This is to certify that the

dissertation entitled

THREE ESSAYS ON ECONOMETRICS

presented by

CHIROK HAN

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in ECONOMICS

Major professor

Date JUNE 14, 2001

THREE ESSAYS ON ECONOMETRICS

By

Chirok Han

AN ABSTRACT OF A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Economics

2001

Professor Peter J. Schmidt

ABSTRACT

THREE ESSAYS ON ECONOMETRICS

By

Chirok Han

This dissertation contains three unrelated essays in econometric theory.

The first chapter considers Generalized Method of Moment-type estimators for which a criterion function is minimized that is not the "standard" quadratic distance measure, but instead is a general L_p distance measure. It is shown that the resulting estimators are root-n consistent, but not in general asymptotically normally distributed, and we derive the limit distribution of these estimators. In addition, we prove that it is not possible to obtain estimators that are more efficient than the "usual" L_2-GMM estimators by considering L_p-GMM estimators. We also consider the issue of the choice of the weight matrix for L_p-GMM estimators.

The second chapter is concerned with the asymptotic properties of the instrumental variable estimators with irrelevant instruments. The estimator is neither consistent nor asymptotically normal, but converges in distribution to a random variable which depends on the covariance of the regressors and the error term. The density of the asymptotic distribution is calculated and it is shown that the mean of the asymptotic distribution is equal to the probability limit of the OLS estimator.

The last chapter is an extension of Ahn, Lee and Schmidt (2001) to allow a *parametric* function for time-varying coefficients on the individual effects. It is shown that the main results of Ahn, Lee and Schmidt (2001) hold for our model, too. Least squares is consistent, given white noise errors, but less efficient than a GMM estimator.

ACKNOWLEDGMENTS

I would first like to thank Professor Peter J. Schmidt, the chairperson of my dissertation committee. He taught my first econometrics course and his enthusiasm and knowledge of this subject inspired me to pursue my present studies. I am grateful that he kindly acted as the Chairperson of my dissertation committee and guided me the for the course of my studies and writing of this dissertation. I would also like to thank Professor Robert M. de Jong, who taught me a great deal on methodological details and helpfully guided me in writing the first chapter of this thesis. I also would like to express my gratitude toProfessor Jeffrey M. Wooldridge for his timely comments and insightful critiques. Without my committee members' help, this thesis would not have been possible, although I am responsible for any remaining mistakes.

I am greatly indebted to my wife, Youngmi, for her love, support, and patience. My four-year-old son, Yoojeong, also helped by his constant love, which is a continuing great source of joy. My special thanks also go to my parents and my parents-in-law, who have supported and encouraged me in various ways with unconditional love and care.

I wish to thank my friends for their time, friendship and support. Some of them are Hoon Kim, Seok Hyeon Kim, Douglas Harris, Hong Peng Ong, and Neil Megan. I also would like to thank all the individual Korean students in the Department of Economics.

I would like to extend a special word of thanks to the staff in the Department of Economics, especially, Margaret Lynch, Linda Wirick and Pamela Dorton who skillfully helped me to handle many administrative items.

Finally, I acknowledge with gratitude Reverend Hyo Nam Hwang, Reverend Jung Kee Lee, and all the members of the Lansing Korean United Methodist Church for their concern, help and friendship. Thanks and glory to God.

TABLE OF CONTENTS

Chapter 1

The Properties of L_p–GMM Estimators

1.1 Introduction

Since Lars Peter Hansen's (1982) original formulation, Generalized Method of Moment (GMM) estimation has become an extremely important and popular estimation technique in economics. This is due to the fact that economic theory usually implies moment conditions that are exploited in the GMM technique, while typically economic theory is uninformative about the exact stochastic structure of economic processes. GMM estimation provides an estimator when a certain set of moment conditions $Eg(y, \theta_0) = 0$ is *a priori* known to hold. When the number of moment conditions exceeds the number of parameters, we cannot hope to obtain an estimator by setting the empirical equivalent $\bar{g}(\theta)$ of our moment condition equal to zero, but instead we will need to make $\bar{g}(\theta)$ as close to zero as possible in some sense. The usual GMM formulation minimizes a quadratic measurement of distance. Hansen (1982) established the large sample properties of these GMM estimators under mild regularity conditions.

The above exposition raises the natural questions of what happens if distance measures other than a quadratic one is used and whether or not those other distance measures can give better estimators. The answer to the latter question is no, as Chamberlain (1987) has shown that the optimal GMM (in the usual sense) estimators attain the efficiency bound. Apart from this general remark on the efficiency of optimal GMM estimators, there have been attempts such as Manski (1983) and Newey (1988) to directly treat the use of non-quadratic measures of distance between population and empirical moments. In those articles results are stated that imply that under mild assumptions, estimators that minimize a general discrepancy function are consistent and asymptotically normally distributed. Based on these results, Newey (1988) concludes that (under regularity conditions) estimators using two different measures of distance are asymptotically equivalent if the corresponding Hessian matrices are asymptotically equal. This implies that it is impossible to obtain better estimators by modifying the quadratic criterion function, given the assumptions of that paper. This conclusion gives a direct argument for the use of quadratic distance measure beside Chamberlain's general argument.

However, when considering L_p-GMM as defined below, it turns out that only the L_2 norm satisfies the assumptions of Manski (1983) and Newey (1988), and other values of p in $[1, \infty)$ do not. The problems are the following. When $p = 1$, the L_p norm is not differentiable at 0; when $p \in (1, 2)$, it is continuously differentiable but is not twice differentiable at 0; when $p \in (2, \infty)$, it is continuously twice differentiable, but the Hessian matrix evaluated at the true parameter becomes zero (and therefore singular). Therefore, the papers by Newey and Manski have no implications for L_p-GMM for values of p other than 2. When considering L_p-GMM, it turns out that the "standard" asymptotic framework will fail. Also, the least absolute deviations type asymptotic framework also does not directly apply. Linton (1999) has recently pointed out in an example in *Econometric Theory* that the estimator minimizing the L_1 distance of the sample moments from zero can have a non-normal limit distribution. In this chapter, we will establish the limit distribution of general L_p-GMM estimators, and we show that L_p-GMM estimators are root-n consistent, but in general need not have an asymptotically normal distribution. In addition, we prove a theorem that shows that L_p-GMM estimators cannot be more efficient than L_2-GMM estimators, thereby strengthening Newey's conclusion to L_p-GMM estimators. Finally, we discuss the problem of finding the optimal weight matrix for L_p-GMM estimators.

Section 1.2 defines our estimator and gives the main theorem for consistency and asymptotic distribution, whose proof is given in the Mathematical Appendix (Section 1.A). Section 1.3 discusses the efficiency of L_2-GMM among all L_p-GMM estimators. Section 1.4 describes the problem of the selection of the weight matrix. In addition, this section gives some interesting results for the case when $p = 1$ and $p = 3$, including Linton's (1999) example. The conclusions section (Section 1.5) is followed by a Mathematical Appendix in which all the proofs are gathered.

1.2 Main theorem

In this section, the main result of this chapter on which the remainder of our discussion of this chapter is based will be stated. Let $y_1, y_2, \ldots$ be a sequence of i.i.d. random vectors in $\mathbb{R}^m$. Let $g(y_i, \theta)$ be a set of q moment conditions with parameter $\theta \in \Theta \subset \mathbb{R}^k$, that is, let $g(y_i, \theta)$ be a random vector in $\mathbb{R}^q$ that satisfies

$$Eg(y_i, \theta_0) = 0. \tag{1.1}$$

Let $\bar{g}(\theta) = n^{-1} \sum_{i=1}^{n} g(y_i, \theta)$. The L_p norm $||\cdot||_p$ is defined as

$$||x||_p = (\sum_{j=1}^{q} |x_j|^p)^{1/p} \tag{1.2}$$

for $p \in [1, \infty)$. The L_p-GMM estimator $\hat{\theta}_n$ is assumed to satisfy

$$||\bar{g}(\hat{\theta}_n)||_p = \inf_{\theta \in \Theta} ||\bar{g}(\theta)||_p. \tag{1.3}$$

Let $|x| = \max_{i,j} |x_{ij}|$ if x is a $k_1 \times k_2$ matrix. Let $\Omega = Eg(y_i, \theta_0)g(y_i, \theta_0)'$, and $D = E(\partial/\partial\theta')g(y_i, \theta_0)$. The regularity assumption below will be needed to establish our results:

Assumption 1.1.

(i) Θ *is a compact and convex subset in* $\mathbb{R}^k$*;*

(ii) θ_0 *is an interior point of* Θ*;*

(iii) $Eg(y_i, \theta) = 0$ *iff* $\theta = \theta_0$*, i.e.,* θ_0 *uniquely satisfies the moment conditions;*

(iv) $g(y, \theta)$ *is continuous in* θ *for each* $y \in \mathbb{R}^m$*, and is measurable for each* $\theta \in \Theta$*;*

(v) $E \sup_{\theta \in \Theta} |g(y_i, \theta)| < \infty$*;*

(vi) $\Omega = Eg(y_i, \theta_0)g(y_i, \theta_0)'$ *is finite;*

(vii) $g(y, \theta)$ *has first derivative with respect to* θ *which is continuous in* $\theta \in \Theta$ *for each* $y \in \mathbb{R}^m$ *and measurable for each* $\theta \in \Theta$*,* $E \sup_{\theta \in \Theta} |(\partial/\partial\theta')g(y_i, \theta)| < \infty$*, and* D *is of full column rank.*

(viii) $||x + D\xi||_p$ *achieves its minimum at a unique point of* ξ *in* $\mathbb{R}^k$ *for each* $x \in \mathbb{R}^q$*.*

Note that part *(viii)* of Assumption 1.1 is nonstandard and far from innocent in the L_1 case. Consider for example sequences of random variables y_{i1} and y_{i2} that are independent of each other and are $N(0, 1)$ distributed, and consider the L_1-GMM estimator that minimizes

$$|\bar{y}_1 - \theta| + |\bar{y}_2 - \theta|. \tag{1.4}$$

Part *(viii)* of Assumption 1.1 will not hold in this case, because any value in the interval $[\min(\bar{y}_1, \bar{y}_2), \max(\bar{y}_1, \bar{y}_2)]$ will minimize the criterion function. Therefore, our result does not establish the limit distribution of the L_p-GMM estimator for this case. However, if we consider the weighted criterion function

$$|\bar{y}_1 - \theta| + c|\bar{y}_2 - \theta| \tag{1.5}$$

for any $c \in [0, \infty)$ except for $c = 1$, part (*viii*) of Assumption 1.1 will be satisfied.

The following theorem now summarizes the asymptotic properties of L_p-GMM estimators. Note that we do not yet explicitly consider weight matrices at this point, but such a treatment can be easily done with the result below at hand.

Theorem 1.2. *Let Y be a random vector in $\mathbb{R}^q$ distributed $N(0, \Omega)$. Then under Assumption 1.1, $\hat{\theta}_n \to \theta_0$ a.s., and*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{\xi \in \mathbb{R}^k} \|Y + D\xi\|_p. \tag{1.6}$$

The proof of this theorem, like all the proofs of this chapter, can be found in the Appendix (Section 1.A). As a special case of the above theorem, the usual L_2-GMM estimator can be considered. $\|Y + D\xi\|_2^2 = (Y + D\xi)'(Y + D\xi)$ is minimized by $\tilde{\xi} = -(D'D)^{-1}D'Y$, so applying Theorem 1.2, we get

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} -(D'D)^{-1}D'Y \sim N[0, (D'D)^{-1}D'\Omega D(D'D)^{-1}] \tag{1.7}$$

which coincides with usual analysis.

In examples below, we will show that for general values of p, normality need not result for the L_p-GMM estimator. We will be able to establish though that the limit distribution is symmetric around 0 and possesses finite second moments.

1.3 Efficiency of L_2-GMM

In this section and in the remainder of this chapter, we consider L_p-GMM estimation with a weight matrix W, i.e., L_p-GMM estimators that minimize the distance from zero of "weighted" average of moment conditions $\|W\bar{g}(\theta)\|_p$, where W is a $q \times q$ nonrandom and nonsingular matrix. It is straightforward to extend our analysis to the case of estimated matrices W, and we will not pursue that issue here. Clearly, whenever $Eg(y_i, \theta_0) = 0$ we will have $EWg(y_i, \theta_0) = 0$, and therefore our previous analysis applies. Below, we will keep using the notations Y, D, and Ω defined previously.

Let $\tilde{\xi}$ minimize $\|W(Y + D\xi)\|_p$. Applying Theorem 1.2, we see that $\hat{\theta}_n$, which minimizes $\|W\bar{g}(\theta)\|_p$, is strongly consistent (since $Wg(y_i, \theta_0)$ is also a set of legitimate moment conditions) and $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \tilde{\xi}$.

To facilitate the efficiency discussion, we need to show asymptotic unbiasedness of L_p-GMM estimators. This is established by noting that the limiting distribution of $n^{1/2}(\hat{\theta}_n - \theta_0)$ is symmetric and has a finite second moment. The following theorem states the unbiasedness result:

Theorem 1.3. *Under Assumption 1.1, L_p-GMM estimators are asymptotically unbiased.*

Because of the asymptotic unbiasedness of our estimators, we can compare weighted L_p-GMM estimators by their asymptotic variances. This property is crucial to prove the following theorem. This result states that optimal L_2-GMM estimators are asymptotically efficient among the class of weighted L_p-GMM estimators.

