Asymptotic Properties of the Weighted Least Squares Estimator Under Moments Restriction

Bayram, Deniz and Dayé, Modeste

University of Namur

2014

Asymptotic Properties of the Weighted Least Squares Estimator Under Moments Restriction*

July 2014

Deniz Bayram Modeste Dayé

1 Introduction

The aim of this work is to review the paper by Hellerstein & Imbens (1982) focusing on the use of auxiliary data and a formal derivation of the asymptotic properties of the underlying Weighted Least Squares estimator. Hellerstein & Imbens had introduced very broadly a GMM model, based on empirical likelihood estimators. The necessity for the specific framework in this paper arises for the purposes of data set combination. The essential idea is to include auxiliary information in the form of weights in the regression analysis of the primary data.

The regression on the primary data is estimated by means of weighted least squares. The set of weights is based on the auxiliary data and estimated via empirical likelihood method. These weights are constructed such that certain moments from the weighted primary data are equal to the corresponding moments derived from the auxiliary data.

The literature for data set combination is huge, however it can be very broadly classified into two, based on its fundamental objectives. The first class of models including Angrist and Krueger (1992) and Arellano and Meghir (1992) combine different data sets in order confront the situation that there is no single data set that contains all variables of interest (Ridder and Moffitt, 2007).

In the second class of models including Imbens and Lancaster (1994) and this paper, generally, the purpose of weights is to increase the precision of the estimates and achieve

*Based on Hellerstein & Imbens (1982).

asymptotic efficiency gains. For this purpose a small sample with all relevant variables is typically combined with a larger sample with fewer variables. Another purpose of weights may be shifting the distribution of the primary sample towards the auxiliary sample distribution. In this report, we focus on the first theorem which assumes that population from which the primary sample was drawn and the population from which the moments are constructed are the same. However if two populations differ and the latter is the population of interest, this shift would be necessary.

Although we focus in this report on Theorem 1, the paper gives also extensions of the model by relaxing two assumptions. Theorem 2 gives the large sample results for the case where population from which the sample was drawn (sampled population) is different from the population of interest (target population). Theorem 3 allows for sampling error in the estimation of the moments from the auxiliary data.

The paper also presents an application of the model on Mincer wage equations. Returns to schooling are estimated using NLS data set, which contains data of IQ and ability test scores. The NLS is one of the few data sets which contain measures that can account for the well-known omitted variable bias, however the size of the data set is small. In order to increase precision, they use the moments from Census data and construct weights for the weighted wage regressions of the NLS data.

2 Derivation of the estimator

Let's consider a random sample $Z = (Y, X)$ with an unknown probability density function $f(x, y)$ and assume that we have knowledge on

$$\mathbf{h}^* = \mathbf{E}[\bar{\mathbf{h}}(\mathbf{Y}, \mathbf{X})] = \int \bar{\mathbf{h}}(\mathbf{Y}, \mathbf{X})\mathbf{dF}(\mathbf{X}, \mathbf{Y}) \tag{1}$$

which is the expectation of an R dimensional function of X and Y in the same population. From this setting, we can write the following moment restriction

$$E[h(Y, X)] = 0 \quad \text{with} \quad h(Y, X) = \bar{h}(Y, X) - h^* \ .$$

Let's now consider the following linear regression

$$\boldsymbol{w_n y_n = w_n x_n^{'} \theta + \epsilon_n \quad n = 1, ..., N} \ . \tag{2}$$

where w_n is the weight put on observation n and is a scalar . Using the strict exogeneity condition of the explanatory variables $E[x_n \epsilon_n] = 0$, we can derive the Weighted Least Squares (WLS) estimator as follows:

$$\begin{aligned} E[x_n \epsilon_n] &= 0 \\ E[x_n(w_n y_n - w_n x_n^{'} \theta)] &= 0 \\ E[w_n x_n y_n] &= E[w_n x_n x_n^{'}]\theta \end{aligned}$$

Hence, under invertibility of $E[w_n x_n x_n^{'}]$, taking the sample analogue of each side and solving for θ we get

$$\boldsymbol{\hat{\theta}_{WLS} = \Big[\sum_{n=1}^{N} \hat{w}_n x_n x_n^{'} \Big]^{-1} \Big[\sum_{n=1}^{N} \hat{w}_n x_n y_n \Big]} \ . \tag{3}$$

It is worth noting that in our linear regression, the sample analogue of $E[w_n x_n y_n] = E[w_n x_n x_n^{'}]\theta$ (using the strict exogeneity as above) is nothing else than the first order condition of

$$\min_{\theta} \sum_{n=1}^{N} (w_n y_n - w_n x_n{'} \theta)^2 \ .$$

We will come back to this later in section 3.

