)]$ differs from zero, the target of estimation θ_{st}^{*} is the probability limit of the estimator based on unweighted estimation using a random sample from an artificial population with probability density function $f_{st}(y, x)$. This distribution can be interpreted as the distribution closest to the sampled distribution (i.e., $f_s(y, x)$), in an empirical-likelihood sense, subject to the restrictions that it has the moments $E_{st}[h(y, x)]$ in common with the target distribution (i.e., $f_t(y, x)$), that is, subject to $E_{st}[h(Y, X)] = E_t[h(Y, X)] = 0$. Under some conditions, this artificial population has the same distribution as the target population. Formally, if the distribution in the sample, $f_s(y, x)$, can be written as

$$f_s(y, x) = f_t(y, x) \cdot (1 + \gamma' h(y, z))$$

for some vector γ, then matching on the moments $E[h(Y, X)]$ will lead to an artificial distribution $f_{st}(y, x)$ that is identical to the target distribution $f_t(y, x)$. This follows because the probability limit λ_{st}^{*} of $\hat{\lambda}$ will in this case be equal to γ. As an example of this, consider the choice-based sampling example (Example 2) introduced in section II. In this case the distribution in the target population is

$$f_t(y, x) = \Phi(x'\theta)^y \cdot (1 - \Phi(x'\theta))^{1-y} \cdot p(x).$$

The distribution in the sampled population is

$$f_s(y, x) = \left[\frac{r}{q} \cdot \Phi(x'\theta)\right]^y \cdot \left[\frac{1 - r}{1 - q} \cdot (1 - \Phi(x'\theta))\right]^{1-y} \cdot p(x).$$

We match on the marginal probability of the event $Y = 1$, or $h(y, x) = y - q$, implying that the probability limit λ_{st}^{*} of $\hat{\lambda}$ is the solution to the equation

$$0 = E_s\left[\frac{h(Y, X)}{1 + \lambda \cdot h(Y, X)}\right] = E_s\left[\frac{Y - q}{1 + \lambda \cdot (Y - q)}\right]$$

$$= \int \sum_{y=0}^{1} \frac{y - q}{1 + \lambda \cdot (y - q)} \left[\frac{r}{q} \cdot \Phi(x'\theta)\right]^y \times \left[\frac{1 - r}{1 - q}(1 - \Phi(x'\theta))\right]^{1-y} \cdot dP(x).$$

The solution is

$$\lambda_{st}^{*} = \frac{r - q}{q(1 - q)},$$

implying that the "intermediate" distribution, $f_{st}(y, x)$ equals $f_t(y, x)$:

$$f_{st}(y, x) = \frac{f_s(y, x)}{1 + \lambda \cdot h(y, x)}$$

$$= \frac{1}{1 + (y - q)(r - q)/(q(1 - q))} \left[\frac{r}{q} \cdot \Phi(x'\theta)\right]^y \times \left[\frac{1 - r}{1 - q}(1 - \Phi(x'\theta))\right]^{1-y} \cdot p(x)$$

$$= \Phi(x'\theta)^y \cdot (1 - \Phi(x'\theta))^{1-y} \cdot p(x) = f_t(y, x).$$

In this example, the Lagrange multiplier λ_{st}^{*} that forces the weighted sample moment $\Sigma w_n y_n/N$ to match the target moment q reweights the sample in the limit exactly back to the target distribution. In practice, it is unlikely that matching on a few moments will lead to an artificial population with exactly the same distribution as the target population. However, as more and more moments are matched, the artificial distribution will get close to the target distribution. In particular, it may be possible to obtain enough of a resemblance between the artificial distribution and the target

distribution with only a few matched moments so that $\text{plim}(\hat{\theta}_{\text{WLS}}) = \theta^*_{st} = \theta^*_t$, even though $f_{st}(y, x) \neq f_s(y, x)$. The extreme example of this occurs when θ depends only on a finite number of moments of the joint distribution of Y and X. Matching exactly on those moments leads to an artificial distribution $f_{st}(y, x)$ that can be different from $f_t(y, x)$, even though it will be the case that $\theta^*_{st} = \theta^*_t$.

An interesting connection emerges with the missing data literature. (See Little and Rubin, 1987, for a survey.) Suppose that the first dataset consists of observations on (z_1, z_2), and the second dataset consists of observations on z_2 alone. If we match on a large number of expectations of functions of z_2, and if the sequence of these functions spans a large enough space, the intermediate distribution will converge to

$$f_{st}(z) = f_s(z_1|z_2) \cdot f_t(z_2).$$

This will equal the target distribution if the conditional distribution of the "missing variable" z_1 conditional on the "observed variable" z_2 is the same in the target and sampled distribution, i.e., if $f_t(z_1|z_2) = f_s(z_1|z_2)$. This condition implies that, if we consider the two datasets together with z_1 missing for some of the observations, the missing data are *missing at random* according to the definition of Rubin (1977).

V. Accounting for Sampling Error in the Moment Restrictions

In the previous sections we assumed that the extra information was in the form of a vector h^*, which is exactly equal to the expectation in the target population of a known function $\bar{h}(\cdot)$ of the random variables Y and X. We imposed the restriction $0 = E_t[h(Y, X)] = E_t[\bar{h}(Y, X) - h^*]$, taking h^* as fixed, even though h^* was actually estimated using a sample from the target population. This may be an adequate procedure when the second dataset is much larger than the first dataset and when the sampled and target distribution are not too different. However, when the techniques developed in this paper are applied to combinations of similarly sized datasets, or to samples from very different distributions, the sampling error in the estimation of the moments of the second dataset should be taken into account. In this section, we generalize the results to the case in which we do not know $h^* = E_t[\bar{h}(Y, X)]$ with certainty.

