TECHNICAL WORKING PAPER SERIES

IMPOSING MOMENT RESTRICTIONS FROM AUXILIARY DATA BY WEIGHTING

Guido W. Imbens
Judith K. Hellerstein

Technical Working Paper 202

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1996

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, and 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 for comments and suggestions and David Neumark for generously providing the NLS sample. We also wish to thank the NSF for financial support through grants 91-22477 and 95-11718. This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper 202
August 1996

IMPOSING MOMENT RESTRICTIONS FROM AUXILIARY DATA BY WEIGHTING

ABSTRACT

In this paper we analyze estimation of coefficients in regression models under moment restrictions where the moment restrictions are derived from auxiliary data. Our approach is similar to those that have been used in statistics for analyzing contingency tables with known marginals. These techniques are useful in cases where data from a small, potentially non-representative data set can be supplemented with auxiliary information from another data set which may be much larger and/or more representative of the target population of interest. The moment restrictions yield weights for each observation that can subsequently be used in weighted regression analysis. We discuss the interpretation of these weights both under the assumption that the target population (from which the moments are constructed) and the sampled population (from which the sample is drawn) are the same, as well as under the assumption 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 earnings, education, and experience, records data on two test scores that may be considered proxies for ability. The NLS is a small data set, 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 earnings 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 on 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.

Guido W. Imbens
Department of Economics
Littauer 117
Harvard University
Cambridge, MA 02138
and NBER
guido_imbens@harvard.edu

Judith K. Hellerstein
Department of Economics
University of Maryland
College Park, MD 20742
and NBER

1. 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.

While sampling weights that accompany longitudinal datasets in principle, at least initially, represent carefully documented stratification schemes, typically they 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 that is neither difficult to implement nor computationally burdensome.

We investigate estimation of wage regressions using the National Longitudinal Survey Young Men's Cohort (NLS) with weights derived from readily available Census records. These weights serve two purposes. First, the weights can increase precision. Second, in cases where 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 may 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

[1] For example, in a sample of twenty papers which utilize data from the NLS, we found only one reference to the use of the NLS sample weights. Keane, Moffitt and Runkle (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." It is not clear whether or not this conclusion is based on a formal test of the difference between the estimates.

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 as 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 simple wage regressions, we estimate returns to education, experience, and ability using weighted least squares.

There are important links between the weighting methods employed in this paper and various strands of the econometrics and statistics literature. First, our methods are most closely related to recent alternatives to GMM estimation based on empirical likelihood methods (Back and Brown, 1990; Imbens, 1993; Qin and Lawless, 1994; Imbens, Johnson, and Spady, 1995). 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 and Rubin, 1987). A key difference with this literature is that we do not use the unit level Census data, only averages of a particular set 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 and Stephan, 1942; Ireland and Kullback, 1968; Little and Wu, 1991), where we relax the multinomial nature of the contingency table problem and do not assume 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. While we therefore do not achieve some of the efficiency gains they report from the using auxiliary information, it aids in the interpretation of the results when target and sampled population differ.

2. 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 estimation can increase the precision of the estimates beyond that of a consistent but inefficient estimator. The second example highlights how weighting can shift the estimand from the estimand that would obtain for a large sample from the sampled population to the estimand that would obtain for a large sample from a different, target population.

EXAMPLE 1

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

$$(1) \quad \bar{\alpha} = \bar{z} = \frac{1}{N} \sum_{n=1}^{N} z_n,$$

with normalized variance $V(Z)$.

Now consider estimation of α^* given prior knowledge of $p^* = Pr(Z > 0)$. While $\bar{\alpha}$ 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 subsamples indexed by $Z > 0$ and $Z \leq 0$:

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

where $\bar{z}_1 = (\sum \delta(z_n) \cdot z_n) / \sum \delta(z_n)$ and $\bar{z}_0 = (\sum (1 - \delta(z_n)) \cdot z_n) / \sum (1 - \delta(z_n))$. 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:

$$(3) \quad \hat{\alpha} = \frac{1}{N} \sum_{n=1}^{N} w_n \cdot z_n,$$

with weights

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

In this representation $\hat{p} = \overline{\delta(z)} = \sum \delta(z_n)/N$ is the fraction of observations with $z_n > 0$. The normalized variance of the limiting distribution of $\hat{\alpha}$ is in large samples equal to $E[V(Z|\delta(Z))]$. In large samples the difference between the normalized variances of $\bar{\alpha}$ 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 sub–populations from $\hat{p}$ to p^* and $1-\hat{p}$ to $1-p^*$ respectively, and in the process leads to a more precise estimator.