Theorem 1.4. *Under Assumption 1.1, an optimal L_2-GMM estimator is asymptotically efficient among the class of weighted L_p-GMM estimators, i.e., the asymptotic variance of an optimal L_2-GMM estimator is less than or equal to that of any weighted L_p-GMM estimator.*

The above theorem provides us with the knowledge that the central message from the result by Newey (1988)—that there is no potential for efficiency improvement by considering discrepancy functions other than quadratic—can be extended towards L_p-GMM estimators. Basically, Theorem 1.4 is obtained by noting that the expression for the limit distribution can be viewed as a finite sample estimation problem in its own right, for which the Cramér-Rao underbound applies.

1.4 Further remarks on weight matrices

In this section, we will discuss various issues involving the choice of the weight matrix W and discuss several examples. We will not be able to prove optimality of a particular nonsingular weight matrix for general L_p-GMM, but instead we will sketch some of the issues below.

It is well-known that the optimal weight matrix W for $p = 2$ satisfies $W\Omega W' = I$ (or $W'W = \Omega^{-1}$). This result can be easily obtained using our first theorem too, for

$$||W(Y + D\xi)||_2^2 = (Y + D\xi)'W'W(Y + D\xi) \tag{1.8}$$

is minimized by

$$\tilde{\xi} = -(D'W'WD)^{-1}D'W'WY \tag{1.9}$$

and its variance is minimized when $W'W = \Omega^{-1}$. Therefore the optimal L_2-GMM estimator has the asymptotic distribution

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, (D'\Omega^{-1}D)^{-1}]. \tag{1.10}$$

Can this efficiency be attained for p other than 2? In general, the answer is yes. It can be achieved for general p by weighting cleverly. Consider

$$W^* = (\Omega^{-1}D \;\vdots\; W_2)' \tag{1.11}$$

where W_2 is of size $q \times (q - k)$, chosen to be orthogonal to D, i.e., $W_2'D = 0$ and chosen such that W^* is nonsingular. This weight matrix always exists when $q > k$.[1] Then

$$||W^*(Y + D\xi)||_p^p = \left\| \begin{pmatrix} D'\Omega^{-1}(Y + D\xi) \\ W_2'(Y + D\xi) \end{pmatrix} \right\|_p^p$$

[1] This weight matrix needs $W_2'D = 0$ and $|W^*| \neq 0$, so there are $(q - k)k + 1$ restrictions. But W^* has $q(q - k)$ free parameters. The number of parameters is greater than or equal to the number of restrictions when $q > k$.

$$= ||D'\Omega^{-1}Y + D'\Omega^{-1}D\xi||_p^p + ||W_2'Y||_p^p \tag{1.12}$$

is minimized by $\tilde{\xi} = -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}Y \sim N[0, (D'\Omega^{-1}D)^{-1}]$ for any $p \geq 1$. So the W^*-weighted L_p-GMM estimator $\hat{\theta}_n$ with W^* chosen as in Equation (1.11) has the asymptotic distribution (1.10), and therefore the weight matrix W^* is optimal for any p.

For $p = 2$, there are two different types of optimal weights. One is given by (1.11) (say, $D'\Omega^{-1}$ type) and the other is characterized by $W\Omega W' = I$ (note that a scalar multiplication of an optimal weight matrix is again optimal). In general, each of these neither implies nor is implied by the other, but they give one and the same asymptotic distribution. Furthermore, the optimal weight of the second type is not unique, since any orthogonal transformation of an optimal weight is again optimal. (When $W\Omega W' = I$, $V = HW$ also satisfies $V\Omega V' = I$ provided $H'H = HH' = I$.) This is, of course, because the W-weighted L_2 distance $||Wx||_2 = (x'W'Wx)^{1/2}$ depends only on the product $W'W$ but not W itself.

But when $p \neq 2$, two different orthogonal transformations W and V of $\Omega^{-1/2}$ are not expected to give equivalent asymptotic distribution, even though both $W\Omega W' = I$ and $V\Omega V' = I$ hold. Here are a few examples.

(i) Our first example is for $p = 1$, $q = 2$, and $k = 1$. Suppose that it is known that (y_{1i}, y_{2i}) is i.i.d. across i with mean $(\theta_0, 2\theta_0)$ and covariance I. Then the moment condition is $E(y_{1i} - \theta_0, y_{2i} - 2\theta_0)' = 0$, and therefore $D = -(1 \quad 2)'$, and $\Omega = I$. Consider two weight matrices: $W = I$ and

$$V = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}. \tag{1.13}$$

It can be seen that V is an optimal weight matrix here for $p \in [1, \infty)$, since the same limit distribution as for optimal L_2-GMM is obtained using V. In the case $W = I$, the W-weighted L_1-GMM estimator can be obtained by minimizing the criterion function

$$|\bar{y}_1 - \theta| + |\bar{y}_2 - 2\theta|, \tag{1.14}$$

and the minimizer equals $(1/2)\bar{y}_2$. This implies that the rescaled and centered W-weighted L_1-GMM estimator is asymptotically distributed as $N(0, 1/4)$, while the rescaled and centered V-weighted L_p-GMM estimator is asymptotically distributed $N(0, 1/5)$.

(*ii*) Here is a more interesting example for $p = 1$, $q = 3$, and $k = 1$. Suppose that y_{1i}, y_{2i} and y_{3i} are mutually independent and i.i.d. across i, have a mean θ_0, and a variance of 1. Then, the moment condition is

$$E(y_{1i} - \theta_0, y_{2i} - \theta_0, y_{3i} - \theta_0)' = 0,$$

implying that $D = -(1, 1, 1)'$ and $\Omega = I$. Consider the two weight matrices $W = I$ and

$$V = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ -\sqrt{2/3} & 1/\sqrt{6} & 1/\sqrt{6} \end{pmatrix}. \tag{1.15}$$

Again, V can be shown to be an optimal weight matrix in this example. In the case of $W = I$, this situation could result when we are minimizing the criterion function

$$|\bar{y}_1 - \theta| + |\bar{y}_2 - \theta| + |\bar{y}_3 - \theta|. \tag{1.16}$$

Note that both W and V are chosen to be orthogonal. The W-weighted L_1-GMM estimator (after centering and scaling) converges in distribution to $\text{argmin}_\xi \, ||N(0, I_3) + D\xi||_1$. The minimizing argument $\tilde{\xi}$, which is the (unique) median of three independent standard normal random variables, has distribution

$$P(\tilde{\xi} \le x) = 6 \int_{-\infty}^{x} \Phi(t)[1 - \Phi(t)]\phi(t)dt = \Phi(x)^2[3 - 2\Phi(x)] \tag{1.17}$$

(see Linton (1999)). This distribution is not normal, and simulations for three standard normals illustrate that the density of the median has sharper center and thicker tail than a (properly rescaled) normal ($N(0, 2/3)$). The result of using V as the weight matrix is different. We have $VD = (-\sqrt{3}\ 0\ 0)'$ and the V-weighted L_1-GMM estimator (after centering and scaling) converges in distribution to $\text{argmin}_\xi\{\|Z + VD\xi\|_1 = |Z_1 - \sqrt{3}\xi| + |Z_2| + |Z_3|\}$ where $Z = (Z_1, Z_2, Z_3)' \sim N(0, I)$. Note that basically, this optimal weight matrix will eliminate two out of three absolute value elements of the criterion function of Equation (1.16). The solution $\tilde{\xi}$ is distributed $N(0, 1/3)$. This example shows that two weight matrices W and V that satisfy $W\Omega W' = I$ and $V\Omega V' = I$ can give asymptotics different not only in variance but in the type of limit distribution, since the one distribution is non-normal, while the other is normal.

(*iii*) The only tractable example for $p > 2$ that we could find is the following. Consider the above case of $q = 3$, $k = 1$, $\Omega = I$, and $D = -(1\ 1\ 1)'$. The weight matrix V of (1.15) is again optimal and the V-weighted L_3-GMM estimator is asymptotically normal. In the case when the weight is $W = I$ so the objective function to be minimized is

$$|\bar{y}_1 - \theta|^3 + |\bar{y}_2 - \theta|^3 + |\bar{y}_3 - \theta|^3, \tag{1.18}$$

the W-weighted L_3-GMM estimator (after centering and rescaling) converges in distribution to $\tilde{\xi} = \text{argmin}_\xi \|Y + D\xi\|_3$ where $Y = (Y_1, Y_2, Y_3)' \sim N(0, I_3)$. Let $(Y_{(1)}, Y_{(2)}, Y_{(3)})$ be the order statistic of (Y_1, Y_2, Y_3), and $(\delta_{(1)}, \delta_{(2)}, \delta_{(3)})$ be the order statistic of $(|Y_1 - Y_2|, |Y_2 - Y_3|, |Y_3 - Y_1|)$. Then it turns out that

$$\begin{aligned} \tilde{\xi} &= \bar{Y} + \text{sgn}(Y_{(1)} + Y_{(3)} - 2Y_{(2)})[(2/3)(\delta_{(3)} + \delta_{(2)}) - (2\delta_{(3)}\delta_{(2)})^{1/2}] \\ &= Y_{(2)} - \text{sgn}(Y_{(1)} + Y_{(3)} - 2Y_{(2)})[\delta_{(3)} - (2\delta_{(3)}\delta_{(2)})^{1/2}] \end{aligned} \tag{1.19}$$

where $\text{sgn}(a) = 1\{a > 0\} - 1\{a < 0\}$ and $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$. In simulations, this distribution cannot be distinguished from a normal.

The natural question now arises whether we can get optimality by a nonsingular weight matrix W satisfying $W\Omega W' = I$. In short, the answer is yes provided $D'\Omega^{-1}D$ equals a scalar or a scalar matrix (a scalar times the identity matrix). The question here is whether we can construct $(\Omega^{-1}D \;\vdots\; W_2)'$ (where $W_2'D = 0$) by an orthogonal transformation of $\Omega^{-1/2}$, that is, whether there exists an orthogonal matrix H of size $q \times q$ such that $H\Omega^{-1/2} = \lambda(\Omega^{-1}D \quad W_2)'$ and $W_2'D = 0$. If such a weight matrix H is exist, it will have the form $H = \lambda(\Omega^{-1/2}D \quad \Omega^{1/2}W_2)'$ for which (a) $W_2'D = 0$, (b) H is nonsingular, i.e., $|H| \neq 0$, and (c) $HH' = I$, i.e.,

$$HH' = \lambda^2 \begin{pmatrix} D'\Omega^{-1}D & D'W_2 \\ W_2'D & W_2'\Omega W_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \tag{1.20}$$

(a) imposes $(q-1)k$ restrictions, and (b) imposes 1 extra restriction. When $k = 1$, (c) is equivalent to $W_2'\Omega W_2 = (D'\Omega^{-1}D)I_{q-1}$ due to (a), which imposes $(q-1)(q-2)/2$ more restrictions. Therefore, when $k = 1$, we have $q(q-1)+1$ free parameters (for W_2 and λ) and $(q-1)+1+(q-1)(q-2)/2 = q(q-1)/2+1$ restrictions. So the number of parameters to be set is greater than or equal to the number of restrictions, whence we conclude that we can find W_2 satisfying (a), (b), and (c). When $k > 1$, (c) can not be satisfied unless $D'\Omega^{-1}D$ is a scalar matrix, but if $D'\Omega^{-1}D$ is so, W_2 and λ satisfying (a), (b), and (c) can be found. The V matrices in the examples above are constructed in this way and are optimal for those problems.

Note that this rule does not depend on the specific value of p, and that the reason the weight W satisfying $W\Omega W' = I$ is optimal for $p = 2$ does not lie in that the optimal weight of type $D'\Omega^{-1}$ can be obtained by an orthogonal transformation of $\Omega^{-1/2}$, but in the specific properties of L_2 distance.

1.5 Conclusion

In this chapter we derived an abstract expression for the limit distribution of estimators which minimizes the L_p distance between population moments and sample moments, as follows:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{\xi \in \mathbb{R}^k} \|Y + D\xi\|_p \tag{1.21}$$

where $Y \sim N[0, Eg(y_i, \theta_0)g(y_i, \theta_0)']$ and $D = E(\partial/\partial\theta')g(y_i, \theta_0)$. This asymptotic representation allows a generalization of the well-known GMM framework of Hansen (1982) towards the L_p distance. As mentioned in the introduction, Manski (1983) and Newey (1988) generalized GMM to allow arbitrary distance (or, more generally, discrepancy) function. But unfortunately they need the second order differentiability of the distance functions and the nonsingularity of a Hessian matrix evaluated at true parameter. Only the L_2 distance satisfies these conditions among all L_p distances.

However, our analysis can not give an explicit form for the asymptotic distribution, but only allows the above abstract representation in terms of the argmin functional. Nonetheless, our method directly supports the result of Chamberlain (1987) that the optimal L_2-GMM estimator is efficient among the class of L_p-GMM estimators. Interestingly, our analysis reduced the analysis of efficiency issues of L_p-GMM estimators to the analysis of the small sample properties of estimators minimizing the L_p distance between Y and $-D\xi$, i.e., $\operatorname{argmin}_{\xi \in \mathbb{R}^k} \|Y + D\xi\|_p$.