Suppose we have an estimate $\hat{h}$ of h^*, based on an average of $\bar{h}(y_i, x_i)$ over a random sample of size M from the target population. Based on such a random sample, the estimate $\hat{h} = 1/M \sum_{j=1}^{M} \bar{h}(y_j, x_j)$ for h^* would satisfy

$$\sqrt{M}(\hat{h} - h^*) \xrightarrow{d} \mathcal{N}(0, \Delta_h)$$

with

$$\Delta_h = E_t[\bar{h}(Y, X) - h^*] \cdot [\bar{h}(Y, X) - h^*]'.$$

We assume that the extra information is in the form of the estimate $\hat{h}$ and its approximate variance Δ_h/M. In addition, we assume that $\hat{h} - h^*$ is independent of the first sample $\{(y_n, x_n)\}_{n=1}^{N}$. We estimate θ by treating $h(y, x) = \bar{h}(y, x) - \hat{h}$ as the moment to be restricted to have expectation zero. We investigate the behavior of the estimator as N and M (the number of observations in both datasets) go to infinity with their ratio converging to a constant $k = M/N$. This is the only interesting case because if N and M converge at different rates, then in large samples the sampling variation in the larger data set can be ignored. To facilitate comparison with the exposition in the previous sections, we assume that M/N is exactly equal to some integer k. We can therefore think of having N observations z, where z_n consists of $(y_n, x_n, \bar{h}_{n1}, \ldots, \bar{h}_{nk})$, i.e., the pair (y_n, x_n) and k observations $(\bar{h}_{n1}, \ldots, \bar{h}_{nk})$. In this setup, the estimating equations for $\hat{\theta}$, $\hat{\lambda}$ and $\hat{h}$ are

$$0 = g(\hat{\theta}, \hat{\lambda}, \hat{h}) = \frac{1}{N}\sum_{n=1}^{N} \begin{pmatrix} x_n \cdot (y_n - \hat{\theta}'x)/(1 + \hat{\lambda}'(\bar{h}(y_n, x_n) - \hat{h}) \\ (\bar{h}(y_n, x_n) - \hat{h})/(1 + \hat{\lambda}'(\bar{h}(y_n, x_n) - \hat{h})) \\ \frac{1}{k}\sum_{j=1}^{k} (\bar{h}_{nj} - \hat{h}) \end{pmatrix}.$$

Solving this leads to $\hat{h} = \sum_{n=1}^{N} \sum_{j=1}^{k} h_{nj}/(N \cdot k)$, and $\hat{\theta}$ and $\hat{\lambda}$ solving the same equations as before, given in Theorem 2, with $h(y, x)$ replaced by $\bar{h}(y, x) - \hat{h}$. The following theorem describes the large sample properties of the estimator under these conditions.

Theorem 3. *When N and M go to infinity, with M/N = k, we have, under regularity conditions,*

$$\begin{pmatrix} \hat{\theta} \\ \hat{\lambda} \\ \hat{h} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \theta^*_{st} \\ \lambda^*_{st} \\ h^* \end{pmatrix}.$$

The variance/covariance matrix has the standard form for generalized method of moments estimators:

$$\sqrt{N}\begin{pmatrix} \hat{\theta} - \theta^*_{st} \\ \hat{\lambda} - \lambda^*_{st} \\ \hat{h} - h^* \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Gamma^{-1}\Delta(\Gamma')^{-1}\right)$$

where, as before,

$$\tilde{X} = X \Big/ \sqrt{1 + \lambda_{st}^{*\prime}(\bar{h}(Y, X) - h^*)},$$

$$\tilde{\epsilon} = (Y - X'\theta_{st}^*) \Big/ \sqrt{1 + \lambda_{st}^{*\prime}(\bar{h}(Y, X) - h^*)},$$

$$h = \bar{h}(Y, X) - h^*$$

and

$$\tilde{h} = (\bar{h}(Y, X) - h^*)/(1 + \lambda_{st}^{*\prime}(\bar{h}(Y, X) - h^*)).$$

We then have

$$\Gamma = \begin{pmatrix} -E_s\tilde{X}\tilde{X}' & -E_s\tilde{X}\tilde{\epsilon}\tilde{h}' & E_s\tilde{X}\tilde{\epsilon}\lambda_{st}^{*\prime}/(1 + \lambda_{st}^{*\prime}h) \\ 0 & -E_s\tilde{h}\tilde{h}' & -\mathcal{J}_R E_s 1/(1 + \lambda_{st}^{*\prime}h)^2 \\ 0 & 0 & -\mathcal{J}_R \end{pmatrix}$$

and

$$\Delta = \begin{pmatrix} E_s\tilde{\epsilon}^2\tilde{X}\tilde{X}' & E_s\tilde{\epsilon}X\tilde{h}' & 0 \\ E_s\tilde{\epsilon}\tilde{h}\tilde{X}' & E_s\tilde{h}\tilde{h}' & 0 \\ 0 & 0 & \Delta_h/k \end{pmatrix}.$$

In particular, the large sample variance of $\sqrt{N}(\hat{\theta} - \theta_{st}^*)$ *equals*

$$E_s[\tilde{X}\tilde{X}']^{-1}(E_s[\tilde{\epsilon}^2\tilde{X}\tilde{X}'] - E_s[\tilde{\epsilon}\tilde{X}\tilde{h}']E_s[\tilde{h}\tilde{h}']^{-1} \times E_s[\tilde{\epsilon}\tilde{h}\tilde{X}'])E_s[\tilde{X}\tilde{X}']^{-1} + V\frac{\Delta_h}{k}V', \tag{9}$$

where

$$V = E_s[\tilde{X}\tilde{X}]^{-1}E_s[\tilde{X}\tilde{\epsilon}\tilde{h}']E_s[\tilde{h}\tilde{h}']^{-1}E_s\left[\frac{\mathcal{J}_R}{(1 + \lambda_{st}^{*\prime}h)^2}\right] + E_s[\tilde{X}\tilde{X}']^{-1}E_s\left[\frac{\tilde{X}\tilde{\epsilon}\lambda_{st}^{*\prime}}{1 + \lambda_{st}^{*\prime}h}\right].$$

Proof: See appendix.

If M is very large relative to N (i.e., if Δ_h/k is close to zero), the variance is dominated by the first term in equation (9), which is the variance given in Theorem 2 for the case where h^* is known without sampling error. If, in addition, we substitute $\lambda_{st}^* = 0$, we obtain the variance for the case where $f_s(y, x) = f_t(y, x)$ given in Theorem 1. The second term in the variance of $\hat{\theta}$ can be substantial, however, even with relatively large k, if the target and sampled distribution are very different. In that case, the weights will have a high variance, leading to an increase in the variance of $\hat{\theta}$ due the presence of a factor involving squares of the weights in the variance formula in Theorem 3.