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

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

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 equations

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

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

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

Instead, the researcher is assumed to have a *choice-based sample*, where with (known) probability r an observation is drawn randomly from the stratum, or subpopulation, with $y = 1$ and 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 where what is available instead is 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 by solving the weighted likelihood equations:

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

where

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

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 formally 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.

3. 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, X a vector of dimension K. The population quantity of interest, θ^*, is the vector of linear regression coefficients $E[XX']^{-1}E[XY]$. 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

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

We do not assume that the errors, $\varepsilon = Y - X'\theta^*$, are homoskedastic, and therefore the variance is the Huber (1980) 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:

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

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

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

If there are no restrictions of the form $\sum w_n h(y_n, x_n) = 0$ the weights $\hat{w}_n$ equal $1/N$ and 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:*

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

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

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

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

Given these moment functions, the standard GMM approach (Hansen, 1982; Newey and 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 (EL) estimator, which is discussed in the context of GMM problems as an alternative to the conventional two–step estimators by Back and Brown (1990), Imbens (1993), and Qin and Lawless (1994).

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

4. 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 in fact weighted estimation is typically motivated by the presumption that the population from which the sample was drawn differs from the population of interest, as in Example 2. In addition, the weights provided in the NLS 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 where target and sampled population 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, or $E_t[h(Y,X)] = 0$. 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 $\sum_n x_n(y_n - x_n'\theta) = 0$ 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

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

In addition,

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

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 show 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 expectation of $h(y,x)$ in common with the target distribution, 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

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

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 2. In this case the distribution in the target population is

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

The distribution in the sampled population is

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

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

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

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

The solution is

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

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

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

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

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

In this example the Lagrange multiplier λ^*_{st} that forces the weighted sample moment $\sum w_n y_n/N$ to match the target moment q reweighs 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 with the missing data literature emerges here. 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).

5. 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 the sampled and target distribution are not too different. When the techniques developed in this paper are applied to combinations of similarly sized datasets, or to datasets with very different distributions, however, 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 where 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 $\hat{h}(y_i, x_i)$ over a random sample of size M from the target population. Based on such a random sample the estimate $\hat{h} = \frac{1}{M}\sum_{j=1}^{M} h(y_j, x_j)$ for h^* would satisfy

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

with

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

We therefore 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 $h(y,x) = \hat{h}(y,x) - \tilde{h}$ 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 smaller dataset 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, $\tilde{X}=X/\sqrt{1+\lambda^{*\prime}_{st}(\bar{h}(Y,X)-h^*)}$,
$\tilde{\epsilon}=(Y-X'\theta^*_{st})/\sqrt{1+\lambda^{*\prime}_{st}(\bar{h}(Y,X)-h^*)}$ *and in addition* $h=\bar{h}(Y,X)-h^*$ *and* $\tilde{h}=(\bar{h}(Y,X)-h^*)/(1+\lambda^{*\prime}_{st}(\bar{h}(Y,X)-h^*))$, *we have*

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

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*

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

where

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

PROOF: See Appendix.

If M is very large relative to N, i.e. if Δ_h/k is very small, the variance is dominated by the first term in (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 $\hat{\theta}$ because of the presence of a factor involving squares of the weights in the variance formula in Theorem 3.

6. 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 14–24. 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 3438 (65.8%) of the men 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 war. (Rhoton, 1984).

There are three issues to consider 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 (1984), the original NLS sample in 1966 differed from the 1966 CPS. This suggests that the NLS need never have been representative of the U.S. population. The second is the issue 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 are interested in using the data on IQ in our empirical analysis below, this is a concern. 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 (Rothon and Nagi, 1991), age, and education (Griliches et al., 1978) which are all

relevant to 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 multi-stage 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 non-random nature of the attrition. Details on the weighting procedures are given in NLS Users' Guide (1995).