As a final remark, note that it is interesting to consider potential robustness properties of the L_p-GMM procedure. The asymptotic results that were presented in this chapter all rely on central limit theorems and existence of second moments, so in this sense, we probably should not expect the L_p-GMM method to have robustness properties of any type. However, since the objective function in the case of L_1-GMM effectively puts less weight on "outlier" moments, one might expect that L_1-GMM may be less vulnerable to the inclusion of an incorrect moment condition than "standard" L_2-GMM estimators. No attempt will be made however in this chapter to formalize this intuition.

1.A Mathematical Appendix

In order to establish the theorems, we will need several results that will be stated as lemmas. Lemma 1.5 is used to prove Theorem 1.3 (the asymptotic unbiasedness of L_p-GMM estimators).

Lemma 1.5. *Let a random vector Y in $\mathbb{R}^q$ with finite q have a normal distribution. Let D be a real nonrandom matrix of size $q \times k$ $(q \geq k)$ with full column rank. Then for any $p \in [1, \infty)$,*

$$\tilde{\xi} = \operatorname*{argmin}_{\xi \in \mathbb{R}^k} ||Y + D\xi||_p \tag{1.22}$$

will have a well-defined finite covariance matrix.

Proof. First note that, because $D'D$ has full column rank under Assumption 1.1,

$$\begin{aligned} \tilde{\xi}'\tilde{\xi} &\leq \tilde{\xi}'D'D\tilde{\xi}/\lambda_{min}(D'D) \\ &\leq (||Y + D\tilde{\xi}||_2 + ||Y||_2)^2/\lambda_{min}(D'D) \\ &\leq c_{p/2}^{2/p}(||Y + D\tilde{\xi}||_p + ||Y||_p)^2/\lambda_{min}(D'D) \\ &\leq c_{p/2}^{2/p}(2||Y||_p)^2/\lambda_{min}(D'D) \end{aligned} \tag{1.23}$$

where the first inequality follows from full column rank of D, the second inequality is the triangle inequality, the third is the inequality

$$(\sum_{i=1}^{q} b_i^2)^{1/2} \leq c_{p/2}^{1/p}(\sum_{i=1}^{q} |b_i|^p)^{1/p} \tag{1.24}$$

which is a consequence of Loève's c_r inequality (see Davidson (1994, p. 140), Equation (9.63)), and the fourth follows by the fact that $||Y + D\xi||_p$ is minimized at $\xi = \tilde{\xi}$. The result then follows because all moments of the normal distribution are finite. □

The first step towards the proof of Theorem 1.2 is the strong consistency proof for L_p-GMM estimators, which can be accomplished by invoking several theorems from Bierens

(1994).

Lemma 1.6. *Under Assumption 1.1, the L_p-GMM estimator $\hat{\theta}_n$ is strongly consistent.*

Proof. First, conditions (*i*) and (*iv*) of Assumption 1.1 ensure the existence and measurability of $\hat{\theta}_n$ by Theorem 1.6.1 of Bierens (1994). The above conditions together with condition (*v*) of Assumption 1.1 imply that $\bar{g}(\theta)$ converges to $Eg(y_i, \theta)$ almost surely uniformly on Θ by Theorem 2.7.5 of Bierens (1994). Hence, $\|\bar{g}(\theta)\|_p \to \|Eg(y_i, \theta)\|_p$ a.s. uniformly on Θ since $\|\cdot\|_p$ is continuous. Finally, this uniform convergence result and the uniqueness of θ_0 by condition (*iii*) of Assumption 1.1 give the stated result by Theorem 4.2.1 of Bierens (1994). □

To prove the main assertion of Theorem 1.2, we will use Theorem 2.7 of Kim and Pollard (1990). We restate Kim and Pollard's theorem as our next lemma.

Lemma 1.7. *Let $Q, Q_1, Q_2, \cdots$ be real-valued random processes on $\mathbb{R}^k$ with continuous paths, and $\hat{\xi}_n$ be random vector in $\mathbb{R}^k$, such that*

(*i*) $Q(\xi) \to \infty$ *as* $|\xi| \to \infty$;

(*ii*) $Q(\cdot)$ *achieves its minimum at a unique point in* $\mathbb{R}^k$;

(*iii*) Q_n *converges weakly to* Q *on any set* $\Xi = [-M, M]^k$;

(*iv*) $\hat{\xi}_n = O_P(1)$;

(*v*) $\hat{\xi}_n$ *minimizes* $Q_n(\xi)$.

Then $\hat{\xi}_n \xrightarrow{d} \operatorname{argmin}_{\xi \in \mathbb{R}^k} Q(\xi)$.

Proof. See Theorem 2.7 of Kim and Pollard (1990). □

To apply Kim and Pollard's theorem and show that its conditions are satisfied in our situation, we need the following three lemmas. For these lemmas, we need the following definitions. Define

$$\hat{\xi}_n = n^{1/2}(\hat{\theta}_n - \theta_0), \quad \text{for } n = 1, 2, \ldots, \tag{1.25}$$

where $\hat{\theta}_n$ is L_p-GMM estimator. Define $\mathbb{R}^q$-valued random functions $h, h_1, h_2, \ldots$ by

$$h_n(\xi) := \begin{cases} n^{1/2}\bar{g}(\theta_0 + \xi n^{-1/2}), & \text{if } \theta_0 + \xi n^{-1/2} \in \Theta \\ n^{1/2}\bar{g}(\theta^0), & \text{otherwise} \end{cases} \tag{1.26}$$

where $\theta^0 = \text{argmax}_{\theta \in \Theta} \|\bar{g}(\theta)\|_p$, for $n = 1, 2, \ldots$, and $h(\xi) = Y + D\xi$ where Y is a $\mathbb{R}^q$-valued random vector distributed $N(0, \Omega)$. Let

$$Q_n(\xi) = \|h_n(\xi)\|_p \quad \text{and} \quad Q(\xi) = \|h(\xi)\|_p. \tag{1.27}$$

The lemmas that we need for the proof of our central result are then the following:

Lemma 1.8. *Suppose the conditions of Assumption 1.1 are satisfied. Then* $n^{1/2}(\hat{\theta}_n - \theta_0) = O_P(1)$.

Proof. The Taylor expansion of $\bar{g}(\theta)$ around $\theta = \theta_0$ implies that

$$\bar{g}(\hat{\theta}_n) = \bar{g}(\theta_0) + (\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0), \tag{1.28}$$

where $\tilde{\theta}_n$ is a mean value in between $\hat{\theta}_n$ and θ_0. From the above Taylor series, from the triangular inequality for the L_p norm, and from the fact that $\hat{\theta}_n$ minimizes $\|\bar{g}(\theta)\|_p$, we have

$$\begin{aligned} \|n^{1/2}(\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)\|_p &\leq \|n^{1/2}\bar{g}(\hat{\theta}_n)\|_p + \|n^{1/2}\bar{g}(\theta_0)\|_p \\ &\leq 2\|n^{1/2}\bar{g}(\theta_0)\|_p. \end{aligned} \tag{1.29}$$

But condition (*vi*) of Theorem 1.2 implies that $n^{1/2}\bar{g}(\theta_0)$ converges in distribution by central limit theorem, and therefore is $O_P(1)$. Therefore,

$$\|n^{1/2}(\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)\|_p = O_P(1). \tag{1.30}$$

Condition (*vii*) of Theorem 1.2 implies that $(\partial/\partial\theta')\bar{g}(\theta)$ follows a strong uniform law of large numbers, which combined with the consistency of $\hat{\theta}$ implies that $(\partial/\partial\theta')\bar{g}(\tilde{\theta}_n) \xrightarrow{\text{a.s.}} E(\partial/\partial\theta')g(y_i, \theta_0) = D$. Now let $\tilde{D} = (\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)$. Then it follows that $\tilde{D}'\tilde{D} \xrightarrow{\text{a.s.}} D'D$. Since $D'D$ is strictly positive definite, $\tilde{D}'\tilde{D}$ becomes strictly positive definite for n large enough. Therefore, for n large enough,

$$n(\hat{\theta}_n - \theta_0)'(\hat{\theta}_n - \theta_0) \leq n(\hat{\theta}_n - \theta_0)'\tilde{D}'\tilde{D}(\hat{\theta}_n - \theta_0)/\tilde{\lambda}_{min}, \qquad (1.31)$$

where $\tilde{\lambda}_{min}$ is the smallest eigenvalue of $\tilde{D}'\tilde{D}$. Because $\tilde{\lambda}_{min} \xrightarrow{\text{a.s.}} \lambda_{min} > 0$ where λ_{min} is the smallest eigenvalue of $D'D$, we get $\tilde{\lambda}_{min} \geq 0.5\lambda_{min}$ eventually (for n large enough) almost surely. Therefore, as n increases, the right hand side of (1.31) eventually becomes less than $4n(\hat{\theta}_n - \theta_0)'\tilde{D}'\tilde{D}(\hat{\theta}_n - \theta_0)/\lambda_{min}$. By Equation (1.30) and because of the equivalence of L_p and L_q norms for $p, q \in [1, \infty)$, this expression is $O_P(1)$, which completes the proof. □

Lemma 1.9. *Consider random functions $Q, Q_1, Q_2, \ldots$ defined by (1.27). Under Assumption 1.1, the finite-dimensional distributions of Q_n converge to the finite-dimensional distributions of Q.*

Proof. With fixed ξ, condition (*ii*) of Assumption 1.1 (θ_0 is an interior point of Θ) ensures that $\theta_0 + \xi n^{-1/2}$ belongs to Θ for n large enough. When this happens, by the Taylor expansion,

$$h_n(\xi) = n^{1/2}\bar{g}(\theta_0 + \xi n^{-1/2}) = n^{1/2}\bar{g}(\theta_0) + (\partial/\partial\theta')\bar{g}(\theta_0 + \tilde{\xi}n^{-1/2})\xi, \qquad (1.32)$$

with $\tilde{\xi}$ lying in between ξ and 0. Condition (*vi*) of Assumption 1.1 (finiteness of the second moment of $g(y_i, \theta_0)$) implies that $n^{1/2}\bar{g}(\theta_0) \xrightarrow{d} Y$, and condition (*vii*) of Theorem 1.2 implies that $(\partial/\partial\theta)\bar{g}(\theta_0 + \tilde{\xi}n^{-1/2})\xi \to D\xi$ a.s. similar to the proof of Lemma 1.8.

To conclude the proof and show the convergence of the finite-dimensional distributions of h_n to h, we can use the Cramér-Wold device (see for example Billinsley (1968), Theorem 7.7), which states that

$$(h_n(\xi_1)' \cdots h_n(\xi_r)')' \xrightarrow{d} (h(\xi_1)' \cdots h(\xi_r)')' \qquad (1.33)$$

if and only if

$$\sum_{j=1}^{r} \lambda_j' h_n(\xi_j) \xrightarrow{d} \sum_{j=1}^{r} \lambda_j' h(\xi_j) \qquad (1.34)$$

for each $\lambda_1 \in \mathbb{R}^q, \ldots, \lambda_r \in \mathbb{R}^q$. And (1.34) is to be easily shown using the result of the first part of this proof.

Finally, note that since $\|\cdot\|_p$ is continuous, the finite-dimensional distribution of $Q_n = \|h_n\|_p$ converge to those of $Q = \|h\|_p$ by the continuous mapping theorem. □

Lemma 1.10. *Under Assumption 1.1, $Q_n(.)$ defined by Equation (1.27) is stochastically equicontinuous on any set $\Xi = [-M, M]^k$.*

Proof. Using the triangular inequality for $\|\cdot\|_p$, we have

$$\begin{aligned} &|Q_n(\xi_1) - Q_n(\xi_2)| \\ &= |\,\|h_n(\xi_1)\|_p - \|h_n(\xi_2)\|_p| \\ &\le \|h_n(\xi_1) - h_n(\xi_2)\|_p \\ &= \|(\partial/\partial\theta')\bar{g}(\theta_0 + \tilde{\xi}_1 n^{-1/2})\xi_1 - (\partial/\partial\theta')\bar{g}(\theta_0 + \tilde{\xi}_2 n^{-1/2})\xi_2\|_p \qquad (1.35) \end{aligned}$$

where $\tilde{\xi}_i$ lies in between ξ_i and 0 for $i = 1, 2$. By the strong uniform law of large numbers for $(\partial/\partial\theta')\bar{g}(\theta)$ and the convergence to zero of $\tilde{\xi}_i n^{-1/2}$ uniformly over all ξ_1 and ξ_2,

$$\sup_{\xi_1 \in \Xi, |\xi_1 - \xi_2| < \delta} |Q_n(\xi_1) - Q_n(\xi_2)| \xrightarrow{\text{a.s.}} \sup_{\xi_1 \in \Xi, |\xi_1 - \xi_2| < \delta} \|D(\xi_1 - \xi_2)\|_p \qquad (1.36)$$

under the conditions of Assumption 1.1. Therefore, by nonsingularity of $D'D$, it follows that for all $\eta > 0$

$$\lim_{\delta\to 0}\limsup_{n\to\infty} P\Big(\sup_{\xi_1\in\Xi, |\xi_1-\xi_2|<\delta} |Q_n(\xi_1) - Q_n(\xi_2)| > \eta\Big) = 0, \tag{1.37}$$

which is the stochastic equicontinuity condition. □

Proof of Theorem 1.2. The strong consistency result of this theorem is proven in Lemma 1.6. For the proof of the main assertion of this theorem, we will show that for the Q_n, Q and $\hat{\xi}_n$ as defined above, all the conditions of Lemma 1.7 are implied by the conditions of Theorem 1.2. First, note that Q_n, Q, and $\hat{\xi}_n$, defined by (1.25) and (1.27), satisfy conditions (*i*)–(*v*) of Lemma 1.7 under the conditions of Theorem 1.2. Condition (*v*) of Lemma 1.7 is guaranteed by the definitions of $\hat{\theta}_n$, $\hat{\xi}_n$, and Q_n. It is also not difficult to notice that condition (*i*) of Lemma 1.7 is trivially satisfied since D is of full column rank. And condition (*ii*) of Lemma 1.7 is just supposed by condition (*viii*) of Theorem 1.2. The weak convergence condition is verified by showing stochastic equicontinuity and finite-dimensional convergence, which together with compactness of the parameter space is well-known to imply weak convergence. Lemmas 1.8, 1.9, and 1.10 therefore ensure that the conditions of Lemma 1.7 are all implied by the conditions of Theorem 1.2, and therefore convergence in distribution of our estimator is proven by invoking Lemma 1.7. □