VI. The Composition of the NLS

The NLS (that is, the NLS Young Men's Cohort) sample of 5,225 young men was drawn in 1966 to represent the civilian, noninstitutionalized population of men ages 14 through 24. Even at its inception, the NLS was a relatively small sample, but the benefits of the survey are that it is longitudinal and contains detailed information about each individual in the sample. The individuals selected into the sample were interviewed almost annually until 1981, after which the survey was discontinued. The NLS suffers from a very high attrition rate: By 1980, the year of the data we use below, only 3,438 (65.8%) of the men originally surveyed remained in the sample. Some of the attrition in early years was due to the fact that a number of the men entered the military and were thus excluded from the sample, but attrition rates remained high even after the Vietnam conflict (Rhoton, 1984).

Three issues must be considered when addressing the "representativeness" of data from the 1980 NLS. The first issue is the representativeness of the original NLS sample as drawn in 1966. According to Rhoton, the original NLS sample in 1966 differed from the 1966 Current Population Survey (CPS). This suggests that the NLS may never have been representative of the U.S. population. The second issue is of missing data for certain individuals in the NLS. Almost 2,000 observations in the original NLS cohort do not have information on IQ scores. Griliches et al. (1978) show that IQ is not missing completely at random; the probability of reported IQ for a given observation is correlated with variables such as age and education. Given that we use the data on IQ in our empirical analysis below, this is an issue.

Finally, there is the issue of attrition. If attrition from the NLS were entirely random or were a function solely of factors uncorrelated with any variables of interest, one would worry about attrition only to the extent that it further reduced the NLS sample size. Attrition in the NLS was not random, however, and it was correlated with factors such as income (Rhoton & Nagi, 1991), age, and education, (Griliches et al., 1978) all of which are all relevant for human capital regressions. The NLS does contain sampling weights that were updated each year of the survey in an attempt to continue to keep the NLS representative of the U.S. population for that age cohort (with the exception that no attempt was made to account for immigration). The original sample weights in 1966 were constructed to reflect the original multistage, clustered-sample design and to adjust for differential response rates across segments of the population and oversampling of blacks. This adjustment process was bound

to be somewhat imperfect since the U.S. population used as the base comparison group was the 1960 Census extrapolated forward to 1966. In subsequent years, further adjustments were made to the sampling weights to try to account for the nonrandom nature of attrition. This was done by dividing the original sample into cells defined by race, education of father, and years in place of residence at the first interview; calculating the response rate within a cell; and adjusting by cell the sampling weights of remaining respondents. If the original (weighted) 1966 sample was not representative of the U.S. population, this adjustment of sampling weights would not make later years of the survey representative. Moreover, it is not clear that the cell adjustments adequately capture the nonrandom nature of the attrition. Details on the weighting procedures are given in NLS Users' Guide (1995).

VII. Returns to Schooling and Census Information

In this section we apply the above analysis to the estimation of wage equations. Since the work of Mincer (1974), economists have estimated returns to education by running least-squares regressions on variants of the following equation:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{experience}_i + \beta_3 \cdot \text{experience}_i^2 + \epsilon_i. \quad (10)$$

In our analysis, we define the wage measure as the usual hourly wage, and education as the highest grade completed. Experience here is the typical "potential experience" measure (calculated as age minus six minus years of education). The NLS (sub)sample we use consists of observations from 1980 on 815 white men between the ages of 28 and 38. These 815 men are the same group used in Blackburn and Neumark (1992) in their study of returns to education. Because Blackburn and Neumark use information on the complete work histories of men in the NLS, the sample is restricted to men for whom data is also available on labor force participation in every year prior to 1980. Given that one of the special and commonly used features of the NLS (and other panel datasets) is its longitudinal nature, this restriction is a reasonable one for us to maintain when studying the effects of combining the NLS with auxiliary data from the Census. The first two sets of results in table 1 give the least-squares estimates of the coefficients in equation (10) based on our NLS sample. We report both unweighted "unit-weight" estimates and estimates using the sampling weights provided with the NLS.

A large body of literature has considered the biases that result from the possibility that there is variation across individuals in the ability that makes them more likely to get schooling, and which has also an independent effect on earnings. (See Griliches, 1977, for an early overview; this is also the topic of Blackburn and Neumark, 1992.) The NLS is one of only a handful of datasets that contain measures of both ability and schooling. In particular, the NLS reports data on an IQ test score as well as the results of another ability test, KWW (Knowledge of the World of Work). An alternative to equation (10) is then

$$\ln(\text{wages}_i) = \beta_0 + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{experience}_i + \beta_3 \cdot \text{experience}_i^2 + \beta_4 \cdot IQ_i + \beta_5 \cdot \text{KWW}_i + \epsilon_i. \quad (11)$$

The first two sets of results in table 2 give least-squares estimates of equation (11) based on our NLS sample, again unweighted and weighted with the NLS-provided weights.

The NLS is a relatively small dataset, however, and estimates based on equation (11) using the NLS are not nearly as precise as they would be if derived from data from the CPS or the Census. Of course, ability measures are not available in those larger datasets. The information we use from the Census consists of the means, variances, and covariances of (the log of) hourly wages, education, experience, and experience-squared. These are calculated from a subsample of 127,345 observations from the 1% Public Use Microdata Sample (PUMS) of the 1980 Census data that was constructed to mimic as closely as possible the selection process that was used to obtain the NLS sample. We extracted data for nonselfemployed working white men between the ages of 28 and 38, earning at least half of the minimum wage (to remove outliers), and working in nonfarm occupations. Since the NLS obtained test scores from high schools attended by the sample respondents, we selected only those men in the Census with nine or more years of education. The NLS topcodes education at 18, so we did the same for our Census sample. Finally, since the NLS topcodes annual earnings at $50,000, a level lower than the official Census topcode, we topcoded annual earnings in the Census at $50,000.[2] We then used annual earnings data along with data on weeks worked and usual hours worked per week to construct hourly wages. With these data, we estimated moments of the joint log hourly wage, years of

TABLE 1.—RETURNS TO SCHOOLING WITHOUT ABILITY MEASURES

	NLS		Weighted NLS		Census	
Var	coeff.	s.e.	coeff.	s.e.	coeff.	s.e.
const	0.398	(0.219)	0.398	(0.228)	0.355	(0.020)
educ	0.081	(0.008)	0.081	(0.008)	0.076	(0.001)
exper	0.085	(0.025)	0.083	(0.027)	0.075	(0.002)
exper^2	−0.002	(0.001)	−0.002	(0.001)	−0.002	(0.000)
sample size	815		815		127,345	

[2] We are grateful to Robert Moffitt for bringing this to our attention.