The weights $\hat{w}_n$ in (3) are estimated optimally by the Empirical Likelihood (EL) method which will basically consist of solving a maximization program of the sum of the log of the weights, given that those weights add up to one and the weighted moment restrictions are fulfilled.

$$\max_{w_n} \sum_{n=0}^{N} \ln w_n \quad \text{s.t.} \quad \sum_{n=1}^{N} w_n = 1 \quad \text{and} \quad \sum_{n=1}^{N} w_n h(y_n, x_n) = 0 \tag{4}$$

The Lagragian of problem (4) is given by

$$\mathcal{L}(\alpha, \lambda, \theta) = \sum_{n=1}^{N} \ln w_n + \alpha(1 - \sum_{n=1}^{N} w_n) - N\lambda' \sum_{n=1}^{N} w_n h(y_n, x_n) \tag{5}$$

where $\lambda = (\lambda_1, ..., \lambda_R)'$ is a R dimensional Lagrange multiplier vector attached to the moment restrictions and α is a scalar Lagrange multiplier attached to the constraint on the sum of the weights.

Taking the first order condition (FOC) with respect to w_n gives:

$$\frac{1}{w_n} - \alpha - \lambda' N h(y_n, x_n) = 0 \tag{6}$$

Multiplying (6) with w_n and applying summation over all terms, we get $\alpha = N$ given that $\sum_{n=1}^{N} w_n h(y_n, x_n) = 0$. Using this value of α we can rewrite (6) as follows:

$$\begin{aligned} & \frac{1}{w_n} - N - N\lambda' h(y_n, x_n) = 0 \\ \text{and} \quad & \frac{1}{w_n} = N(1 + \lambda' h(y_n, x_n)) = 0 \end{aligned}$$

Hence we obtain

$$N w_n = \frac{1}{(1 + \lambda' h(y_n, x_n))} \tag{7}$$

In order to obtain $\hat{w}_n$, an estimate of w_n the optimal weights, we need an estimate of λ, the Lagrange multiplier vector. Applying $\sum_{n=1}^{N} h(y_n, x_n)$ to (7) and using the fact $\sum_{n=1}^{N} w_n h(y_n, x_n) = 0$ (second constraint in (4)) , we can get :

$$\mathbf{0} = \sum_{\mathbf{n=1}}^{\mathbf{N}} \frac{\mathbf{h(y_n, x_n)}}{\mathbf{(1 + \lambda' h(y_n, x_n))}} \tag{8}$$

Consequently $\hat{\lambda}$ solves (8) and using this estimate, the estimated optimal weight in (7) reads

$$\hat{w}_n = \frac{1}{N(1 + \hat{\lambda}' h(y_n, x_n))} \tag{9}$$

Therefore in a certain sense, all the previous steps consists of estimating the weights optimally, which requires an estimate of the Lagrange multiplier in order to fully get $\hat{\theta}_{WLS}$ could be summarized by just transforming the constrained problem in (4) into an unconstrained one using (7). In fact, noting that w_n is a decreasing function of λ and so does $_n$, plugging (7) into the objective function in (4) we can successively write:

$$\max_{w_n} \sum_{n=1}^{N} \ln w_n = \min_{\lambda} \sum_{n=1}^{N} \ln \frac{1}{N(1 + \lambda' h(y_n, x_n))} \tag{10}$$

$$= \min_{\lambda} \sum_{n=1}^{N} -ln[N(1 + \lambda' h(y_n, x_n))] \tag{11}$$

Here, neglecting N which is a constant and can always be normalized, the unconstrained problem in (11) is equivalent to:

$$\min_{\lambda} \sum_{n=1}^{N} -\ln[1 + \lambda' h(y_n, x_n)]. \tag{12}$$

From (12), it is easy to see that the FOC yields exactly the equation in (8) which gives $\hat{\lambda}$ and implicitly $\hat{w}_n$ from (7) and $\hat{\theta}$ from (3).