Proof of Theorem 1.3. By Lemma 1.5, $\tilde{\xi}$ has a finite mean. And by Theorem 1.2, $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \tilde{\xi} = \text{argmin}_\xi \|Y + D\xi\|_p$ where $Y \sim N(0, \Omega)$ and $D = E(\partial/\partial\theta')g(y_i, \theta_0)$. From the symmetry of Y, it follows that $\|Y + D\xi\|_p$ is distributed identically to $\|Y + D(-\xi)\|_p$, which implies identical distributions of $\tilde{\xi}$ and $-\tilde{\xi}$. Therefore, the mean of $\tilde{\xi}$ is 0. □

Proof of Theorem 1.4. Let $\hat{\theta}_n$ be the W-weighted L_p-GMM estimator. By Theorem 1.2,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \underset{\xi}{\text{argmin}}\, \|W(Y + D\xi)\|_p. \tag{1.38}$$

So the problem here is to show that $\tilde{\xi}_2 = \text{argmin}_\xi ||\Omega^{-1/2}(Y + D\xi)||_2$ has smaller variance than any other $\tilde{\xi}_p = \text{argmin}_\xi ||W(Y + D\xi)||_p$. Now, let us view the minimization problem $\min_\xi ||Y + D\xi||_p$ as generating estimators $\tilde{\xi}_p$ of the unknown parameter ξ, where $Y \sim N(-D\xi, \Omega)$ with known Ω. The result of Theorem 1.3 now states that all L_p-GMM estimators will be asymptotically unbiased, and the argument can be easily extended to show global unbiasedness of $\tilde{\xi}_p$ for ξ (as required for the application of the Cramér-Rao lower variance bound). The likelihood function

$$L(Y, D; \xi) = (2\pi)^{-q/2}|\Omega|^{-1/2}\exp\{-(1/2)(Y + D\xi)'\Omega^{-1}(Y + D\xi)\} \tag{1.39}$$

satisfies all the required regularity conditions for Cramér-Rao inequality (see Theil (1971), p.384). And it now follows that the asymptotic distribution of the optimal L_2-GMM estimator

$$\underset{\xi \in \mathbb{R}^k}{\text{argmin}} ||\Omega^{-1/2}(Y + D\xi)||_2 = -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}Y \tag{1.40}$$

attains the Cramér-Rao variance lower bound of $(D'\Omega^{-1}D)^{-1}$, since it equals the maximum likelihood estimator. The result then follows. □

Chapter 2

The Asymptotic Distribution of the Instrumental Variable Estimators When the Instruments Are Not Correlated with the Regressors

2.1 Introduction

A number of recent papers, including Bound, Jaeger and Baker (1995) and Staiger and Stock (1997), have considered instrumental variable (IV) estimators when the instruments are weak, in the sense that the correlation between the instruments and the regressors is low. In this chapter, we consider the extreme case that the instruments are completely irrelevant. In this case we can prove the following interesting result: the mean of the asymptotic distribution of the IV estimator is the same as the probability limit of the OLS estimator. Thus, as might be expected, irrelevant instruments do not remove the least squares bias.

To be specific, consider the linear model $y = X\beta + \varepsilon$ (in matrix notation) where ε is a $T \times 1$ random vector with mean zero, X is a $T \times K$ random matrix of regressors, and β is a $K \times 1$ parameter. It is well known that when X_t is correlated with ε_t, the ordinary least squares (OLS) estimator is not consistent. More specifically, under the regularity conditions that ensure the convergence of the statistics $T^{-1}X'X$ and $T^{-1}X'\varepsilon$ in probability, the OLS estimator converges in probability as $T \to \infty$ to $\beta_0 + (EX_tX_t')^{-1}EX_t\varepsilon_t$, which is different from β_0, the true parameter, unless $EX_t\varepsilon_t = 0$.

To obtain a consistent estimator, one possibility is instrumental variable estimation. Good instruments Z $(T \times L)$ are those which satisfy:

(*i*) $T^{-1}Z'Z$ converges in probability to a nonrandom, nonsingular matrix;

(*ii*) $T^{-1}Z'X$ converges in probability to a nonrandom matrix with full column rank;

(*iii*) $T^{-1/2}Z'\varepsilon$ converges in distribution to a normal random vector with zero mean.

When the instruments are good, the IV estimator is consistent and asymptotically normal.

Here we are concerned with the case that condition (*ii*) fails. Suppose that $L \geq K$, so that there are enough instruments, but the instruments (Z) are not strongly correlated with the regressors (X). Specifically, let the reduced form for X be:

$$X = Z\Pi + V. \tag{2.1}$$

Staiger and Stock (1997) consider the case that $\Pi = \Pi_T = C/\sqrt{T}$, with C a $L \times K$ matrix of constants. They call this the case of *weak instruments*. In this case the correlation between X_t and Z_t is of order $T^{-1/2}$, and condition (*ii*) fails. Staiger and Stock show that with weak instruments $\hat{\beta}_{IV}$, the IV estimator, does not have a probability limit but rather $\hat{\beta}_{IV} - \beta_0$ converges to a non-normal random variable. The mean of the asymptotic distribution of $\hat{\beta}_{IV} - \beta_0$ is non-zero, so that with weak instruments there is asymptotic bias. This bias is in the same direction as the bias of OLS.

In this chapter we consider the case of *irrelevant instruments*, which are uncorrelated with the regressors. This is a special case of Staiger and Stock, corresponding to $C = 0$ so that $\Pi = 0$ in the reduced form (2.1) for all T. In this case we show that the mean of the asymptotic distribution of $(\hat{\beta}_{IV} - \beta_0)$ is the same as $(\text{plim}\,\hat{\beta}_{OLS} - \beta_0)$, the asymptotic bias of the OLS estimator.

2.2 The limit distribution

Consider a linear model in matrix notation

$$y = X^{\circ}\beta + W\gamma + \varepsilon \tag{2.2}$$

where y and X° are respectively a $T \times 1$ vector of dependent variables and a $T \times K$ matrix of the endogenous regressors, W is a $T \times G$ matrix of exogenous regressors, the first column of which is a vector of ones, ε is the vector of errors, and β and γ are the parameters to be estimated.

Consider a $T \times L$ random matrix Z° of "instruments." For any matrix A with full column rank, let $P_A = A(A'A)^{-1}A'$. Let $X = (I - P_W)X^{\circ}$ and $Z = (I - P_W)Z^{\circ}$. Thus X is the part of the endogenous regressors not explained by the exogenous regressors, and similarly Z is the part of the "instruments" not explained by the exogenous regressors.

We make the following "high level" assumptions.

Assumption 2.1. *$T^{-1}X'X$, $T^{-1}\varepsilon'\varepsilon$, and $T^{-1}Z'Z$ converge in probability to finite, nonrandom, nonsingular matrices, and $T^{-1}X'\varepsilon$ converges to a nonrandom matrix.*

Let $\Sigma = \operatorname{plim} T^{-1}(X,\varepsilon)'(X,\varepsilon)$. It has submatrices Σ_{XX}, $\Sigma_{X\varepsilon}$, and $\sigma_{\varepsilon\varepsilon}$, which are the probability limits of $T^{-1}X'X$, $T^{-1}X'\varepsilon$, and $T^{-1}\varepsilon'\varepsilon$, respectively. Also let $\Omega = \operatorname{plim} T^{-1}Z'Z$. Assumption 2.1 can be regarded as the implication of a law of large numbers under more primitive assumptions on the sequences. For example, when the sequence $(\varepsilon_t, X_t^{\circ\prime}, Z_t^{\circ\prime})'$ is i.i.d. and its second moment exists, $\Sigma_{XX} = EX_t^{\circ}X_t^{\circ\prime} - EX_t^{\circ}W_t' \cdot (EW_tW_t')^{-1}EW_tX_t^{\circ\prime}$, $\Sigma_{X\varepsilon} = EX_t^{\circ}\varepsilon_t - EX_t^{\circ}W_t'(EW_tW_t')^{-1}EW_t\varepsilon_t = EX_t^{\circ}\varepsilon_t$, $\sigma_{\varepsilon\varepsilon} = E\varepsilon_t^2$, and $\Omega = EZ_t^{\circ}Z_t^{\circ\prime} - EZ_t^{\circ}W_t'(EW_tW_t')^{-1}EW_tZ_t^{\circ\prime}$.

Let $\rho = \Sigma_{XX}^{-1/2}\Sigma_{X\varepsilon}\sigma_{\varepsilon\varepsilon}^{-1/2}$ which is a multivariate correlation coefficient. A key assumption is the irrelevance of Z as instruments for X, as follows:

Assumption 2.2. $T^{-1/2}Z'(X,\varepsilon) \xrightarrow{d} \Omega^{1/2}(\xi\Sigma_{XX}^{1/2}, \eta\sigma_{\varepsilon\varepsilon}^{1/2})$ *where* $\text{vec}(\xi,\eta)$ *is a multivariate centered normal with* $E\,\text{vec}(\xi)\,\text{vec}(\xi)' = I$*,* $E\eta\eta' = I$ *and* $E\,\text{vec}(\xi)\eta' = \rho \otimes I$*.*

Note that Assumption 2.2 implies $T^{-1}Z'X \xrightarrow{p} 0$, which may agree with an intuitive definition of irrelevant instruments. Also, this assumption can be regarded as the implication of a central limit theorem under more primitive assumptions, as above.

Now let $\hat{\beta}_{IV}$ be the estimate of β in equation (2.2), when estimation is by IV using (Z°, W) as instruments. It is readily shown that

$$\hat{\beta}_{IV} - \beta_0 = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'\varepsilon. \tag{2.3}$$

By dividing $Z'Z$ by T, and $Z'X$ and $Z'\varepsilon$ by $T^{1/2}$, we observe that $\hat{\beta}_{IV} - \beta_0$ is a function φ of $(T^{-1}Z'Z, T^{-1/2}Z'X, T^{-1/2}Z'\varepsilon)$, where $\varphi : \mathbb{R}^{L\times L} \times \mathbb{R}^{L\times K} \times \mathbb{R}^{L\times 1} \to \mathbb{R}^{K\times 1}$ is defined by $\varphi(\Omega, A, b) = (A'\Omega^{-1}A)^{-1}A'\Omega^{-1}b$. Obviously, φ is measurable and is almost surely continuous in the limit. Here continuity is assured by the nonsingularity of the limit of $T^{-1}Z'Z$ and the almost sure full column rank of the limit of $T^{-1/2}Z'X$. Therefore, we apply the continuous mapping theorem to get the following result.

Theorem 2.3. *Under Assumptions 2.1 and 2.2,*

$$\hat{\beta}_{IV} \xrightarrow{d} \tilde{\beta}_{asy} = \beta_0 \dot{+} \Sigma_{XX}^{-1/2}(\xi'\xi)^{-1}\xi'\eta\sigma_{\varepsilon\varepsilon}^{1/2}, \tag{2.4a}$$

or equivalently,

$$\hat{\delta} = \Sigma_{XX}^{1/2}(\hat{\beta}_{IV} - \beta_0)\sigma_{\varepsilon\varepsilon}^{-1/2} \xrightarrow{d} \tilde{\delta}_{asy} = (\xi'\xi)^{-1}\xi'\eta. \tag{2.4b}$$

We note that the result in (2.4a) is the same as equation (2.5) of Staiger and Stock (1997, p.562) when $C = 0$ (and therefore $\lambda = 0$ in (2.3a) and (2.3b)).

We now calculate the density of $\tilde{\delta}_{asy}$ as follows.

Theorem 2.4. *Under Assumptions 2.1 and 2.2, the density of $\tilde{\delta}_{asy}$ is*

$$f(d) = C_{K,L} \cdot (1 - \rho'\rho)^{-K/2} \left| (I,\ d) \begin{pmatrix} I & \rho \\ \rho' & 1 \end{pmatrix}^{-1} \begin{pmatrix} I \\ d' \end{pmatrix} \right|^{-(L+1)/2} \tag{2.5}$$

where $C_{K,L} = 2^{-(L-1)(K-1)/2} \pi^{-K/2} \Gamma\left(\frac{L+1}{2}\right) \Gamma\left(\frac{L-K+1}{2}\right)^{-1}$.