TABLE 2.—RETURNS TO SCHOOLING WITH ABILITY USING CENSUS MOMENTS

	NLS		Weighted NLS		NLS/Census			
Var	coeff.	s.e.	coeff.	s.e.	coeff.	s.e. (1)	s.e. (2)	s.e. (3)
const	0.294	(0.235)	0.324	(0.240)	0.016	(0.179)	(0.109)	(0.114)
educ	0.054	(0.010)	0.051	(0.011)	0.068	(0.006)	(0.006)	(0.006)
exper	0.068	(0.025)	0.063	(0.027)	0.076	(0.024)	(0.008)	(0.009)
exper^2	−0.002	(0.001)	−0.002	(0.001)	−0.002	(0.001)	(0.000)	(0.000)
iq	0.004	(0.001)	0.003	(0.001)	0.005	(0.001)	(0.001)	(0.001)
kww	0.008	(0.003)	0.009	(0.003)	−0.003	(0.003)	(0.003)	(0.003)

education, and age distributions of the relevant target population of the U.S. population in 1980.

In the third set of results in table 1, we report results based on the Census data from estimating the wage equation without ability measures. This regression can also be interpreted as one based on the NLS data with weights derived from the Census where we match on all first, second, and cross moments of log earnings, experience, and experience-squared. In matching on these moments, we exactly recover the regression based on Census data alone.

As mentioned above, the first two sets of results in table 2 come from estimating wage equations with ability measures using the unit-weighted NLS and the NLS with NLS sampling weights. The last sets of results reported in table 2 are from regressions using the NLS combined with moments from the Census. We match on the thirteen moments consisting of first, second, and cross moments of the common variables: log(wage), education, experience, and experience-squared. Standard errors for all estimates are given in parentheses. The standard errors for the weighted NLS with published weights are calculated using standard weighted least-squares methods, which do not take into account the manner in which the weights are constructed (because properly accounting for effects of the weights is not possible on the basis of the available information).

The first set of standard errors for the NLS/Census estimates (reported in the column labeled s.e. (1) of the NLS/Census results in table 2) is estimated based on Theorem 1 under the assumption that the NLS sample is drawn randomly from the same population as the Census. The second set of standard errors (s.e. (2) of table 2) are estimated without this assumption, based on the results in Theorem 2. The third set of standard errors (s.e. (3)) also take into account the sampling variation in the Census-based moment estimates. They are based on the results in Theorem 3.

The first thing to note is that the NLS results do not change much when the regression is weighted by the sampling weights published with the NLS. The estimates are virtually unchanged from the unweighted ones, as is also noted by Keane et al. (1988). This is not surprising given that the NLS weights are all relatively close to one. The second point is that the results are quite different when the regression is weighted by the weights constructed using the Census data. As figure 1 illustrates, the NLS weights are very different from the Census weights, and the Census weights are much more skewed toward large values.

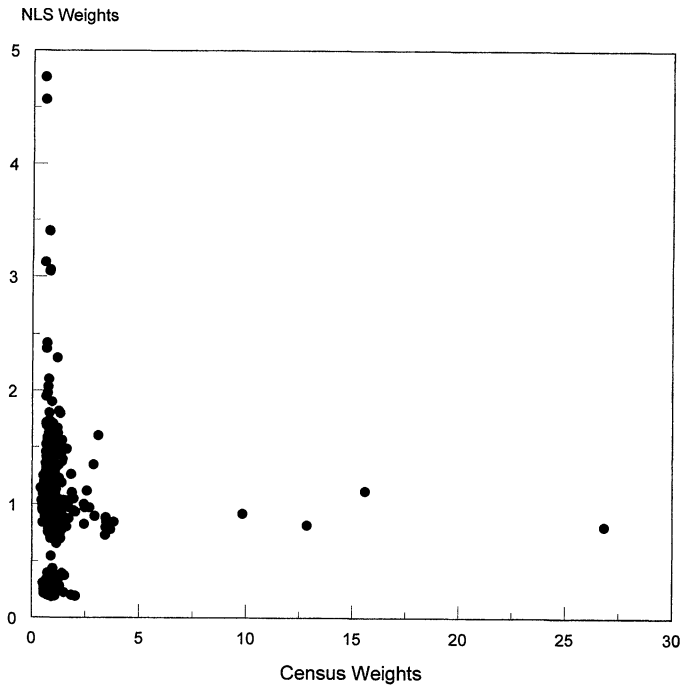
In fact, the rank correlation between the NLS sampling weights and the weights constructed from the Census data is negative at −0.166 and significantly different from zero at standard levels of significance. These results show that the NLS sample is quite different from the Census even after we used selection criteria in the Census (using only white males aged 28 to 38 with at least 9 years of education, topcoding education, and topcoding wages) to try to make the samples as comparable as possible. Moreover, the sampling weights in the NLS do not reweight the sample to reflect the Census. The third point is that using the Census moments leads to an increase in precision where the estimated returns to education are equivalent to having a primary dataset more than twice the size of the NLS sample. Not surprisingly, the sampling variation in the Census moments does not contribute significantly to the sampling variation in the parameter estimates. This is attributable to the facts that the Census is 165 times larger than the NLS sample, and that the match between the distributions of the Census and the NLS is close enough that no NLS observation has a weight that dominates the last term in the variance formula in Theorem 3.