The whole problem of the determination of $\hat{\theta}$ can be summarized by the following system of equations:

$$\begin{cases} \sum_{n=1}^{N} x_n \epsilon_n = 0 \\ \sum_{n=1}^{N} \dfrac{h(y_n, x_n)}{1 + \lambda' h(y_n, x_n)} = 0 \end{cases} \tag{13}$$

In this system, $\epsilon_n = w_n y_n - w_n x_n^{'} \theta = \dfrac{1}{1 + \lambda' h(y_n, x_n)}(y_n - x_n^{'}\theta)$ with w_n as it is found in (7). The first equation of the system comes from the strict exogeneity condition and the second equation comes from (8). The solution to the system (13) is thus $\{\hat{\theta}_{WLS}, \hat{\lambda}\}$.

3 The Theorem and the Proof

The theorem we are interested in (Theorem 1 in the paper) states the consistency and the asymptotic normality of the Weighted Least Squares (WLS) estimator derived above.

3.1 Theorem

Given regularity conditions, the estimator $\hat{\theta}_{WLS}$ of θ^* has the following asymptotic properties:

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

and

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

In the following sections we will establish the consistency and the asymptotic normality of $\hat{\theta}_{WLS}$.

3.2 Proof

3.2.1 Consistency of $\hat{\theta}_{WLS}$

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

Here we basically have to check whether the regularity conditions for consistency holds for the estimator derived. From what we have said on the system (13) in the previous section , let's posit

$$\boldsymbol{\rho(y,x,\theta,\lambda)} = \begin{pmatrix} \dfrac{\boldsymbol{x(y-x'\theta)}}{\boldsymbol{1+\lambda' h(y,x)}} \\ -\dfrac{\boldsymbol{h(y,x)}}{\boldsymbol{1+\lambda' h(y,x)}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\rho_1(y,x,\theta,\lambda)} \\ \boldsymbol{\rho_2(y,x,\theta,\lambda)} \end{pmatrix} \tag{1}$$

The ρ's are simply the underlying (moment) functions from the system (13). Moreover, setting $\sum_{n=1}^{N} \rho(y_n, x_n, \theta, \lambda) = 0$ given (13) generates the FOC of the following program

$$\min_{\boldsymbol{\theta,\lambda}} \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \begin{pmatrix} \boldsymbol{(w_n y_n - w_n x_n{}'\theta)^2} \\ \boldsymbol{-ln[1+\lambda' h(y_n,x_n)]} \end{pmatrix} = \min_{\boldsymbol{\theta,\lambda}} \boldsymbol{Q_n(\theta,\lambda)} \tag{2}$$

where equation (7) in section 2 defines the relationship between w_n and λ; and $Q_n(\theta, \lambda)$ is defined as our sample criterion function.

Let's then define the population criterion function as follows:

$$Q(\theta,\lambda) = \mathbb{E}\left[\begin{pmatrix} (w_n y_n - w_n x_n{}'\theta)^2 \\ -ln[1+\lambda' h(y_n,x_n)] \end{pmatrix} \right] = \mathbb{E}[m(y_n,x_n,\theta,\lambda)] \tag{3}$$

From the program in (2) the FOC are indeed given by $\sum_{n=1}^{N} \rho(y_n, x_n, \theta, \lambda) = 0$.
The high level assumptions for consistency are first the identification of the true parameter

of interest, θ^* and second, the uniform weak convergence of the sample criterion function $Q_n(\theta, \lambda)$ to the population criterion function $Q(\theta, \lambda)$.

$\Longrightarrow$ **Regularity conditions for well-separatedness and identification of θ^***

Let's assume that the parameter space Θ is closed and bounded (compact) in $\mathbb{R}^R$. We have that $m(z_n, \theta, \lambda)$ in (3) where $z_n = (y_n, x_n)$, is continuous in θ and λ, $\forall\ z_n$ such that $\lambda' h(y_n, x_n) > -1$. Moreover, assuming $E \sup_{(\theta,\lambda)\in\Theta} \left\| m(z_n, \theta, \lambda) \right\| < \infty$, we can claim that $Q(\theta, \lambda)$ is continuous. Hence two of the primitive conditions for identification are verified. Concerning the third primitive condition i.e. whether the population criterion function $Q(\theta, \lambda)$ is uniquely minimized, we have to check using (1), where setting

$$\mathbb{E}\left[\begin{pmatrix} \rho_1(y, x, \theta, \lambda) \\ \rho_2(y, x, \theta, \lambda) \end{pmatrix} \right] = \mathbb{E}\left[\begin{pmatrix} \dfrac{x(y - x'\theta)}{1 + \lambda' h(y, x)} \\ -\dfrac{h(y, x)}{1 + \lambda' h(y, x)} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{4}$$

gives the FOC from minimizing the population criterion function (see (1) and (3)).