Proof. See Appendix. □

Given K (the dimension of X_t) and L (the dimension of Z_t), the density depends upon ρ only. As is mentioned in Phillips (1980), this density is similar to the multivariate t distribution. The first moment of $\tilde{\delta}_{asy}$ exists as long as L is strictly greater than K, and more generally its integer moments exist up to the degree of over-identification. (See Phillips (1980, p. 870).)

2.3 The relationship with the OLS estimator

We are now in a position to prove our main result.

Theorem 2.5. *Suppose $L > K$. Then under Assumptions 2.1 and 2.2, the mean of $\tilde{\beta}_{asy}$ is equal to the probability limit of the OLS estimator.*

Proof. We observe that the density of $\tilde{\delta}_{asy}$ in (2.4b) is symmetric around ρ, the correlation coefficient of the endogenous regressors and the error. Furthermore, if $L > K$, the mean of $\tilde{\delta}_{asy}$ exists. Therefore, if $L > K$, $E\tilde{\delta}_{asy} = \rho$. Then

$$\begin{aligned} E\tilde{\beta}_{asy} &= \beta_0 + \Sigma_{XX}^{-1/2} E\tilde{\delta}_{asy}\sigma_{\varepsilon\varepsilon}^{1/2} \\ &= \beta_0 + \Sigma_{XX}^{-1/2} \rho\sigma_{\varepsilon\varepsilon}^{1/2} \\ &= \beta_0 + \Sigma_{XX}^{-1}\Sigma_{X\varepsilon} \\ &= \operatorname{plim} \hat{\beta}_{OLS}. \end{aligned} \tag{2.6}$$

An alternative proof that does not depend on the exact form of the density of $\tilde{\delta}_{asy}$ is as follows. When the mean of $\tilde{\delta}_{asy} = (\xi'\xi)^{-1}\xi'\eta$ exists,

$$E(\xi'\xi)^{-1}\xi'\eta = E(\xi'\xi)^{-1}\xi' E(\eta|\xi) \tag{2.7}$$

by the law of iterated expectations. But since $E\,\mathrm{vec}(\xi)\,\mathrm{vec}(\xi)'$, $E\,\mathrm{vec}(\xi)\eta'$, and $E\eta\eta'$ are respectively equal to $I \otimes I$, $\rho \otimes I$, and I,

$$E\eta|\,\mathrm{vec}(\xi) = (\rho' \otimes I)(I \otimes I)^{-1}\,\mathrm{vec}\,\xi = \xi\rho. \tag{2.8}$$

(For the operations involved with the Kronecker product and vec operators, see Magnus and Neudecker (1988, Ch. 2).) Hence, $E(\xi'\xi)^{-1}\xi'\eta = \rho$. It follows that $E\tilde{\beta}_{asy} = \beta_0 + \Sigma_{XX}^{-1/2} E\tilde{\delta}_{asy}\sigma_{\varepsilon\varepsilon}^{1/2} = \beta_0 + \Sigma_{XX}^{-1}\Sigma_{X\varepsilon}$, which is equal to the probability limit of the OLS estimator, as in the original proof. □

2.4 Conclusion

In this chapter, we answered some questions about the IV estimator using irrelevant instruments in linear models. We saw that the IV estimator is not consistent but converges to a nondegenerate distribution which is similar to a multivariate t distribution. When the number of instruments (excluding the exogenous regressors) is strictly greater than the number of endogenous regressors, the mean of the asymptotic distribution exists and is equal to the probability limit of the OLS estimator.

2.A Proof of Theorem 2.4

First, observe that the rows of the $L \times (K+1)$ matrix (ξ, η) are a random sample from $N(0, J)$ where $J = \left(\begin{smallmatrix} I & \rho \\ \rho' & 1 \end{smallmatrix}\right)$. Thus, $(\xi, \eta)'(\xi, \eta)$ has a $K+1$ dimensional central Wishart distribution with L degrees of freedom on the covariance matrix J. When $L \geq K+1$, its density at the point $\xi'\xi = B_1$, $\xi'\eta = b_2$, and $\eta'\eta = b_3$ is

$$\begin{aligned} g(B_1, b_2, b_3) = & 2^{-L(K+1)/2}\Gamma_{K+1}(\tfrac{L}{2})^{-1}(1-\rho'\rho)^{-L/2} \times \\ & |B|^{\frac{1}{2}(L-K-2)} \exp\{-\frac{1}{2} \operatorname{tr} J^{-1}B\} \end{aligned} \tag{2.9}$$

where $B = \left(\begin{smallmatrix} B_1 & b_2 \\ b_2' & b_3 \end{smallmatrix}\right)$ and Γ_n is the multivariate gamma function defined as

$$\Gamma_n(a) = \pi^{n(n-1)/4} \prod_{j=1}^{n} \Gamma(a - \tfrac{j-1}{2}). \tag{2.10}$$

(See Johnson and Kotz (1972, p.162).)

Following Phillips (1980), consider the one-to-one mapping ψ on the set of $K+1$ dimensional, real, symmetric, positive definite matrices defined as

$$\psi : \begin{pmatrix} B_1 & b_2 \\ b_2' & b_3 \end{pmatrix} \to \begin{pmatrix} B_1 & B_1^{-1}b_2 \\ b_2'B_1^{-1} & b_3 - b_2'B_1^{-1}b_3 \end{pmatrix}. \tag{2.11}$$

Then the inverse ψ^{-1} is

$$\psi^{-1} : \begin{pmatrix} A_1 & d \\ d' & a_3 \end{pmatrix} \to \begin{pmatrix} A_1 & A_1 d \\ d'A_1 & a_3 + d'A_1 d \end{pmatrix} \tag{2.12}$$

whose Jacobian turns out to be $|A_1|$. Therefore, by the change-of-variable technique, the density of the symmetric random matrix, which is defined such that the upper-left $K \times K$ diagonal block is $\xi'\xi$, the lower-right 1×1 diagonal block is $\eta'\eta - \eta'\xi(\xi'\xi)^{-1}\xi'\eta$, and the upper-right $K \times 1$ off-diagonal block is $\tilde{\delta}_{asy} = (\xi'\xi)^{-1}\xi'\eta$, evaluated at the point such that

$$\xi'\xi = A_1, \ (\xi'\xi)^{-1}\xi'\eta = d, \text{ and } \eta'\eta - \eta'\xi(\xi'\xi)^{-1}\xi'\eta = a_3, \tag{2.13}$$

where A_1 is symmetric, positive definite and a_3 is positive, becomes

$$\begin{aligned} h(A_1, d, a_3) &= g(A_1, A_1 d, a_3 + d' A_1 d) \cdot |A_1| \\ &= 2^{-L(K+1)/2} \Gamma_{K+1}(\tfrac{L}{2})^{-1} (1 - \rho'\rho)^{-L/2} \cdot H_1(A_1) \cdot H_3(a_3) \end{aligned} \tag{2.14}$$

where

$$H_1(S) = |S|^{(L-K)/2} \exp\{-\tfrac{1}{2} \operatorname{tr} S[I + (1-\rho'\rho)^{-1}(d-\rho)(d-\rho)']\} \tag{2.15}$$

and

$$H_3(x) = x^{(L-K)/2-1} \exp\{-\tfrac{1}{2} x (1-\rho'\rho)^{-1}\}. \tag{2.16}$$

The density of $\tilde{\delta}_{asy}$ at d is obtained by integrating out A_1 (symmetric and positive definite) and a_3 (positive) from (2.14). From the definition of the $\Gamma(\cdot)$ function, the integral of $H_3(a_3)$ in (2.14) over all positive a_3 is equal to

$$\int_0^\infty H_3(x)dx = 2^{(L-K)/2}(1-\rho'\rho)^{(L-K)/2}\Gamma(\tfrac{L-K}{2}) \tag{2.17}$$

The integral of the matrix argumented function $H_1(S)$ over all symmetric, positive definite matrices is obtained from the results in James (1964). Equations (25), (26), and (28) of James (1964, pp. 479–480) imply that for any nonsingular real symmetric $K \times K$ matrix D,

$$\int_{S>0} |S|^{a-\frac{1}{2}(K+1)} \exp\{-\operatorname{tr} SD\} dS = \Gamma_K(a)|D|^{-a} \tag{2.18}$$

where the integral is taken over all symmetric, positive definite $K \times K$ matrices. Thus, we have the evaluation

$$\begin{aligned} \int_{S>0} H_1(S)dS &= 2^{(L+1)/2}\Gamma_K(\tfrac{L+1}{2}) \times \\ &\quad |I + (1-\rho'\rho)^{-1}(d-\rho)(d-\rho)'|^{-(L+1)/2}. \end{aligned} \tag{2.19}$$

The desired density (2.5) is obtained by combining Equations (2.14), (2.17), and (2.19).

Chapter 3

Estimation of a Panel Data Model with Parametric Temporal Variation in Individual Effects

3.1 Introduction

In this chapter we consider the model:

$$y_{it} = X_{it}'\beta + Z_i'\gamma + \lambda_t(\theta)\alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N, \; t = 1, \ldots, T. \tag{3.1}$$

We treat T as fixed, so that "asymptotic" means as $N \to \infty$. The distinctive feature of the model is the interaction between the time-varying parametric function $\lambda_t(\theta)$ and the individual effect α_i. We consider the case that the α_i are "fixed effects," as will be discussed in more detail below. In this case estimation may be non-trivial due to the "incidental parameters problem" that the number of α's grows with sample size; see, for example, Chamberlain (1980).

Models of this form have been proposed and used in the literature on frontier productions functions (measurement of the efficiency of production). For example, Kumbhakar (1990) proposed the case that $\lambda_t(\theta) = [1 + \exp(\theta_1 t + \theta_2 t^2)]^{-1}$, and Battese and Coelli (1992) proposed the case that $\lambda_t(\theta) = \exp[-\theta(t - T)]$. Both of these papers considered random effects models in which α_i is independent of X and Z. In fact, both of these papers proposed specific (truncated normal) distributions for the α_i, with estimation by maximum likelihood. The aim of the present chapter is to provide a fixed-effects treatment of models of this type.

There is also a literature on the case that the λ_t themselves are treated as parameters. That is, the model becomes:

$$y_{it} = X_{it}'\beta + Z_i'\gamma + \lambda_t\alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T. \tag{3.2}$$

This corresponds to using a set of dummy variables for time rather than a parametric function $\lambda_t(\theta)$, and now $\lambda_t\alpha_i$ is just the product of fixed time and individual effects. This model has been considered by Kiefer (1980), Holtz-Eakin, Newey and Rosen (1988), Lee (1991), Chamberlain (1992), Lee and Schmidt (1993) and Ahn, Lee and Schmidt (2001), among others. Lee (1991) and Lee and Schmidt (1993) have applied this model to the frontier production function problem, in order to avoid having to assume a specific parametric function $\lambda_t(\theta)$. Another motivation for the model is that a fixed-effects version allows one to control for unobservables (e.g. macro events) that are the same for each individual, but to which different individuals may react differently.

Ahn, Lee and Schmidt (2001) establish some interesting results for the estimation of model (3.2). A generalized method of moments (GMM) estimator of the type considered by Holtz-Eakin, Newey and Rosen (1988) is consistent given exogeneity assumptions on the regressors X and Z. Least squares applied to (3.2), treating the α_i as fixed parameters, is consistent provided that the regressors are strictly exogenous and that the errors ε_{it} are white noise. The requirement of white noise errors for consistency of least squares is unusual, and is a reflection of the incidental parameters problem. Furthermore, if the errors are white noise, then a GMM estimator that incorporates the white noise assumption dominates least squares, in the sense of being asymptotically more efficient. This is also a somewhat unusual result, since in the usual linear model with normal errors, the moment conditions implied by the white noise assumption would not add to the efficiency of estimation.

The results of Ahn, Lee and Schmidt apply only to the case that the λ_t are unrestricted, and therefore do not apply to the model (3.1). However, in this chapter we show that essentially the same results do hold for the model (3.1). This enables us to use a parametric function $\lambda_t(\theta)$, and to test the validity of this assumption, while maintaining only weak assumptions on the α_i. This may be very useful, especially in the frontier production function setting. Applications using unrestricted λ_t have yielded temporal patterns of efficiency that seem unreasonably variable and in need of smoothing, which a parametric function can accomplish.

The plan of the chapter is as follows. Section 3.2 restates the model and lists our assumptions. Section 3.3 considers GMM estimation under basic exogeneity assumptions, while Section 3.4 considers GMM when we add the conditions implied by white noise errors. Section 3.5 considers least squares estimation and the sense in which it is dominated by GMM. Finally, Section 3.6 contains some concluding remarks.

3.2 The model and assumptions

The model is given in equation (3.1) above. We can rewrite it in matrix form, as follows. Let $y_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (X_{i1}, \ldots, X_{iT})'$, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$. Thus y_i is $T \times 1$, X_i is $T \times K$, ε_i is $T \times 1$, β is $K \times 1$, γ is $g \times 1$, and α_i is a scalar. (In this chapter, all the vectors are column vectors, and the data matrices are "vertically tall.") Define a function $\lambda : \Theta \to \mathbb{R}^T$, where Θ is a compact subset of $\mathbb{R}^p$, such that $\lambda(\theta) = (\lambda_1(\theta), \ldots, \lambda_T(\theta))'$. Note that T is fixed. In matrix form, our model is:

$$y_i = X_i\beta + 1_T Z_i'\gamma + \lambda(\theta)\alpha_i + \varepsilon_i, \quad i = 1, \ldots, N. \tag{3.3}$$

$\lambda(\theta)$ must be normalized in some way such as $\lambda(\theta)'\lambda(\theta) \equiv 1$ or $\lambda_1(\theta) \equiv 1$, to rule out trivial failure of identification arising from $\lambda(\theta) = 0$ or scalar multiplications of $\lambda(\theta)$. Here we choose the normalization $\lambda_1(\theta) \equiv 1$.