It is important to recall that the Census weights we use are functions of the first and second moments of education, experience and its square, *and* the log of hourly wage. This wage measure, of course, is the dependent variable. This use of Census wage data allows us to help correct for the differences between the NLS and Census that are functions of unobservable characteristics that affect wages, as well as the observables on the right-hand side of the wage equation. The use of the dependent variable in constructing the weights means that the coefficient estimates can always be sensitive to weighting, regardless of the functional form of the wage equation. Indeed, even the coefficients on variables not included in the auxiliary moment function (such as iq and kww) can be affected by this weighting.

To formalize this argument, consider the effect of weighting by x versus weighting by y on the conditional density of y given x. Using a weighting function $w_1(x)$ that depends solely on x implies that the conditional density of y given x in the target population (i.e., weighted) is the same as the sample conditional density (i.e., the unweighted):

$$f_t(y|x) = \frac{w_1(x)f_s(y, x)}{\int_y w_1(x)f_s(y, x)\,dy} = \frac{w_1(x)f_s(y|x)f_s(x)}{w_1(x)f_s(x)} = f_s(y|x).$$

FIGURE 1.—NLS AND CENSUS WEIGHTS



In contrast, weighting by a function of y, $w_2(y)$, can change the conditional density:

$$f_t(y|x) = \frac{w_2(y)f_s(y,x)}{\int_y w_2(y)f_s(y,x)\,dy} = f_s(y|x)\cdot\frac{w_2(y)}{\int_y w_2(y)f_s(y|x)dy},$$

which in general will differ from $f_s(y|x)$.

The significance of the differences between estimates based solely on the NLS and those incorporating Census information is investigated directly in tables 3 through 5. The restrictions implied by the Census information can be tested in a number of ways. The results of two of these methods are given in table 3; the results of a third method are in table 4. (Imbens et al. (1998) discuss a number of alternative testing procedures.) The first method simply compares directly the Census moments used in the restrictions to the corresponding NLS moments. Results from this test are reported in the first set of columns in table 3, which presents the difference between the NLS and Census moments, the corresponding standard errors, and the t-statistics. The second way to examine the impact of the restrictions is to consider the estimates of the Lagrange multipliers formed when constructing the weights for the weighted regression. The second set of columns in table 3 reports the estimates of the Lagrange multipliers λ and the corresponding standard errors and t-statistics. The last row of the table gives the statistics for the tests of the hypotheses that all NLS moments are equal to Census moments, and that all Lagrange multipliers are equal to zero. For all tests, the variances are calculated under the null hypothesis that the target and sampled distributions are equal.

A third way to investigate the difference between the Census and NLS sample is to consider directly the differences in parameter estimates. In table 4 we present the differences in the NLS and combined NLS-Census estimates, with associated standard errors. These standard errors do not assume that target and sampled distributions are equal, and take into account the sampling error in the Census moments. For comparison purposes, we also report the raw estimates from table 2 again. The last row presents a test statistic for the null hypothesis that the NLS and combined NLS-Census estimates are equal. The test statistic has, under

TABLE 3.—TESTS OF EQUALITY OF NLS AND CENSUS DATA

	NLS—Census Moments			Lagrange Multipliers		
Moment	est.	s.e.	t-stat	est.	s.e.	t-stat
educ	−0.35	(0.08)	−4.5	1.02	(0.49)	2.1
log(wage)	0.15	(0.02)	10.2	−1.85	(1.57)	−1.2
exper	0.86	(0.13)	6.5	5.69	(1.59)	3.6
educ^2	−10.65	(2.21)	−4.8	−0.01	(0.01)	−0.9
log(wage)^2	0.53	(0.06)	8.5	−0.16	(0.14)	−1.2
exper^2	21.17	(3.74)	5.7	−0.47	(0.14)	−3.3
educ × log(wage)	1.30	(0.30)	4.3	0.10	(0.05)	1.8
exper × log(wage)	3.67	(0.35)	10.6	0.20	(0.18)	1.1
$\text{exper}^2 \times \text{log(wage)}$	70.27	(8.57)	8.2	−0.00	(0.01)	−0.7
educ × exper	8.47	(1.46)	5.8	−0.16	(0.01)	−2.5
$\text{educ} \times \text{exper}^2$	246.67	(41.98)	5.9	0.01	(0.00)	2.8
exper^3	429.69	(83.98)	5.1	0.02	(0.01)	2.5
exper^4	8.34×10^3	(1.77×10^3)	4.7	−0.00	(0.00)	−2.0
chi-square tests (dof)	311.84 (13)			109.5 (13)		

TABLE 4.—TESTS OF EQUALITY OF UNIT- AND CENSUS-WEIGHTED ESTIMATES

	NLS est.	NLS/Census est	NLS—Census Estimates dif	s.e.	t-stat
const.	0.294	0.016	0.278	(0.288)	1.1
educ.	0.054	0.068	−0.015	(0.010)	−1.2
exper.	0.068	0.076	−0.008	(0.028)	−0.3
exper.^2	−0.002	−0.002	−0.000	(0.000)	−0.3
iq	0.004	0.005	−0.001	(0.002)	−0.3
kww	0.008	−0.003	0.010	(0.005)	2.0
chi-square test (dof)			104.97 (6)		

TABLE 5.—OBSERVATIONS WITH FIVE HIGHEST AND LOWEST CONSTRUCTED WEIGHTS

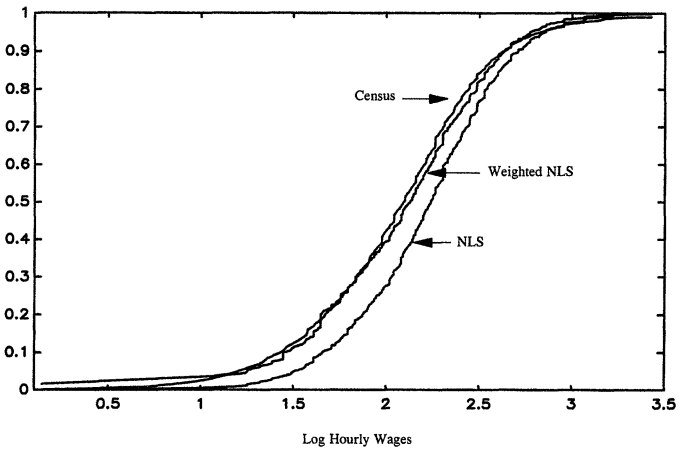
weight	educ	log(wage)	exper	IQ	KWW
26.8	18	1.65	5	107	30
15.6	17	1.45	13	106	34
12.9	12	0.14	11	108	46
3.8	16	2.58	6	119	24
3.7	16	2.49	6	114	28
0.50	12	2.52	20	109	44
0.49	12	2.53	20	96	32
0.49	12	2.53	20	110	37
0.46	12	2.79	20	95	48
0.41	12	3.22	20	82	41