From (1), (3) and (4), we can see that $E[\rho_2(y, x, \theta, \lambda)] = 0$ for a unique value which is $\lambda = 0$, given that $E[h(y, x)] = 0$. Moreover $E[\partial\rho_2(x, y, \lambda)/\partial\lambda] = \dfrac{hh'}{(1 + \lambda' h)^2} > 0$ (always positive definite) and thus also at $\lambda = 0$, it clearly appears that the population criterion function is uniquely minimized at $\lambda = 0$ (the true value of λ) through ρ_2. Besides, given the equation (7) in section 2, we can also claim that the true values of λ and θ, say, $0 and \theta^*$ uniquely minimize the population criterion function in (3) through the first order conditions ρ_1 (given the invertibility of $E[xx']^{-1}$ assumed earlier) for θ and ρ_2 for λ.

Hence we have checked the primitive conditions for identification of (θ^*, λ).

Now we will establish the uniform weak convergence.

$\Longrightarrow$ **Regularity conditions for uniform weak convergence of the sample criterion function**

We have some independently and identically distributed data $z_n = (y_n, x_n)$. We have assumed above that Θ is compact, $E\sup_{(\theta,\lambda)\in\Theta}\left\|m(z_n,\theta,\lambda)\right\| < \infty$ and shown that $m(z_n,\theta,\lambda)$ is continuous. So the stochastic equicontinuity of $Q_n(\theta,\lambda)$ holds and $E[m(y_n,x_n,\theta,\lambda)$ is continuous in (θ,λ) at the true values of the parameters $(\theta^*,0)$. Hence the uniform weak convergence of $Q_n(\theta,\lambda)$ follows.

With this, we have established the high level assumptions for consistency of (θ,λ). Thus, we conclude:

$$\hat{\lambda} \xrightarrow{p} 0$$
$$\text{and} \quad \hat{\theta}_{WLS} \xrightarrow{p} \theta^*$$

In the following section we will establish the asymptotic normality of $\hat{\theta}_{WLS}$.

3.2.2 Asymptotic normality of $\hat{\theta}_{WLS}$

We have shown previously that $\hat{\theta}_{WLS} \xrightarrow{p} \theta^*$. Now let's take a first order Taylor series expansion of

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

around $(\theta,\lambda) = (\theta^*,0)$ and neglect the remainder

We have,

$$\rho(y,x,\theta,\lambda) = \begin{pmatrix} x\epsilon \\ h \end{pmatrix} - \begin{pmatrix} xx' & x\epsilon h' \\ 0 & hh' \end{pmatrix} \begin{pmatrix} \theta-\theta^* \\ \lambda \end{pmatrix} \tag{5}$$

Taking summation over both sides of (5) and multiplying by $1/N$, given the equation (13) in section 2, we obtain

$$0 = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} x\epsilon \\ \frac{1}{N}\sum_{n=1}^{N} h \end{pmatrix} - \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} xx^{'} & \frac{1}{N}\sum_{n=1}^{N} x\epsilon h^{'} \\ 0 & \frac{1}{N}\sum_{n=1}^{N} hh^{'} \end{pmatrix} \begin{pmatrix} \theta - \theta^{*} \\ \lambda \end{pmatrix} \tag{6}$$

Since the matrix multipliying the parameter vector in (6) is positive definite and thus invertible we get

$$\begin{pmatrix} \theta - \theta^{*} \\ \lambda \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} xx^{'} & \frac{1}{N}\sum_{n=1}^{N} x\epsilon h^{'} \\ 0 & \frac{1}{N}\sum_{n=1}^{N} hh^{'} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} x\epsilon \\ \frac{1}{N}\sum_{n=1}^{N} h \end{pmatrix} \tag{7}$$

Multiplying both sides of (7) by $\sqrt{N}$ reads,

$$\sqrt{N}\begin{pmatrix} \theta - \theta^{*} \\ \lambda \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} xx^{'} & \frac{1}{N}\sum_{n=1}^{N} x\epsilon h^{'} \\ 0 & \frac{1}{N}\sum_{n=1}^{N} hh^{'} \end{pmatrix}^{-1} \sqrt{N}\begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} x\epsilon \\ \frac{1}{N}\sum_{n=1}^{N} h \end{pmatrix} \tag{8}$$