Let $W_i = (X'_{i1}, \ldots, X'_{iT}, Z'_i)'$. We make the following "orthogonality" and "covariance" assumptions.

Assumption 3.1 (Orthogonality). $E(W'_i, \alpha_i)'\varepsilon'_i = 0$.

Assumption 3.2 (Covariance). $E\varepsilon_i\varepsilon'_i = \sigma^2_\varepsilon I_T$.

Assumption 3.1 says that ε_{it} is uncorrelated with α_i, Z_i, and $X_{i1}, \ldots, X_{iT}$, and therefore contains an assumption of strict exogeneity of the regressors. Note that it does not restrict the correlation between α_i and $[Z_i, X_{i1}, \ldots, X_{iT}]$, so that we are in the fixed-effects framework. Assumption 3.2 asserts that the errors are white noise.

We also assume the following regularity conditions.

Assumption 3.3 (Regularity).

(i) $(W'_i, \alpha_i, \varepsilon'_i)'$ *is independently and identically distributed over* i*;*

(ii) ε_i *has finite fourth moment, and* $E\varepsilon_i = 0$*;*

(iii) $(W'_i, \alpha_i)'$ *has finite nonsingular second moment matrix;*

(iv) $EW_i(Z'_i, \alpha_i)$ *is of full column rank;*

(v) $\lambda(\theta)$ *is twice continuously differentiable in* θ.

The first four of these conditions correspond to assumptions (BA.1)–(BA.4) of Ahn, Lee and Schmidt (2001), who give some explanation. Condition (*v*) is new, and self-explanatory.

3.3 GMM under the Orthogonality Assumption

Let $u_{it} = u_{it}(\beta, \gamma) = y_{it} - X_{it}'\beta - Z_i'\gamma$, and $u_i = u_i(\beta, \gamma) = (u_{i1}, \ldots, u_{iT})'$. Since $u_{it} = \lambda_t(\theta)\alpha_i + \varepsilon_{it}$, it follows that $u_{it} - \lambda_t(\theta)u_{i1} = \varepsilon_{it} - \lambda_t(\theta)\varepsilon_{i1}$, which does not depend on α_i. This is a sort of generalized within transformation to remove the individual effects. The Orthogonality Assumption (Assumption 3.1) then implies the following moment conditions:

$$EW_i[u_{it}(\beta, \gamma) - \lambda_t(\theta)u_{i1}(\beta, \gamma)] = 0, \ t = 2, \ldots, T. \tag{3.4}$$

These moment conditions can be written in matrix form, as follows. Define $G(\theta) = [-\lambda_*(\theta), I_{T-1}]'$, where $\lambda_* = (\lambda_2, \ldots, \lambda_T)'$. The generalized within transformation corresponds to multiplication by $G(\theta)'$, and the moment conditions (3.4) can equivalently be written as follows:

$$Eb_{1i}(\beta, \gamma, \theta) = E[G(\theta)'u_i(\beta, \gamma) \otimes W_i] = 0. \tag{3.5}$$

(This corresponds to equation (7) of Ahn, Lee and Schmidt (2001), but looks slightly different because our W_i is a column vector whereas theirs is a row vector.) This is a set of $(T-1)(TK+g)$ moment conditions.

Some further analysis is needed to establish that (3.5) contains *all* of the moment conditions implied by the Orthogonality Assumption. Let $\Sigma_{WW} = EW_iW_i'$, $\Sigma_{W\alpha} = EW_i\alpha_i$, and $\sigma_\alpha^2 = E\alpha_i^2$. Given the model (3.3), the Orthogonality Assumption holds if and only if the following moment conditions hold:

$$E[u_i(\beta, \gamma) \otimes W_i - \lambda(\theta) \otimes \Sigma_{W\alpha}] = 0. \tag{3.6}$$

We could use these moment conditions as the basis for GMM estimation. Alternatively, we can remove the parameter $\Sigma_{W\alpha}$ by applying a nonsingular linear transformation to (3.6) in such a way that the transformed set of moment conditions is separated into two subsets, where the first subset does not contain $\Sigma_{W\alpha}$ and the second subset is exactly identified for $\Sigma_{W\alpha}$, given (β, γ, θ). The following transformation accomplishes this.

$$E \begin{bmatrix} G' \otimes I_d \\ \lambda' \otimes I_d \end{bmatrix} [u_i \otimes W_i - \lambda \otimes \Sigma_{W\alpha}] = 0 \tag{3.7}$$

where $d \equiv TK + g$ for notational simplicity; similarly, G, λ and u_i are shortened expressions for $G(\theta)$, $\lambda(\theta)$ and $u_i(\beta, \gamma)$. This is a nonsingular transformation, since (G, λ) is nonsingular, and therefore GMM based on (3.7) is asymptotically equivalent to GMM based on (3.6). Now split (3.7) into its two parts:

$$E(G' u_i \otimes W_i) = 0 \tag{3.8}$$

$$E(\lambda' u_i) W_i - (\lambda' \lambda) \Sigma_{W\alpha} = 0. \tag{3.9}$$

Here (3.9) is exactly identified for $\Sigma_{W\alpha}$, given β, γ and θ, in the sense that the number of moment conditions in (3.9) is the same as the dimension of $\Sigma_{W\alpha}$. Also $\Sigma_{W\alpha}$ does not appear in (3.8). It follows (e.g., Ahn and Schmidt (1995), Theorem 1) that the GMM estimates of β, γ and θ from (3.8) alone are the same as the GMM estimates of β, γ and θ if we use both (3.8) and (3.9), and estimate the full set of parameters $(\beta, \gamma, \theta, \Sigma_{W\alpha})$. But (3.8) is the same as (3.5), which establishes that (3.5) contains all the useful information about β, γ and θ implied by the Orthogonality Assumption.

Let $\bar{b}_1(\beta, \gamma, \theta) = N^{-1} \sum_{i=1}^{N} b_{1i}(\beta, \gamma, \theta)$. Then the optimal GMM estimator $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\theta}$ based on the Orthogonality Assumption solves the problem

$$\min_{\beta, \gamma, \theta} N \bar{b}_1(\beta, \gamma, \theta)' V_{11}^{-1} \bar{b}_1(\beta, \gamma, \theta) \tag{3.10}$$

where $V_{11} = E b_{1i} b_{1i}'$ evaluated at the true parameters. As usual, V_{11} can be replaced by any consistent estimate. A standard estimate would be

$$\hat{V}_{11} = \frac{1}{N} \sum_{i=1}^{N} b_{1i}(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta}) b_{1i}(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})' \tag{3.11}$$

where $(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})$ is an initial consistent estimate of (β, γ, θ) such as GMM using identity weighting matrix. Under certain regularity conditions (Hansen (1982), Assumption 3) the resulting GMM estimator is $\sqrt{N}$-consistent and asymptotically normal.

To express the asymptotic variance of the GMM estimator analytically, we need a little more notation. Let S_X be the $T(TK+g) \times K$ selection matrix such that $X_i = (I_T \otimes W_i)'S_X$, and let S_Z be the $T(TK+g) \times g$ selection matrix such that $1_T Z_i' = (I_T \otimes W_i)'S_Z$. S_X and S_Z have the following forms:

$$S_X = (I_K \; O \; \cdots \; O \; O_{K\times g} \vdots O \; I_K \; \cdots \; O \; O_{K\times g} \vdots \cdots \vdots O \; O \; \cdots \; I_K \; O_{K\times g})' \tag{3.12}$$

$$S_Z = (O_{g\times K} \; \cdots \; O_{g\times K} \; I_g \vdots \cdots \vdots O_{g\times K} \; \cdots \; O_{g\times K} \; I_g)' = 1_T \otimes (O_{g\times TK}, I_g)' \tag{3.13}$$

where O's without dimension subscript stand for $O_{K\times K}$. Define $\Lambda_* = \partial\lambda_*(\theta_0)/\partial\theta'$. The variance of the asymptotic distribution of the GMM estimates of β, γ and θ equals $(B_1' V_{11}^{-1} B_1)^{-1}$ where $V_{11} = Eb_{1i}b_{1i}'$ as above and

$$B_1 = [(G \otimes \Sigma_{WW})'S_X, (G \otimes \Sigma_{WW})'S_Z, \Lambda_* \otimes \Sigma_{W\alpha}]. \tag{3.14}$$

This result can be obtained either by direct calculation, or by applying the chain rule to B_1 calculated in Ahn, Lee and Schmidt (2001, p. 251). This asymptotic variance form is obtained from the Orthogonality Assumption only and does not need any further assumption.

A practical problem with this GMM procedure is that it is based on a rather large set of moment conditions. Some considerable simplifications are possible if we make the following assumption of no conditional heteroskedasticity (NCH) of ε_i:

$$E(\varepsilon_i \varepsilon_i' | W_i) = \Sigma_{\varepsilon\varepsilon}. \tag{NCH}$$

Under the NCH assumption,

$$V_{11} = E[G(\theta_0)'\varepsilon_i\varepsilon_i' G(\theta_0) \otimes W_i W_i'] = G(\theta_0)'\Sigma_{\varepsilon\varepsilon}G(\theta_0) \otimes \Sigma_{WW}. \tag{3.15}$$

Σ_{WW} can be consistently estimated by $\hat{\Sigma}_{WW} = N^{-1}\sum_{i=1}^{N} W_i W_i'$. Also, for any sequence (β_N, γ_N) that converges in probability to (β_0, γ_0), we have

$$\frac{1}{N}\sum_{i=1}^{N} u_i(\beta_N, \gamma_N)u_i(\beta_N, \gamma_N)' \xrightarrow{p} \Sigma_{\varepsilon\varepsilon} + \sigma_\alpha^2 \lambda(\theta_0)\lambda(\theta_0)'. \tag{3.16}$$

Since $G(\theta)'\lambda(\theta) = 0$, for any initial consistent estimate $(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})$,

$$G(\tilde{\theta})' \left(N^{-1}\sum_{i=1}^{N} u_i(\tilde{\beta}, \tilde{\gamma})u_i(\tilde{\beta}, \tilde{\gamma})' \right) G(\tilde{\theta}) \tag{3.17}$$

will consistently estimate $G(\theta_0)'\Sigma_{\varepsilon\varepsilon}G(\theta_0)$. Thus it is easy to construct a consistent estimate of V_{11} as given in (3.15).

In order to consistently estimate the asymptotic variance under NCH, we need to estimate Σ_{WW}, $\Sigma_{W\alpha}$, and $G'\Sigma_{\varepsilon\varepsilon}G$. Estimation of Σ_{WW} and $G'\Sigma_{\varepsilon\varepsilon}G$ was discussed above. We can obtain an estimate of $\Sigma_{W\alpha}$ from the GMM problem (3.7). A direct algebraic calculation gives us that

$$\hat{\Sigma}_{W\alpha} = \frac{1}{N}\sum_{i=1}^{N} W_i \frac{\hat{\lambda}'\hat{u}_i}{\hat{\lambda}'\hat{\lambda}} - \frac{1}{N}\sum_{i=1}^{N} W_i [\widehat{\lambda'\Sigma_{\varepsilon\varepsilon}G}\,(\widehat{G'\Sigma_{\varepsilon\varepsilon}G})^{-1}\hat{G}'\hat{u}_i]/(\hat{\lambda}'\hat{\lambda}) \tag{3.18}$$

where $\hat{u}_i = u_i(\hat{\beta}, \hat{\gamma})$, $\hat{\lambda} = \lambda(\hat{\theta})$, $\hat{G} = G(\hat{\theta})$, and $\widehat{\lambda'\Sigma_{\varepsilon\varepsilon}G}$ is a consistent estimate of $\lambda'\Sigma_{\varepsilon\varepsilon}G$, one possibility of which is $N^{-1}\sum_{i=1}^{N} \hat{\lambda}'\hat{u}_i\hat{u}_i'\hat{G}$.

Finally, under the NCH assumption, the set of moment conditions (3.5) can be converted into an *exactly identified* set of moment conditions that yield an asymptotically equivalent GMM estimate. Specifically, we can replace the moment conditions $Eb_{1i} = 0$ by the moment conditions $EB_1'V_{11}^{-1}b_{1i} = 0$. Routine calculation using the forms of B_1, V_{11} and b_{1i} yields the explicit expression:

$$EX_i'G(G'\Sigma_{\varepsilon\varepsilon}G)^{-1}G'u_i = 0 \tag{3.19a}$$

$$EZ_i 1_T' G(G'\Sigma_{\varepsilon\varepsilon}G)^{-1}G'u_i = 0 \tag{3.19b}$$

$$E\Sigma_{W\alpha}'\Sigma_{WW}^{-1}W_i \cdot \Lambda_*'(G'\Sigma_{\varepsilon\varepsilon}G)^{-1}G'u_i = 0. \tag{3.19c}$$

These three sets of moment conditions respectively correspond to (21a), (21b), and (21c) of Ahn, Lee and Schmidt (2001, p. 229). We can replace the nuisance parameters $\Sigma_{\varepsilon\varepsilon}$, $\Sigma_{W\alpha}$ and Σ_{WW} by consistent estimates, as given above (based on some initial consistent GMM estimates of β, γ and θ). The point of this simplification is that we have drastically reduced the set of moment conditions: there are $(T-1)(TK+g)$ moment conditions in b_{1i} (equation (3.5)) but only $K+g+p$ moment conditions in (3.19).