the null, a chi-squared distribution with six degrees of freedom (dof). Even though the t-statistics themselves in the last column of table 3 and in table 4 are not particularly large, they are highly correlated in each case and all of the chi-squared statistics in table 3 and 4 clearly reject the hypothesis that the NLS and Census distributions are equal. The fact that the Census and NLS samples differ significantly in the distributions of wages, education and experience was also reported by Gottschalk and Moffitt (1992) in a comparison of the NLS and the CPS. The methodology presented here provides a clear interpretation of the differences between the samples. While the raw differences in moments suggest that there are particularly large differences in average wages between the NLS and Census, the Lagrange multiplier estimates reported in table 3 suggest that the binding restrictions arise when forcing the first, second, and cross moments of experience and its square in the two samples to be equal.

The regression results in table 2 illustrate that the effect of ability bias on estimates of the return to education is quite sensitive to the datasets employed. Just using the NLS sample suggests that the effect of omitting ability measures on estimates of returns to education is $0.081 - 0.054 = 0.027$, or that almost three percentage points of the estimated return of 8.7% is due to ability bias. If we change the distribution of wages, education, and experience to be closer to that of the Census, the estimate of the effect of ability bias is $0.076 - 0.068 = 0.008$, less than one-third the decrease found using only the NLS data. The weighted interpretation of the new estimator makes it clear that the key difference between the unit- and Census-weighted estimates is the difference across the two samples in the distribution of wages, experience, and education. Since the ability measures are not independent of these three variables, the regression estimates differ considerably depending on which wages, education, and experience distribution we use.

If we do not make the assumption that the two populations—the sampled population from which the sample is drawn and the target population from which the moments are constructed—are the same, the weighted estimator corresponds to an artificial population with probability density function $f_{st}(y, x)$ defined in section IV. It is interesting to investigate some aspects of this distribution. In figures 2 through 4, we present estimates of the distribution functions of education, experience, and the logarithm of hourly wages for the Census distribution, the NLS distribution, and the Census-weighted NLS distribution. For all of the distributions, it is clear that by forcing the Census-weighted NLS distribution to have the same first, second, and cross moments of log wages, education, experience, and experience-squared as the Census distribution, the Census-weighted NLS and the Census are much closer than the unweighted NLS and Census. These figures also show that the men with low wages are clearly underrepresented in the NLS relative to the Census, as are, to a lesser extent, men with high levels of education.

FIGURE 2.—WAGE DISTRIBUTIONS



We can also see these effects by inspecting the weights directly. In table 5 we give an indication of differences between observations with high and low weights. We report the values of all variables for the observations with the five highest and lowest weights. It is apparent from this table (and the rest of the weights, which we do not report) that the segment of the Census population that is overrepresented in the NLS (the men with low weights) consists of men with relatively high wages, low education, and higher than average experience. The average IQ score of these observations is lower than the NLS average of 103, while their KWW is higher than the NLS average of 37. Because the weights do not directly depend on IQ and KWW, the association between the weights and the test scores stems from the association between IQ, KWW, and the variables (including wages) used in the construction of the weights.

One would not want to exclude from the sample observations with high weights for two reasons. First, this would induce sample selection since the weights are partially a function of log wages (the dependent variable). Second,

FIGURE 3.—EDUCATION DISTRIBUTIONS

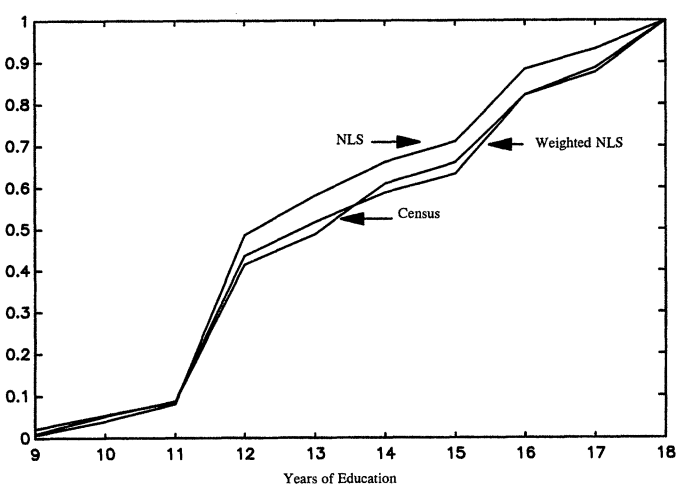
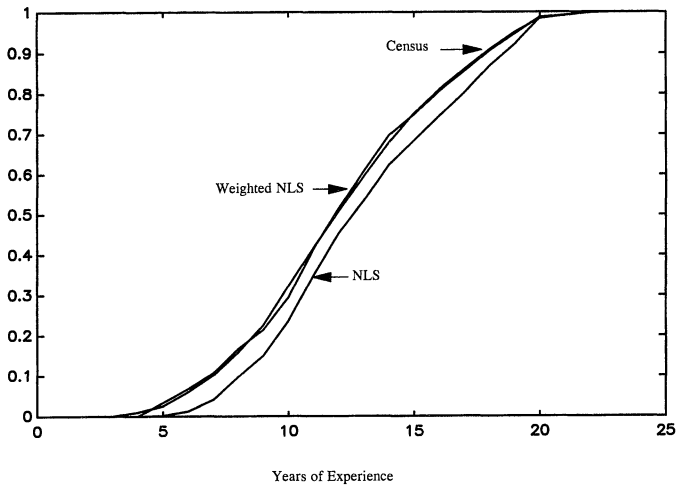


FIGURE 4.—EXPERIENCE DISTRIBUTIONS



even if the weights were not a function of the dependent variable, one would not want to throw out observations with high weights if the parameters of interest are those from the target population (the Census in this case). In the sense we make precise in section IV, excluding observations with high weights pulls the estimates away from the target population and back toward the sampled population (the NLS in this case).