7. Returns to Schooling and Census Information

In this section we apply the above analysis to the estimation of wage equations. Economists have traditionally estimated returns to education using least squares regressions of the following type:

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

Mincer (1974) estimates these and other equations on a large sample from the CPS. We use usual weekly earnings, and interpret education as 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 815 observations of white men between

the ages of 28 and 38. Details of the sample we use are available in Blackburn and Neumark (1991). The first two sets of results in Table 1 gives the least squares estimates of the coefficients for our NLS sample. We report both unweighted "unit weight" estimates and estimates using the weights provided with the NLS.

A large literature (see Griliches [1977] for an overview) has considered the biases resulting from the possibility that there is variation in ability across individuals that makes them more likely to get schooling, and which has also an independent effect on earnings. The NLS is one of only a handful of datasets that contain measures of ability and years of 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 (10) is then

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

The first two sets of results in Table 2 give least squares estimates of (11) based on our NLS sample, again unweighted and weighted with the NLS provided weights. However, the NLS is a relatively small dataset, and estimates based on (11) using the NLS will not be nearly as precise as those based on the larger datasets such as the CPS or the Census.

The information from the Census consists of the means and variances of earnings, education, experience, and the covariances of earnings with education, experience and experience squared. They 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 non-self-employed working white men between the ages of 28–38, earning at least half of the minumum wage (to remove outliers), and working in non-farm occupations. In addition, since the NLS obtained test scores from high schools attended by the sample respondents, we selected only those men in the Census with 9 or more years of education.

Since the NLS topcodes education at 18, we did the same for our Census sample. The Census data were used to estimate moments of the joint wage, years of 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 the results from the regression without ability based on the Census data are given. This regression can also be interpreted as based on the NLS data with weights derived from the Census, because when we match on all first, second and cross moments, we exactly recover the regression based on Census data alone.

In Table 2 we report the main results, from the regression with ability measures, for the NLS alone, for the NLS with NLS sampling weights and for the NLS combined with moments from the Census. We match on thirteen moments consisting of all first, second and cross moments of the common variables log(earnings), education, experience and experience–squared. Standard errors 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 Census weighted estimates, reported in the second column of the Census weighted 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 in second to last column of Table 2 are estimated without this assumption, based on the results in Theorem 2. The third set of standard errors, in the last column of Table 2, account for 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. 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

Table 1: RETURNS TO SCHOOLING WITHOUT ABILITY MEASURES FOR NLS SAMPLE

	Unit Weights		NLS Weights		Census Weights	
Var	coeff.	s.e.	coeff.	s.e.	coeff.	s.e.
const	4.020	(0.226)	4.022	(0.223)	3.880	(0.019)
educ	0.087	(0.008)	0.087	(0.008)	0.085	(0.001)
exper	0.094	(0.026)	0.092	(0.027)	0.088	(0.002)
$exper^2$	-0.003	(0.001)	-0.002	(0.001)	-0.002	(0.000)
sample size	815		815		127,345	

Table 2: RETURNS TO SCHOOLING WITH ABILITY MEASURES FOR NLS SAMPLE

	Unit Weights		NLS Weights		Census Weights			
Var	coeff.	s.e.	coeff.	s.e.	coeff.	s.e. (1)	s.e. (2)	s.e. (3)
const	3.945	(0.241)	3.982	(0.247)	3.618	(0.188)	(0.096)	(0.104)
educ	0.056	(0.011)	0.055	(0.009)	0.071	(0.007)	(0.007)	(0.007)
exper	0.073	(0.026)	0.069	(0.027)	0.080	(0.025)	(0.008)	(0.009)
$exper^2$	-0.002	(0.001)	-0.002	(0.001)	-0.002	(0.001)	(0.000)	(0.000)
iq	0.003	(0.001)	0.003	(0.001)	0.004	(0.001)	(0.001)	(0.001)
kww	0.009	(0.003)	0.010	(0.002)	0.001	(0.003)	(0.003)	(0.003)

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 -0.092 and significantly different from zero at standard levels of significance. These results show that the NLS sample is different from the Census even after we used selection criteria in the Census (using only white males between 28 and 38 years old with at least 9 years of education and education topcoded at 18) to try to make the samples as close as possible. Moreover, the sampling weights in the NLS do not reweight the sample to reflect the Census.