We note that this is a stronger result than the corresponding result (Proposition 1, p. 229) of Ahn, Lee and Schmidt (2001). In order to reach essentially the same conclusion on the reduction of the number of moment conditions, they impose the assumption that ε_i is independent of (W_i, α_i), a much stronger assumption than our NCH assumption.

3.4 GMM under the Orthogonality and Covariance Assumptions

In this section we continue to maintain the Orthogonality Assumption (Assumption 3.1), but now we add the Covariance Assumption (Assumption 3.2), which asserts that $E\varepsilon_i\varepsilon_i' = \sigma_\varepsilon^2 I_T$.

Clearly the Covariance Assumption holds if and only if

$$E(u_i u_i') = \sigma_\alpha^2 \lambda\lambda' + \sigma_\varepsilon^2 I_T. \tag{3.20}$$

Condition (3.20) contains $T(T+1)/2$ distinct moment conditions. It also contains the two nuisance parameters σ_α^2 and σ_ε^2, and so it should imply $T(T+1)/2-2$ moment conditions for the estimation of β, γ and θ. These are in addition to the moment conditions (3.5) implied by the Orthogonality Assumption.

To write these moment conditions explicitly, we need to define some notation. Let $H = \text{diag}(H_2, H_3, \ldots, H_T)$, with H_t equal to the $T \times (T-t)$ matrix of the last $T-t$

columns (the $(t+1)$th through Tth columns) of I_T for $t < T$, and with H_T equal to a $T \times (T-2)$ matrix of the second through $(T-1)$-th columns of I_T.[1] Then we can write the distinct moment conditions implied by the Orthogonality and Covariance Assumptions as follows:

$$Eb_{1i} = E(G'u_i \otimes W_i) = 0 \tag{3.21a}$$

$$Eb_{2i} = EH'(G'u_i \otimes u_i) = 0 \tag{3.21b}$$

$$Eb_{3i} = E\big[G'u_i \otimes \frac{\lambda' u_i}{\lambda'\lambda}\big] = 0. \tag{3.21c}$$

(In these expressions, G is short for $G(\theta)$, λ is short for $\lambda(\theta)$, and u_i is short for $u_i(\beta, \gamma)$.)

The moment conditions b_{1i} in (3.21a) are exactly the same as those in (3.5) of the previous section, and follow from the Orthogonality Assumption.

The moment conditions b_{2i} in (3.21b) correspond to those in equation (12) of Ahn, Lee and Schmidt (2001). Note that it is not the case that $E(G'u_i \otimes u_i) = 0$. Rather, looking at a typical element of this product, we have $E(u_{it} - \lambda_t u_{i1})u_{is}$, which equals zero for $s \neq t$ and $s \neq 1$. The selection matrix H' picks out the logically distinct products of expectation zero, the number of which equals $T(T-1)/2 - 1$. The selection matrix H plays the same role as the definition of the matrices U_{it}° plays in Ahn, Lee and Schmidt (2001). We note that the moment conditions b_{2i} follow from the non-autocorrelation of the ε_{it}; homoskedasticity would not be needed.

The $(T-1)$ moment conditions in b_{3i} in (3.21c) correspond to those in equation (13) of Ahn, Lee and Schmidt (2001). They assert that, for $t = 2, \ldots, T$, $E(u_{it} - \lambda_t u_{i1})(\sum_{s=1}^{T} \lambda_s u_{is}) = 0$, and their validity depends on both the non-autocorrelation and the homoskedasticity of the ε_{it}.

[1] For any matrix B with T rows, $H_t'B$ selects the last $T-t$ rows of B for $t < T$, and $H_T'B$ selects the second through $(T-1)$-th rows of B. For any matrix B with T columns, BH_t selects the last $T-t$ columns of B for $t < T$, and BH_T selects the second through $(T-1)$-th columns of B.

Some further analysis may be useful to establish that (3.21b) and (3.21c) represent all of the useful implications of the Covariance Assumption. We begin with the implication (3.20) of the Covariance Assumption, which we rewrite as

$$E(u_i \otimes u_i) = \sigma_\alpha^2(\lambda \otimes \lambda) + \sigma_\varepsilon^2 \text{vec} I_T. \tag{3.22}$$

Now, let S be the $T^2 \times T(T+1)/2$ selection matrix such that, for a $T \times 1$ vector u, $\text{vech}(uu') = S'(u \otimes u)$, where "vech" is the vector of distinct elements. Then

$$ES'(u \otimes u) = S'[\sigma_\alpha^2(\lambda \otimes \lambda) + \sigma_\varepsilon^2 \text{vec} I_T] \tag{3.23}$$

contains the distinct moment conditions.

Now we transform the moment conditions (3.23) by multiplying them by a nonsingular matrix, in such a way that (i) the first $T(T+1)/2 - 2$ transformed moment conditions are those given in (3.21b) and (3.21c); and (ii) the last two moment conditions are exactly identified for the nuisance parameters (σ_α^2 and σ_ε^2), given the other parameters. This will imply that the last two moment conditions are redundant for the estimation of β, γ and θ, and thus that (3.21b) and (3.21c) contain all of the useful information implied by the Covariance Assumption for estimation of β, γ and θ.

To exhibit the transformation, let G_t be the $(t-1)$th column of G; let e_t^* equal the tth column of I_{T-2} and e_T equal the last column of I_T; and define

$$(H_T^{**})' = [-\lambda_T H_T', \; e_1^* e_T', \; \ldots, \; e_{T-2}^* e_T', \; O_{(T-2)\times T}]. \tag{3.24}$$

(H_T was defined above.) Then

$$[G_2 \otimes H_2, \; \ldots, \; G_{T-1} \otimes H_{T-1}, \; H_T^{**}]'S \cdot S'(u_i \otimes u_i) = H'(G' \otimes I_T)(u_i \otimes u_i), \tag{3.25}$$

which is the same as in b_{2i} in (3.21b). Also, let $J_1^* = I_T - \lambda\lambda'$ and J_t^*, $t = 2, \ldots, T$, is equal to $\text{diag}\{O_{t\times t}, \lambda_t I_{T-t}\}$ plus a $T \times T$ matrix with zero elements except for the tth row which is λ'. Then

$$H_1'[J_1^*, \ldots, J_T^*]S \cdot S'(u_i \otimes u_i) = (\lambda' \otimes G')(u_i \otimes u_i), \tag{3.26}$$

which is equal to b_{3i} in (3.21c).

The point of the above argument is that the transformations preceding $S'(u_i \otimes u_i)$ in (3.25) and (3.26), stacked vertically, construct a $[T(T+1)/2 - 2] \times T(T+1)/2$ matrix of full row rank, and yield the moment conditions b_{2i} and b_{3i}. The remaining two moment conditions that determine the nuisance parameters are

$$E\begin{bmatrix} u_{i1}^2 \\ u_{i2}u_{i1} \end{bmatrix} = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 \\ \lambda_2\sigma_\alpha^2 \end{bmatrix} \tag{3.27}$$

and must be linearly independent of the others (since they involve σ_α^2 and σ_ε^2 while the others do not).

The asymptotic variance of the GMM estimate is complicated because it depends on the moments of ε_{it} up to fourth order. However, we can simplify things with the following "conditional independence of the moments up to fourth order" (CIM4) assumption:

Conditional on (W_i, α_i), ε_{it} is independent over $t = 1, 2, \ldots, T$, with mean zero, and with second, third and fourth moments that do not depend on (W_i, α_i) or on t. (CIM4)

This is a strong assumption; it implies the Orthogonality Assumption, the Covariance Assumption, the NCH assumption, and more. In Appendix A, we calculate the asymptotic variance matrix of the GMM estimate based on (3.21) under the assumption (CIM4).

Let $\Lambda = \partial\lambda(\theta_0)/\partial\theta$ and note that $\Lambda_* = G'\Lambda$. Given assumption (CIM4), the moment conditions (3.19), which are asymptotically equivalent to (3.21a), can be simplified as follows:

$$EX_i' P_G u_i = 0 \tag{3.28a}$$

$$EZ_i 1_T' P_G u_i = 0 \tag{3.28b}$$

$$E\Sigma_{W\alpha}' \Sigma_{WW}^{-1} W_i \cdot \Lambda' P_G u_i = 0. \tag{3.28c}$$

That is, in place of the large set of moment conditions (3.21a), (3.21b) and (3.21c), we can use the reduced set of moment conditions consisting of (3.28), (3.21b) and (3.21c).

A final simplification arises if, conditional on (W_i, α_i), ε_{it} is i.i.d. normal. In this case, (3.21b) can be shown to be redundant given (3.21a) and (3.21c). (See Proposition 4 of Ahn, Lee and Schmidt (2001, p. 231).) Hence, in that case, the GMM estimator using the moment conditions (3.28) and (3.21c) is efficient.

3.5 Least Squares

In this section we consider the concentrated least squares (CLS) estimation of the model. We treat the α_i as parameters to be estimated, so this is a true "fixed effects" treatment. We can consider the following least squares problem:

$$\min_{\beta,\gamma,\theta,\alpha_1,\dots,\alpha_N} N^{-1}\sum_{i=1}^{N}[y_i - X_i\beta - 1_T Z_i'\gamma - \lambda(\theta)\alpha_i]'[y_i - X_i\beta - 1_T Z_i'\gamma - \lambda(\theta)\alpha_i]. \tag{3.29}$$

Solving for $\alpha_1, \dots, \alpha_N$ first, we get

$$\alpha_i(\beta, \gamma, \theta) = [\lambda(\theta)'\lambda(\theta)]^{-1}\lambda(\theta)' u_i(\beta, \gamma) \quad i = 1, \dots, N. \tag{3.30}$$

where $u_i(\beta, \gamma) = y_i - X_i\beta - 1_T Z_i'\gamma$ as before. Then the estimates $\hat{\beta}_{LS}$, $\hat{\gamma}_{LS}$, and $\hat{\theta}_{LS}$ minimizing (3.29) are equal to the minimizers of the *sum of the squared concentrated residuals*

$$\bar{C}(\beta, \gamma, \theta) = N^{-1}\sum_{i=1}^{N} C_i(\beta, \gamma, \theta) = N^{-1}\sum_{i=1}^{N} u_i(\beta, \gamma)' M_{\lambda(\theta)} u_i(\beta, \gamma) \tag{3.31}$$

which is obtained by replacing α_i in (3.29) with (3.30). From the name of (3.31), we call $\hat{\beta}_{LS}$, $\hat{\gamma}_{LS}$ and $\hat{\theta}_{LS}$ the *concentrated least squares estimator.*

Since $G'\lambda = 0$, we have $M_\lambda G = G$ and therefore $M_\lambda = P_G = G(G'G)^{-1}G'$. So the first order conditions of the CLS estimation become

$$\begin{bmatrix} \partial\bar{C}/\partial\beta \\ \partial\bar{C}/\partial\gamma \\ \partial\bar{C}/\partial\theta \end{bmatrix} = -\frac{2}{N}\sum_{i=1}^{N}\begin{bmatrix} X_i'P_G u_i \\ Z_i 1_T' P_G u_i \\ \Lambda' P_G u_i u_i' \lambda(\lambda'\lambda)^{-1} \end{bmatrix} = 0. \qquad (3.32)$$

Interpreting (3.32) as sample moment conditions, we can construct the corresponding (exactly identified) implicit population moment conditions:

$$EX_i'P_G u_i = 0 \qquad (3.33a)$$

$$EZ_i 1_T' P_G u_i = 0 \qquad (3.33b)$$

$$E\Lambda' P_G u_i u_i' \lambda(\lambda'\lambda)^{-1} = 0. \qquad (3.33c)$$

That is, the CLS estimator is asymptotically equivalent to the GMM estimator based on (3.33).

The moment conditions (3.33a) and (3.33b) are satisfied under the Orthogonality Assumption. However, this is not true of (3.33c). The moment conditions (3.33c) require the Covariance Assumption to be valid (unless we make very specific and unusual assumptions about the form of λ and its relationship to the error variance matrix). Thus, the consistency of the CLS estimator requires *both* the Orthogonality Assumption *and* the Covariance Assumption. This is a rather striking result, since the consistency of least squares does not usually require restrictions on the second moments of the errors, and is a reflection of the incidental parameters problem.