The conclusion from the empirical analysis is twofold. First, the effect on returns to education (and experience) of omitting ability from wage regressions using NLS data is not well determined. While including ability in wage regressions from the NLS has a significant effect on the estimated returns to education, the magnitude of this effect gets reduced by two-thirds when we reweight the sample to make the distribution of wages, education, and experience resemble more closely that in the Census. Second, there are important differences between the Census and NLS samples (particularly the NLS sample we use here, that of men who continually responded to the survey year after year), and generalizations of estimates based solely on NLS data to the U.S. population are therefore difficult to justify.

VIII. Conclusion

In this paper we show how moment restrictions derived from auxiliary data can be taken into account when estimating regression coefficients on a primary dataset. We show that efficient estimators can be characterized as weighted versions of the estimators that would apply in the absence of moment restrictions. We investigate the interpretation of these estimators with and without making the assumption that the primary data and the aggregate data reflect the same distribution. An application of this to a wage regression controlling for ability measures—using Census estimates of moments of the wages, education, and experience distribution to form weights—yields some interesting results. Tests of the equality of moments from the NLS and 1% Census samples indicate that the two samples do not reflect the same underlying population. In addition, imposing the restrictions implied by the Census moments considerably changes the wage regression results. This implies that estimates based solely on NLS data may not be very robust and need not generalize to the population at large. By imposing the moment restrictions from the Census, the weighted regression results come "closer" than the unweighted results to those that would be obtained if ability measures were available in the Census. The sense in which this occurs is not necessarily numerical, but rather is that some moments from the weighted NLS are forced to equal the corresponding Census moments.

The methodological implications of our study are relevant for many empirical studies in the social sciences. In many of these studies, there are doubts regarding whether the dataset that is used is truly representative of the population of interest, and consequently there should be hesitation in generalizing results based on the data. The methods developed here can be used to alleviate some of these doubts by weighting the data towards a more representative sample. This may be particularly useful for studies based on longitudinal datasets, where our approach can be used to counter the effects of attrition. An example where this approach would be relevant is the comparisons between NLS, PSID, and CPS in Gottschalk and Moffitt (1992). The weighting approach developed in this paper may in such cases be an alternative to the model-based approach for attrition in, for example, Hausman and Wise (1979), especially when refreshment samples are available (Ridder, 1992).

A number of questions are not answered in this paper. First, we use just one of several different weighting schemes to impose the moment restrictions. Alternatives, such as the exponential-tilting estimator suggested in Haberman (1983) and Imbens (1997), may have different properties in small samples and with few moment restrictions even if the populations are the same. These differences may be even larger if the two populations differ, as discussed in Little and Wu (1991). A second issue is determining the number of moment restrictions to impose in the case, as in our application, where unit-level observations are available in the second dataset. Using too many restrictions may compromise the large sample results that are used for inference, while too few may leave the estimated too far from the target distribution. We could also have relaxed some of the restrictions by imposing only equality of functions of the moments. For example, one might wish to impose equality of the correlation coefficients, while allowing means and variances to differ. While that would introduce additional parameters into the model, it would fit easily into our framework.

REFERENCES

Back, K., and D. Brown, "Estimating Distributions from Moment Restrictions," Working paper, Graduate School of Business, Indiana University.

Blackburn, M., and D. Neumark, "Unobserved Ability, Efficiency Wages, and Inter-industry Wage Differentials," *Quarterly Journal of Economics* 107 (1992), 1421–1436.

Chamberlain, G., "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34 (1987), 305–334.

Cosslett, S. R., "Maximum Likelihood Estimation for Choice-based Samples," *Econometrica* 49 (1981), 1289–1316.

Deming, W. E., and F. F. Stephan, "On a Least Squares Adjustment of a Sampled Frequency When the Expected Marginal Tables are Known," *Annals of Mathematical Statistics,* 11 (1992), 427–444.

Eicker, F., "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *The Annals of Mathematical Statistics,* 34 (1963), 447–456.

Gottschalk, P. and R. Moffitt, "Earnings and Wage Distributions in the NLS, CPS and PSID," Part I of Final Report to the US Department of Labor, "Earnings Mobility and Earnings Inequality in the United States" (1992).

Griliches, Z., "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica* 45 (1977), 1–22.

Griliches, Z., B. H. Hall, and J. A. Hausman, "Missing Data and Self-Selection in Large Panels," *Annales De L'Insee* 30–31 (1978), 137–176.

Haberman, S. J., "Adjustment by Minimum Discriminant Information," *Annals of Statistics* 12 (1983), 971–988.

Hansen, L. P., "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica* 50 (1982), 1029–1054.

Hausman, J., and D. Wise, "Attrition in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica* 47 (1979), 455–473.

Imbens, G. W., "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-based Sampling," *Econometrica,* 60 (1992), 1187–1214.

——— "One-step Estimators for Over-identified Generalized Method of Moments Models," *Review of Economic Studies* 64 (1997), 359–383.

Imbens, G. W., Johnson, P., and R. H. Spady, "Information-Theoretic Approaches to Inference in Moment Condition Models," *Econometrica* 66 (1998), 333–357.

Imbens, G. W., and T. Lancaster, "Combining Micro and Macro Data in Microeconometric Models," *Review of Economic Studies* 61 (1994), 655–680.

Ireland, C. T., and S. Kullback, "Contingency Tables with Known Marginals," *Biometrika* 55 (1968), 179–188.

Keane, M., Moffitt, R., and D. Runkle, "Real Wages of the Business cycle: Estimating the Impact of Heterogeneity with Micro Data," *Journal of Political Economy* 96 (1988), 1232–1266.

Lancaster, T., "A Paradox in Choice-based Sampling," working paper, Department of Economics, Brown University (1991).

Little, R., and D. B. Rubin, *Statistical Analysis with Missing Data* (New York: Wiley, 1987).