A 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. Taking into account the sampling error in the Census

moments does not substantially change this.

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 fact that that the Census is 165 times larger than the NLS sample, and to the fact 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.

The significance of the differences between estimates based solely on the NLS and those incorporating Census information is investigated directly in Tables 3–5. There are a number of ways of testing the restrictions implied by the Census information. The results of two of these methods are given in Table 3; the results of a third method are in Table 4. Imbens, Johnson and Spady (1995) 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. This is done in the first set of columns in Table 3 which reports 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 distribution are equal.

A third way to investigate the difference between the Census and NLS sample is to consider directly the difference in parameter estimates. In Table 4 we present the differences in the unit weight and census weighted estimates, with associated standard errors. These standard errors do not assume that target and sampled distribution are equal, and take into account the sampling error in the Census moments. For comparison purposes we also report

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	1.20	(0.51)	2.3
log(earn)	0.18	(0.02)	12.0	0.42	(1.79)	0.3
exper	0.86	(0.13)	6.5	5.73	(1.70)	3.4
$educ^2$	-10.65	(2.21)	-4.8	-0.01	(0.01)	-0.6
$log(earn)^2$	2.04	(0.18)	11.4	0.05	(0.12)	0.4
$exper^2$	21.17	(3.74)	5.7	-0.44	(0.14)	-3.1
educ × log(earn)	0.43	(0.57)	0.7	-0.02	(0.05)	-0.4
exper × log(earn)	7.33	(0.81)	9.0	-0.02	(0.17)	-0.1
$exper^2$ × log(earn)	155.60	(22.36)	7.0	0.00	(0.01)	0.3
educ × exper	8.47	(1.46)	5.8	-0.16	(0.06)	-2.7
educ × $exper^2$	246.67	(41.98)	5.9	0.01	(0.00)	2.8
$exper^3$	429.69	(83.98)	5.1	0.01	(0.01)	2.1
$exper^4$	8.34×10^3	(1.77×10^3)	5.0	-0.00	(0.00)	-1.6
chi–square tests (d.o.f.)	325.9 (13)			164.6 (13)		

the raw estimates from Table 2 again. The last row presents a test statistic for the null–hypothesis that the unit weight and Census weighted estimates are equal. The test statistic has, under the null, a Chi–squared distribution with six degrees of freedom.

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 Tables 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 earnings, education and experience was also reported in 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 in particular average earnings differ considerably between 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 in the two samples to be

Table 4: TESTS OF EQUALITY OF UNIT WEIGHT AND CENSUS WEIGHTED ESTIMATES

	Unit Weights	Census Weighted	Difference		
	est.	est	dif	s.e.	t–stat
const.	3.945	3.618	0.327	(0.284)	1.1
educ.	0.056	0.071	-0.015	(0.013)	-1.2
exper.	0.073	0.080	-0.007	(0.027)	-0.3
exper.^{2}	-0.002	-0.002	0.000	(0.000)	-0.3
iq	0.003	0.004	-0.001	(0.002)	-0.3
kww	0.009	0.001	0.008	(0.005)	1.5
chi–square test (d.o.f.)				142.6 (6)	

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 0.087–0.056=0.031, or that three percentage points of the estimated return of 8.7% is due to ability bias. If we change the distribution of earnings, education and experience to be closer to that of the Census, the estimate of the effect of ability bias is 0.085–0.071=0.014, less than half the decrease found using only the NLS data.

The weighted interpretation of the new estimator makes it clear that the key difference between the No Weight and Census weighted estimates is the difference across the two samples in the distribution of earnings, experience, and education. Since the ability measures are not independent of these three variables, the regression estimates differ considerably depending on which earnings, 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 4. It is of interest to investigate some aspects of this distribution. In Figures 2–4 we present estimates of the distribution

function of education and the logarithm of weekly earnings for the Census distribution, the NLS distribution and the Census–weighted NLS distribution. For all of the distributions, it clear that by forcing the Census–weighted NLS distribution to have the same mean and variance of log earnings 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 earnings are clearly under–represented in the NLS relative to the Census, as are, to a lesser extent, men with high levels of education.