We would generally believe that least squares should be efficient when the errors are i.i.d. normal. However, similarly to the result in Ahn, Lee and Schmidt (2001), this is not true in the present case. The efficient GMM estimator under the Orthogonality and Covariance Assumptions uses the moment conditions (3.21), while the CLS estimator uses only a subset of these. This can be seen most explicitly in the case that, conditional on (W_i, α_i), the ε_{it} are i.i.d. normal. Then (3.21b) is redundant and (3.21a) can be replaced by (3.28), so that the efficient GMM estimator is based on (3.28a), (3.28b), (3.28c) and (3.21c). The CLS estimator is based on (3.33a), which is the same as (3.28a); (3.33b), which is the same as (3.28b); and (3.33c), which is a subset of (3.21c).[2] So the inefficiency of CLS lies in its failure to use the moment conditions (3.28c) and from its failure to use all of the moment conditions in (3.21c). The latter failure did not arise in the Ahn, Lee and Schmidt (2001) analysis (see footnote 2).

In Appendix B, we calculate the asymptotic variance matrix of the CLS estimator, under the "conditional independence of the moments up to fourth order" (CIM4) assumption of Section 3.4.

3.6 Conclusion

In this chapter we have considered a panel data model with parametrically time-varying coefficients on the individual effects. Following Ahn, Lee and Schmidt (2001), we have enumerated the moment conditions implied by alternative sets of assumptions on the model. We have shown explicitly that our sets of moment conditions capture all of the useful information contained in our assumptions, so that the corresponding GMM estimators exploit these assumptions efficiently.

We have also considered concentrated least squares estimation. Here the incidental

[2]The moment conditions (3.33c) are equivalent to $E\Lambda' G(G'G)^{-1} b_{3i} = 0$. When the number of parameters in θ is less than $T - 1$, the transformation $\Lambda' G(G'G)^{-1}$ loses information. This will be so in most parametric models for $\lambda(\theta)$, though it is not true in the model of Ahn, Lee and Schmidt (2001).

parameters problem is relevant because we are treating the fixed effects as parameters to be estimated. An interesting result is that the consistency of the least squares estimator requires both exogeneity assumptions and the assumption that the errors are white noise. Furthermore, given the white noise assumption, the least squares estimator is inefficient, because it fails to exploit all of the moment conditions that are available.

We show how the GMM estimation problem can be simplified under some additional assumptions, including the assumption of no conditional heteroskedasticity and a stronger conditional independence assumption. Under these assumptions we also give explicit expressions for the variance matrices of the GMM and least squares estimators.

APPENDIX

In this Appendix we derive the asymptotic variances of the efficient GMM estimator and the CLS estimator. We make the "conditional independence of the moments up to fourth order" (CIM4) assumption of Section 3.4.

3.A The asymptotic variance of the GMM estimator

Under the Orthogonality and Covariance Assumptions, the moment conditions we have are $b_{1i} = G'u_i \otimes W_i$, $b_{2i} = H'(G'u_i \otimes u_i)$, and $b_{3i} = (\lambda'\lambda)^{-1}\lambda' u_i \otimes G'u_i$. Let $\delta = (\beta', \gamma', \theta')'$. Let $B_j = -E(\partial b_{ji}/\partial\delta)$ for $j = 1, 2, 3$, evaluated at the true parameters. Let $V_{jk} = Eb_{ji}b'_{ki}$ for $j, k = 1, 2, 3$, evaluated at the true parameters. Define $\kappa_3 = E\varepsilon_{it}^3/\sigma_\varepsilon^2$ and $\kappa_4 = E(\varepsilon_{it}^4 - 3\sigma_\varepsilon^4)/\sigma_\varepsilon^2$. Let $\mu_W = EW_i$; $\Phi = \Phi(\theta) = \lambda_*\lambda_*' + \mathrm{diag}(\lambda_2, \ldots, \lambda_T)$; and $\Phi_* = \lambda_*\lambda_*' + \mathrm{diag}(\lambda_2^2, \ldots, \lambda_T^2)$, where $\lambda_* = (\lambda_2, \ldots, \lambda_T)'$. After some algebra, we get

$$V_{11} = \sigma_\varepsilon^2(G'G \otimes \Sigma_{WW}) \tag{3.34}$$

$$V_{12} = \sigma_\varepsilon^2(G'G \otimes \Sigma_{W\alpha}\lambda')H \tag{3.35}$$

$$V_{13} = \sigma_\varepsilon^2\left[G'G \otimes \Sigma_{W\alpha} + \frac{\kappa_3}{\lambda'\lambda}(\Phi \otimes \mu_W)\right] \tag{3.36}$$

$$V_{22} = \sigma_\varepsilon^2 H'[G'G \otimes (\sigma_\alpha^2\lambda\lambda' + \sigma_\varepsilon^2 I_T)]H \tag{3.37}$$

$$V_{23} = \sigma_\varepsilon^2 H'\left\{\left[\left(\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{\lambda'\lambda}\right)G'G + \frac{\kappa_3}{\lambda'\lambda}\mu_\alpha\Phi\right] \otimes \lambda\right\} \tag{3.38}$$

$$V_{33} = \sigma_\varepsilon^2\left\{\left(\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{\lambda'\lambda}\right)G'G + 2\frac{\kappa_3}{\lambda'\lambda}\mu_\alpha\Phi + \frac{\kappa_4}{(\lambda'\lambda)^2}\Phi_*\right\} \tag{3.39}$$

and

$$B_1 = [(G \otimes \Sigma_{WW})'S_X,\ (G \otimes \Sigma_{WW})'S_Z,\ \Lambda_* \otimes \Sigma_{W\alpha}] \tag{3.40}$$

$$B_2 = H'(I_{T-1} \otimes \lambda)[(G \otimes \Sigma_{W\alpha})'S_X,\ (G \otimes \Sigma_{W\alpha})'S_Z,\ \sigma_\alpha^2\Lambda_*] \tag{3.41}$$

$$B_3 = [(G \otimes \Sigma_{W\alpha})'S_X,\ (G \otimes \Sigma_{W\alpha})'S_Z,\ \sigma_\alpha^2\Lambda_*]. \tag{3.42}$$

With these results, the variance-covariance of the GMM estimator is

$$\operatorname{cov}\sqrt{N}(\hat{\delta}-\delta) = \left[(B_1',\ B_2',\ B_3') \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{12}' & V_{22} & V_{23} \\ V_{13}' & V_{23}' & V_{33} \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix} \right]^{-1}. \tag{3.43}$$

3.B The asymptotic variance of the CLS estimator

By the standard Taylor series expansion technique, we find that the asymptotic variance will be equal to $A_0 B_0^{-1} A_0$ where

$$A_0 = E\frac{\partial^2 C_i}{\partial\delta\partial\delta'}, \quad \text{and} \quad B_0 = E\frac{\partial C_i}{\partial\delta}\frac{\partial C_i}{\partial\delta'} \tag{3.44}$$

evaluated at the true parameter. Let us calculate each of them. Let $\Lambda = \partial\lambda(\theta_0)/\partial\theta' = (0_{p\times 1},\ \Lambda_*')'$. B_0 is the same as in Ahn, Lee and Schmidt (2001, p. 253). Let $\Psi = G(G'G)^{-1}\Phi\cdot(G'G)^{-1}G'$; $\Psi_* = G(G'G)^{-1}\Phi_*(G'G)^{-1}G'$; and $\mu_\alpha = E\alpha_i$. Then

$$E\frac{\partial C_i}{\partial\beta}\frac{\partial C_i}{\partial\beta'} = 4\sigma_\varepsilon^2 S_X'(P_G\otimes\Sigma_{WW})S_X \tag{3.45}$$

$$E\frac{\partial C_i}{\partial\beta}\frac{\partial C_i}{\partial\gamma'} = 4\sigma_\varepsilon^2 S_X'(P_G\otimes\Sigma_{WW})S_Z \tag{3.46}$$

$$E\frac{\partial C_i}{\partial\beta}\frac{\partial C_i}{\partial\theta'} = 4\sigma_\varepsilon^2 S_X'\left[P_G\otimes\Sigma_{W\alpha} + \frac{\kappa_3}{\lambda'\lambda}(\Psi\otimes\mu_W)\right]\Lambda \tag{3.47}$$

$$E\frac{\partial C_i}{\partial\gamma}\frac{\partial C_i}{\partial\gamma'} = 4\sigma_\varepsilon^2 S_Z'(P_G\otimes\Sigma_{WW})S_Z \tag{3.48}$$

$$E\frac{\partial C_i}{\partial\gamma}\frac{\partial C_i}{\partial\theta'} = 4\sigma_\varepsilon^2 S_Z'\left[P_G\otimes\Sigma_{W\alpha} + \frac{\kappa_3}{\lambda'\lambda}(\Psi\otimes\mu_W)\right]\Lambda \tag{3.49}$$

$$E\frac{\partial C_i}{\partial\theta}\frac{\partial C_i}{\partial\theta'} = 4\sigma_\varepsilon^2\Lambda'\left\{\left(\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{\lambda'\lambda}\right)P_G + 2\frac{\kappa_3}{\lambda'\lambda}\mu_\alpha\Psi + \frac{\kappa_4}{(\lambda'\lambda)^2}\Psi_*\right\}\Lambda. \tag{3.50}$$

A_0 is obtained from the following.

$$E\frac{\partial^2 C_i}{\partial\beta\partial\delta'} = 2[S_X'(P_G\otimes\Sigma_{WW})S_X,\ S_X'(P_G\otimes\Sigma_{WW})S_Z,\ S_X'(P_G\otimes\Sigma_{W\alpha})\Lambda] \tag{3.51}$$

$$E\frac{\partial^2 C_i}{\partial\gamma\partial\delta'} = 2[S_Z'(P_G\otimes\Sigma_{WW})S_X,\ S_Z'(P_G\otimes\Sigma_{WW})S_Z,\ S_Z'(P_G\otimes\Sigma_{W\alpha})\Lambda] \tag{3.52}$$

$$E\frac{\partial^2 C_i}{\partial\theta\partial\delta'} = 2[\Lambda'(P_G\otimes\Sigma_{W\alpha}')S_X,\ \Lambda'(P_G\otimes\Sigma_{W\alpha}')S_Z,\ \sigma_\alpha^2\Lambda' P_G\Lambda]. \tag{3.53}$$

Bibliography

Ahn, S. C., Y. H. Lee, and P. Schmidt (2001) 'GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects.' *Journal of Econometrics* 101, 219–255.

Ahn, S. C. and P. Schmidt (1995) 'A Separability Result for GMM Estimation, with Applications to GLS Prediction and Conditional Moment Tests.' *Econometric Reviews* 14, 19–34.

Battese, G. E., and T. J. Coelli (1992) 'Frontier Production Functions, Technical Efficiency and Panel Data: with Application to Paddy Farms in India.' *Journal of Productivity Analysis* 3, 153–170.

Bierens, H. J. (1994) *Topics in Advanced Econometrics*, Cambridge University Press.

Billingsley, P. (1968) *Convergence of Probability Measures*, Wiley.

Bound, J., D. A. Jaeger, and R. M. Baker (1995) 'Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogeneous Explanatory Variable is Weak (in Applications and Case Studies).' *Journal of the American Statistical Association* 90, 443–450.

Chamberlain, G. C. (1980) 'Analysis of Covariance with Qualitative Data.' *Review of Economic Studies* 47, 225–238.

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

Chamberlain, G. C. (1992) 'Efficiency Bounds for Semiparametric Regression.' *Econometrica* 60, 567–596.

Davidson, J. (1994) *Stochastic Limit Theory*, Oxford University Press.

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

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988) 'Estimating Vector Autoregressions with Panel Data.' *Econometrica* 56, 1371–1395.

James, A. T. (1964) 'Distributions of Matrix Variates and Latent Roots Derived from Normal Samples.' *Annals of Mathematical Statistics* 35, 475–501.

Johnson, N. L., and S. Kotz (1972) *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley & Sons.

Kiefer, N. M. (1980) 'A Time Series–Cross Sectional Model with Fixed Effects with an Intertemporal Factor Structure.' Unpublished Manuscript, Cornell University.

Kim, J., and D. Pollard (1990) 'Cube Root Asymptotics.' *The Annals of Statistics* 18, 191–219.

Kumbhakar, S. C. (1990) 'Production Frontiers, Panel Data and Time-varying Technical Inefficiency.' *Journal of Econometrics* 46, 201–212.

Lee, Y. H. (1991) 'Panel Data Models with Multiplicative Individual and Time Effects: Applications to Compensation and Frontier Production Functions.' Unpublished Ph.D. Dissertation, Department of Economics, Michigan State University.

Lee, Y. H., and P. Schmidt (1993) 'A Production Frontier Model with Flexible Temporal Variation in Technical Inefficiency.' In *The Measurement of Productive Efficiency: Techniques and Applications,* edited by H. Fried, C. A. K. Lovell, and S. Schmidt, Oxford University Press.

Linton, O. (1999) 'Problem.' *Econometric Theory* 15, 151.

Magnus, J. R., and H. Neudecker (1988) *Matrix Differentation Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons.

Manski, C. F. (1983) 'Closest Empirical Distribution Estimation.' *Econometrica* 51, 305–319.

Newey, W. K. (1988) 'Asymptotic Equivalence of Closest Moments and GMM Estimators.' *Econometric Theory* 4, 336–340.

Phillips, P. C. B. (1980) 'The Exact Distribution of Instrumental Variable Estimators in an Equation Containing $n + 1$ Endogenous Variables.' *Econometrica* 48, 861–878.

Staiger, D., and J. H. Stock (1997) 'Instrumental Variables Regression with Weak Instruments.' *Econometrica* 65, 557–586.

Theil, H. (1971) *Principles of Econometrics*, Wiley.