Using (8), we can apply the weak law of large numbers for each element of the inverted matrix as follows:

$$\frac{1}{N}\sum_{n=1}^{N} xx^{'} \xrightarrow{p} \mathbb{E}[xx^{'}]; \quad \frac{1}{N}\sum_{n=1}^{N} x\epsilon h^{'} \xrightarrow{p} \mathbb{E}[x\epsilon h^{'}] \quad \text{and} \quad \frac{1}{N}\sum_{n=1}^{N} hh^{'} \xrightarrow{p} \mathbb{E}[hh^{'}]$$

Moreover, by applying central limit theorem applied to the second term of the right hand side of (8), we obtain

$$\sqrt{N}\begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} x\epsilon \\ \frac{1}{N}\sum_{n=1}^{N} h \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V\begin{pmatrix} x\epsilon \\ h \end{pmatrix}\right)$$

We have $\frac{1}{N}\sum_{n=1}^{N} x\epsilon \xrightarrow{p} 0$ and $\frac{1}{N}\sum_{n=1}^{N} h \xrightarrow{p} 0$ due to, respectively, the strict exogeneity condition on the explanatory variables and the moment on function h(see section 2).

We should note that in this subsection and elsewhere, 0 should be seen as a vector of scalar zeros since h , x and θ are of dimension R as emphasized before .

Let us then rewrite (8), given the two previous results we got from (7) and using Slutsky theorem as

$$\sqrt{N}\begin{pmatrix}\theta-\theta^{*}\\ \lambda\end{pmatrix}\xrightarrow{d}\begin{pmatrix}\mathbb{E}[xx^{'}] & \mathbb{E}[x\epsilon h^{'}]\\ 0 & \mathbb{E}[hh^{'}]\end{pmatrix}^{-1}\mathcal{N}\left(\mathbf{0},V\begin{pmatrix}x\epsilon\\ h\end{pmatrix}\right) \tag{9}$$

First, taking the inverse of the triangular block matrix, we get

$$\begin{pmatrix}\mathbb{E}[xx^{'}] & \mathbb{E}[x\epsilon h^{'}]\\ 0 & \mathbb{E}[hh^{'}]\end{pmatrix}^{-1}=\begin{pmatrix}\mathbb{E}[xx^{'}]^{-1} & -\mathbb{E}[xx^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}]\mathbb{E}[hh^{'}]^{-1}\\ 0 & \mathbb{E}[hh^{'}]^{-1}\end{pmatrix} \tag{10}$$

Second,

$$\begin{aligned}V\begin{pmatrix}x\epsilon\\ h\end{pmatrix}&=\mathbb{E}\left[\begin{pmatrix}x\epsilon\\ h\end{pmatrix}\begin{pmatrix}\epsilon x^{'} & h^{'}\end{pmatrix}\right]\\ &=\mathbb{E}\left[\begin{pmatrix}\epsilon^{2}xx^{\prime} & x\epsilon h^{'}\\ h\epsilon x^{'} & hh^{'}\end{pmatrix}\right]\\ &=\begin{pmatrix}\mathbb{E}[\epsilon^{2}xx^{'}] & \mathbb{E}[x\epsilon h^{'}]\\ \mathbb{E}[h\epsilon x^{'}] & \mathbb{E}[hh^{'}]\end{pmatrix}\end{aligned} \tag{11}$$

Using the results from (9) , (10) and (11), we can write

$$\sqrt{N}\begin{pmatrix}\theta-\theta^{*}\\ \lambda\end{pmatrix}\xrightarrow{d}\begin{pmatrix}\mathbb{E}[xx^{'}]^{-1} & -\mathbb{E}[xx^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}]\mathbb{E}[hh^{'}]^{-1}\\ 0 & \mathbb{E}[hh^{'}]^{-1}\end{pmatrix}\mathcal{N}\left(\mathbf{0},\begin{pmatrix}\mathbb{E}[\epsilon^{2}xx^{'}] & \mathbb{E}[x\epsilon h^{'}]\\ \mathbb{E}[h\epsilon x^{'}] & \mathbb{E}[hh^{'}]\end{pmatrix}\right) \tag{12}$$

From (12) it easy to derive the asymptotic normality of each of the parameters θ and λ. For what we are interested in, we can see that at $\theta=\hat{\theta}_{WLS}$, using the Slutsky theorem,