We can also see these effects by inspecting the weights directly. In Table 5 we present the values of all variables for the observations with the highest and lowest weights. It is apparent from this table that segment of the Census population under–represented in the NLS consists of men with relatively high earnings, somewhat low education and higher than average experience. Their average IQ is approximately equal to the NLS average of 103, but their KWW is considerably higher than the NLS average of 37. Because the weights do not directly depend on IQ and KWW this assocation between the weights and KWW stems from the association between KWW and the variables used in the construction of the weights.

Table 5: OBSERVATIONS WITH FIVE HIGHEST AND LOWEST CONSTRUCTED WEIGHTS

weights	educ	log(earn)	exper	age	iq	KWW
26.3	18	5.56	5	29	107	30
15.4	17	4.84	13	36	106	34
13.7	12	4.06	11	29	108	46
7.75	10	4.70	16	32	103	38
7.11	16	5.52	6	28	125	34
0.42	12	6.48	20	28	95	48
0.40	11	6.62	21	38	115	40
0.39	13	7.60	13	32	108	41
0.37	13	7.60	16	35	120	50
0.30	12	7.31	20	38	82	41

The conclusion from this 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, this effect gets cut in half when we reweight the sample to make the distribution of earnings, education and experience resemble more closely that in the Census. Second, there are important differences between the Census and NLS samples, and generalizations of estimates based solely on NLS data to the U.S. population are therefore difficult to justify.

8. 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 earnings, education and experience distribution 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 changes the wage regression results considerably. 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 obtain if ability measures were available in the Census. The sense in which this occurs is that some moments from the weighted NLS are set equal to 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 about whether the dataset 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 possible to impose the moment restrictions. Alternatives, such as the exponential tilting estimator suggested in Haberman (1983) and Imbens (1993), may have different properties in small samples and with few moment restrictions even if the populations are the same, and these differences are likely to be larger. 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 which are used for inference, while too few may leave the estimand 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. We intend to address these issues in further research.

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 $\sum w_n h(y_n, x_n) = 0$, 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 $(\hat{\theta}_{\text{WLS}}, \hat{\lambda})$ can be written as the solution to the system of equations

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

where

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

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 and McFadden, 1994), $\hat{\lambda} \xrightarrow{p} 0$ and consequently $\hat{\theta}_{\text{WLS}} \xrightarrow{p} \theta^*$. 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 $(1/\sqrt{N}) \sum_{n=1}^{N} \rho(y_n, x_n, \theta^*, 0)$ leads in a straightforward manner to the results in the theorem. □

Proof of Theorem 2:

We estimate w_n by maximizing $\sum \ln w_n$ subject to the restrictions $\sum w_n = 1$ and $\sum w_n h(y_n, x_n) = 0$. 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 $(\hat{\theta}_{\text{WLS}}, \hat{\lambda})$ as the solution to the system of equations

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

where

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

Now expanding these equations around the probability limits of $\hat{\lambda}$ and $\hat{\theta}_{\text{WLS}}$, 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. □