Little, R., and M. Wu, "Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ," *Journal of the American Statistical Association* 86 (A13) (1991), 87–95.

Manski, C. F., and S. R. Lerman, "The Estimation of Choice Probabilities from Choice-based Samples," *Econometrica* 45 (1977), 1977–1988.

Mincer, J., *Schooling, Experience, and Earnings* (New York: National Bureau of Economic Research, 1974)

NLS Users' Guide 1995, Center for Human Resource Research, Ohio State University (1995).

Newey, W., and D. McFadden, "Large Sample Estimation and Hypothesis Testing," in Engle and McFadden (eds.), *The Handbook of Econometrics* (Vol. 4, New York: North-Holland, 1994).

Qin, J., and J. Lawless, "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics* 22 (1994), 300–325.

Ridder, G., "An Empirical Evaluation of Some Models for Non-random Attrition in Panel Data," *Structural Change and Economic Dynamics* 3 (1992), 337–355.

Rhoton, P., "Attrition and the National Longitudinal Surveys of Labor Market Experience: Avoidance, Control and Correction," mimeo, Center for Human Resource Research, Ohio State University (1984).

Rhoton, P., and K. Nagi, "Attrition by Wealth in the Original NLS Cohorts," Center for Human Resource Research, Ohio State University (1991).

Rubin, D. B., "Inference and Missing Data," *Biometrika* 63 (1977), 581–592.

White, H., "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48 (1980), 817–838.

APPENDIX

Proof of Theorem 1: The weights satisfy

$$N \cdot \hat{w}_n = 1/(1 + \hat{\lambda}' h(y_n, x_n))$$

where $\hat{\lambda}$, the Lagrange multiplier for the restriction $\Sigma\, w_n h(y_n, x_n) = 0$, is the solution to

$$0 = \sum_{n=1}^{N} \frac{h(y_n, x_n)}{1 + \lambda' h(y_n, x_n)}.$$

This implies that the vector $(\hat{\theta}_{\mathrm{WLS}}, \hat{\lambda})$ can be written as the solution to the system of equations

$$\sum_{n=1}^{N} \rho(y_n, x_n, \theta, \lambda) = 0 \tag{12}$$

where

$$\rho(y, x, \theta, \lambda) = \begin{pmatrix} \rho_1(y, x, \theta, \lambda) \\ \rho_2(y, x, \lambda) \end{pmatrix} = \begin{pmatrix} x \cdot (y - \theta' x)/(1 + \lambda' h(y, x)) \\ h(y, x)/(1 + \lambda' h(y, x)) \end{pmatrix}.$$

First note that $E[\rho_2(Y, X, \lambda)] = 0$ at $\lambda = 0$. This solution is unique because $E\partial\rho_2(Y, X, \lambda)/\partial\lambda < 0$. Therefore, under regularity conditions (Hansen, 1982; Newey & McFadden, 1994), $\hat{\lambda} \xrightarrow{p} 0$, and consequently $\hat{\theta}_{\mathrm{WLS}} \xrightarrow{p} \theta^*$. Second, using a second-order Taylor series expansion of $\rho(y, x, \theta, \lambda)$ around $\theta = \theta^*$ and $\lambda = 0$, and a central limit theorem for $(1/\sqrt{N})$, $\Sigma_{n=1}^{N}$ $\rho(y_n, x_n, \theta^*, 0)$ leads in a straightforward manner to the results in the theorem.

Proof of Theorem 2: We estimate w_n by maximizing $\Sigma \ln w_n$ subject to the restrictions $\Sigma\, w_n = 1$ and $\Sigma\, w_n h(y_n, x_n) = 0$. The solution can be written as

$$w_n = 1/(1 + \hat{\lambda}' h(y_n, x_n)).$$

The solution for $\hat{\lambda}$ solves

$$\max_{\lambda} \sum_{n=1}^{N} \ln (1 + \lambda' h(y_n, x_n)).$$

Assuming there is an interior solution λ_{st}^* to $\max_\lambda E[\ln (1 + \lambda' h(Y, X))]$, $\hat{\lambda}$ will converge to λ_{st}^* which therefore must satisfy

$$E_s\left[\frac{h(Y, X)}{1 + \lambda_{st}^{*\prime} h(Y, X)}\right] = 0.$$

We can still characterize the vector $(\hat{\theta}_{WLS}, \hat{\lambda})$ as the solution to the system of equations

$$\sum_{n=1}^{N} \rho(y_n, x_n, \theta, \lambda)=0$$

where

$$\rho(y, x, \theta, \lambda) = \begin{pmatrix} \rho_1(y, x, \theta, \lambda) \\ \rho_2(y, x, \lambda) \end{pmatrix} = \begin{pmatrix} x \cdot (y - \theta' x)/(1 + \lambda' h(y, x)) \\ h(y, x)/(1 + \lambda^1 h(y, x)) \end{pmatrix}.$$

Now expanding these equations around the probability limits of $\hat{\lambda}$ and $\hat{\theta}_{WLS}$ (which are λ_{st}^* and θ_{st}^* respectively), we get the desired result. It follows that $f_{st}(y, x)$ is a valid probability density function because

$$1 = \int dF_s(y, x) = \int (1 + \lambda_{st}^{*\prime} h(y, x))\, dF_{st}(y, x)$$
$$= \int dF_{st}(y, x) + \int \lambda_{st}^{*\prime} h(y, x)\, dF_{st}(y, x) = \int dF_{st}(y, x).$$

The last equality follows from the fact that

$$\int \lambda_{st}^{*\prime} h(y, x)\, dF_{st}(y, x) = \lambda_{st}^{*\prime} \int \frac{h(y, x)}{1 + \lambda_{st}^{*\prime} h(y, x)}\, dF_s(y, x) = 0.$$

Since $\hat{w}_n \geq 0$, it is also true that $f_{st}(y, x) \geq 0$, and therefore it follows that $f_{st}(y, x)$ is a valid probability density function.

Proof of Theorem 3: The consistency part follows directly from the consistency of $\hat{h}$ for h^* combined with Theorem 2. The variance/covariance matrix follows from standard GMM arguments.