$\sqrt{N}(\hat{\theta}_{WLS} - \theta^*)$ converges to a normal distribution with mean zero and asymptotic variance $Avar(\hat{\theta}_{WLS})$ which is derived successively as follows from (12) as:

$$\begin{aligned}
&\Big(\mathbb{E}[xx^{'}]^{-1} - \mathbb{E}[xx^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}][hh^{'}]^{-1}\Big) \begin{pmatrix} \mathbb{E}[\epsilon^2 xx^{'}] & \mathbb{E}[x\epsilon h^{'}] \\ \mathbb{E}[h\epsilon x^{'}] & \mathbb{E}[hh^{'}] \end{pmatrix} \begin{pmatrix} \mathbb{E}[xx^{'}]^{-1} \\ -\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}]\mathbb{E}[xx^{'}]^{-1} \end{pmatrix} \\
&= \mathbb{E}[xx^{'}]^{-1}\Big(I - \mathbb{E}[x\epsilon h^{'}][hh^{'}]^{-1}\Big) \begin{pmatrix} \mathbb{E}[\epsilon^2 xx^{'}] & \mathbb{E}[x\epsilon h^{'}] \\ \mathbb{E}[h\epsilon x^{'}] & \mathbb{E}[hh^{'}] \end{pmatrix} \begin{pmatrix} I \\ -\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}] \end{pmatrix} \mathbb{E}[xx^{'}]^{-1} \\
&= \mathbb{E}[xx^{'}]^{-1} \Big(\mathbb{E}[\epsilon^2 xx^{'}] - \mathbb{E}[\epsilon xh^{'}]\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[\epsilon hx^{'}] \quad \mathbb{E}[\epsilon xh^{'}] - \mathbb{E}[\epsilon xh^{'}]\Big) \begin{pmatrix} I \\ -\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}] \end{pmatrix} \times \\
&\qquad \mathbb{E}[xx^{'}]^{-1} \\
&= \mathbb{E}[xx^{'}]^{-1} \Big(\mathbb{E}[\epsilon^2 xx^{'}] - \mathbb{E}[\epsilon xh^{'}]\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[\epsilon hx^{'}] \quad 0\Big) \begin{pmatrix} I \\ -\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[x\epsilon h^{'}] \end{pmatrix} \mathbb{E}[xx^{'}]^{-1}
\end{aligned}$$

Finally, the matrix multiplication yields

$$Avar(\hat{\theta}_{WLS}) = \mathbb{E}[xx^{'}]^{-1}\Big(\mathbb{E}[\epsilon^2 xx^{'}] - \mathbb{E}[\epsilon xh^{'}]\mathbb{E}[hh^{'}]^{-1}\mathbb{E}[\epsilon hx^{'}]\Big)\mathbb{E}[xx^{'}]^{-1}$$

which is the asymptotic variance that we had to derive.

Using capital letters for x just because it looks nicer, we get the asymptotic normality stated in the theorem.

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

It is easy to see that, without the weights, one retrieves the basic OLS asymptotic normality in the presence of heteroskedasticity.

4 Conclusion

After showing the usefulness of data combination and the weighting required given information about some known moment conditions in the population, we derived the estimator and provided a formal proof of its asymptotic properties needed for statistical tests.

References

[1] Angrist, J. D., and A. B. Krueger, 1992. "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples." Journal of the American Statistical Association 87: 328-36.

[2] Arellano, M., and C. Meghir, 1992. "Female labour supply and on-the-job-search: An empirical model estimated using complementary data sets." Review of Economic Studies 59: 537-57.

[3] Dominguez , A. N and I. N. Lobato, 2004."Consistent estimation of models defined by conditional moment restrictions" Vol. 72, No. 5, 1601-1615 , Econometrica.

[4] Hellerstein, J. K. and G. W. Imbens, 1999. "Imposing Moment Restrictions from Auxiliary Data by Weighting", Vol. 81 , No. 1, pp. 1-14 , Review of Economics and Statistics.

[5] Imbens, G. W., and T. Lancaster, 1994. "Combining micro and macro data in microeconometric models." Review of Economic Studies 61: 655-80.

[6] Newey, W. K. and D. McFadden, 1994. "Large sample estimation and hypothesis testing," Handbook of Econometrics, in: R. Engle & D. McFadden (ed.) volume 4, chapter 36, Elsevier.

[7] Ramalho,E. & J. Ramalho, 2006."Two-Step Empirical Likelihood Estimation under Stratified Sampling When Aggregate Information Is Available" (ed) Vol 74 No. 5 September, The Manchester School.

[8] Ridder, G. & R. Moffitt, 2007. "The Econometrics of Data Combination," Handbook of Econometrics, in: J.J. Heckman & E.E. Leamer (ed.), volume 6, chapter 75, Elsevier.