Figure 1: NLS and Census Weights

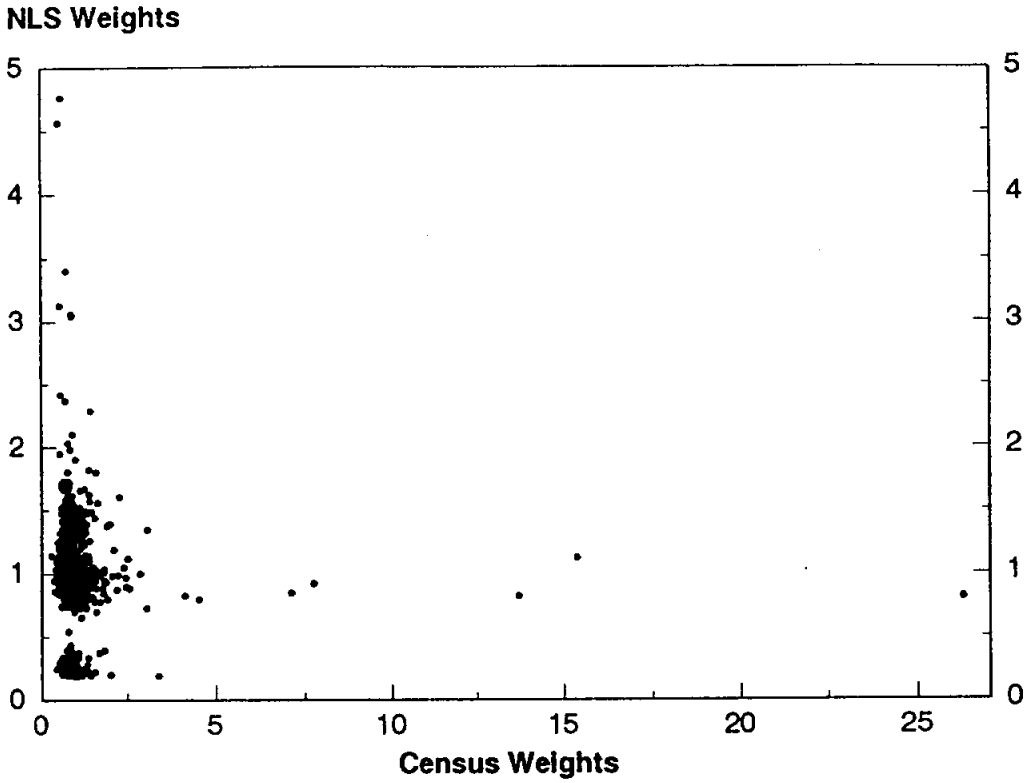


Figure 2: Wage Distributions

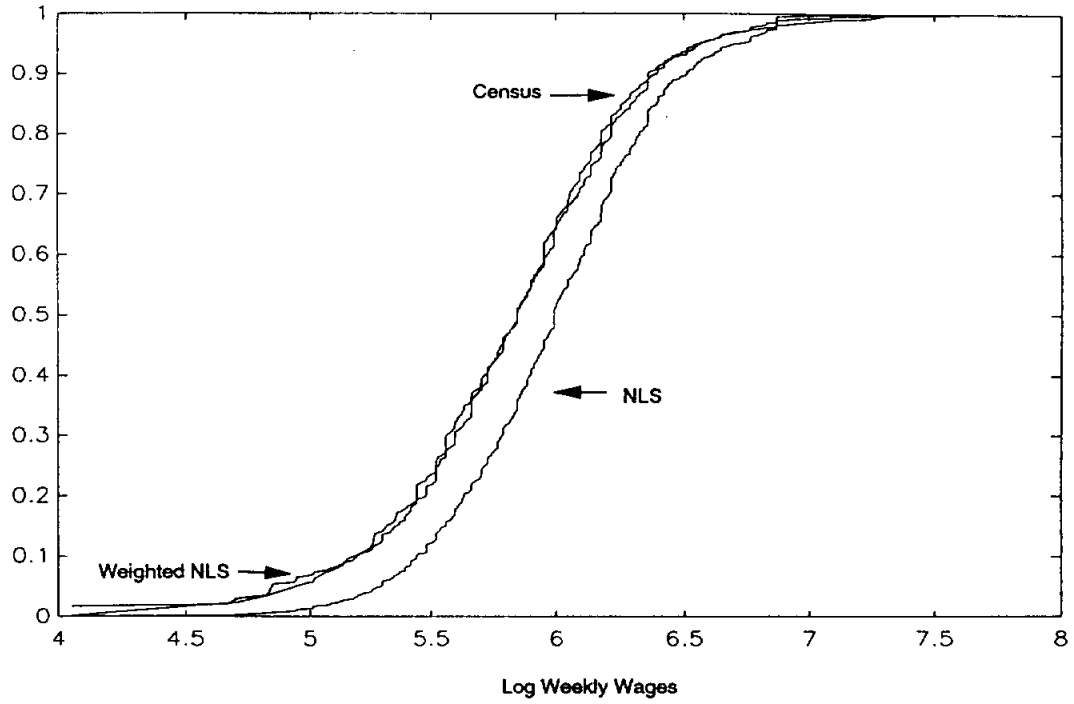


Figure 3: Education Distributions

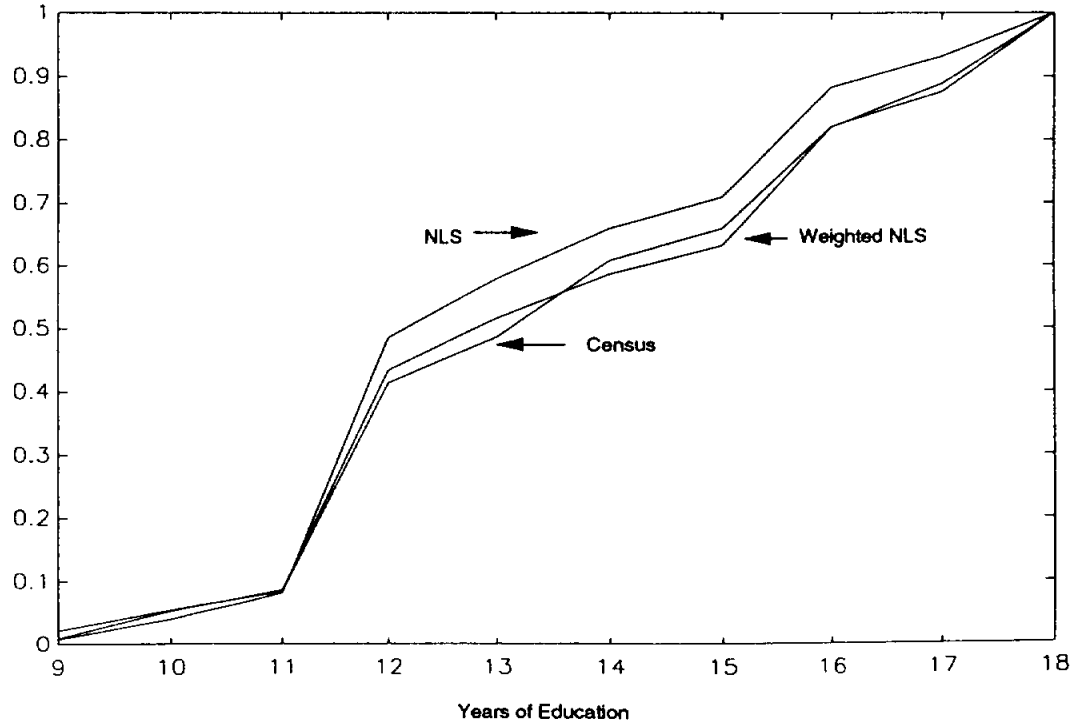
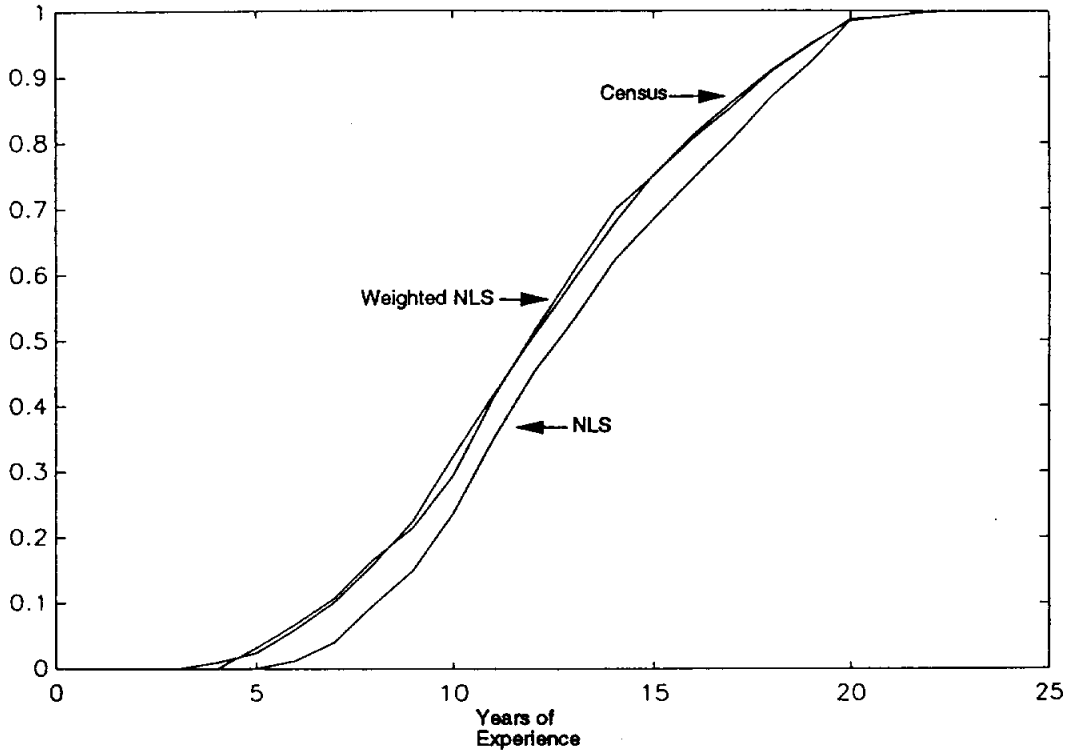


Figure 4: Experience Distributions



REFERENCES

BACK, K., AND D. BROWN, (1990), "Estimating Distributions from Moment Restrictions", working paper, Graduate School of Business, Indiana University.

BLACKBURN, M., AND D. NEUMARK, (1992), "Unobserved Ability, Efficiency Wages, and Inter-industry Wage Differentials," *Quarterly Journal of Economics*, Vol. 107, No 4, 1421-36.

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

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

DEMING, W. E., AND F. F. STEPHAN, (1942), "On a Least Squares Adjustment of a Sampled Frequency When the Expected Marginal Tables are Known," *Annals of Mathematical Statistics*, Vol 11, 427-444.

GOTTSCHALK, P. AND R. MOFFITT, (1992), "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".

GRILICHES, Z. (1977) "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, vol 45, no 1, 1-22.

GRILICHES, Z., B. H. HALL, AND J.A. HAUSMAN (1978) "Missing Data and Self-Selection in Large Panels" *Annales De L'Insee*, no 30-31.

HABERMAN, S. J., (1983), "Adjustment by Minimum Discriminant Information", *Annals of Statistics*, Vol. 12, no 3, 971-988.

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

HAUSMAN, J., AND D. WISE, (1979), "Attrition in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*.

HUBER, P. J., (1980), *Robust Statistics*, Wiley, New York.

IMBENS, G. W., (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-based Sampling," *Econometrica*, vol. 60, no. 5, 1187-1214.

IMBENS, G. W. (1993), "A New Approach to Generalized Method of Moments Estimation," Harvard Institute of Economic Research Working Paper 1633.

IMBENS, G.W., JOHNSON, P., AND R.H. SPADY, (1995), "Information-Theoretic Approaches to Inference in Moment Condition Models," NBER technical working paper.

IMBENS, G. W., AND T. LANCASTER, (1994), "Combining Micro and Macro Data in Microeconometric Models," *Review of Economic Studies.*

IRELAND, C. T., AND S. KULLBACK, (1968), "Contingency Tables with Known Marginals," *Biometrika*, 55, 179–188.

KEANE, M., MOFFIT, R, AND D. RUNKLE, (1988), "Real Wages of the Business cycle: Estimating the Impact of Heterogeneity with Micro Data," *Journal of Political Economy.*

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

LITTLE, R., AND D. B. RUBIN, (1987), *Statistical Analysis with Missing Data*, New York, Wiley.

LITTLE, R., AND M. WU, (1991), "Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ, *Journal of the American Statistical Association*, Vol 86, no 413, 87–95.

MANSKI, C. F., AND S. R. LERMAN, (1977), "The Estimation of Choice Probabilities from Choice-based Samples," *Econometrica*, vol 45, 1977-1988.

MINCER, J., (1974), *Schooling, Experience, and Earnings*, New York, NBER.

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

NEWEY, W., AND D. MCFADDEN, (1994), "Large Sample Estimation and Hypothesis Testing," in Engle and McFadden (eds.), *The Handbook of Econometrics*, Vol. 4.

RIDDER, G., (1992), "An Empirical Evaluation of Some Models for Non–random Attrition in Panel Data," *Structural Change and Economic Dynamics*, Vol 3, no 2, 337–355.

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

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

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

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