UCLA, Department of Economics, Patrik Guggenberger
Economics 203C (Spring, 2009), April 23, 2009

Problem Set IV (due Wed 5/6/09)

1.Take the linear iid IV model

$$y_i = x_i'\beta + U_i,\ E(z_iU_i) = 0,\ \Omega := Ez_iz_i'U_i^2 > 0,$$

where x_i and z_i are d- and k-dimensional vectors, respectively, $k > d$, and consider the IV estimator $\widehat{\beta}$ of β that uses $\widehat{\Omega}^{-1}$ as weighting matrix, where $\widehat{\Omega} \to_p \Omega$. The J statistic of overidentifying restrictions is given by $J_n = n\overline{g}_n(\widehat{\beta})'\widehat{\Omega}^{-1}\overline{g}_n(\widehat{\beta})$, where $g_i(\beta) := (y_i - x_i'\beta)z_i$, $\widehat{U}_i := y_i - x_i'\widehat{\beta}$, $\widehat{\Omega} := \sum_{i=1}^n z_iz_i'\widehat{U}_i^2/n$, and $\overline{g}_n(\beta) := \sum_{i=1}^n g_i(\beta)/n$. For an invertible matrix C defined below, let

$$D_n := I_k - C'(Z'X/n)(\frac{1}{n}X'Z\widehat{\Omega}^{-1}Z'X/n)^{-1}(\frac{1}{n}X'Z\widehat{\Omega}^{-1}C'^{-1}) \text{ and } R := C'Ez_ix_i',$$

where X and Z are the matrices that have rows given by the x_i and z_i, respectively. Show that $J_n \to_d \chi^2_{k-d}$ as $n \to \infty$ by demonstrating each of the following:

(i) There is a matrix C, such that $\Omega^{-1} = CC'$ and $\Omega = C'^{-1}C^{-1}$,
(ii) $J_n = n(C'\overline{g}_n(\widehat{\beta}))'(C'\widehat{\Omega}C)^{-1}C'\overline{g}_n(\widehat{\beta})$,
(iii) $C'\overline{g}_n(\widehat{\beta}) = D_nC'\overline{g}_n(\beta_0)$,
(iv) $D_n \to_p I_k - R(R'R)^{-1}R'$,
(v) $n^{1/2}C'\overline{g}_n(\beta_0) \to_d N \sim N(0, I_k)$,
(vi) $J_n \to_d N'(I_k - R(R'R)^{-1}R')N$,
(vii) $N'(I_k - R(R'R)^{-1}R')N \sim \chi^2_{k-d}$. Hint: $I_k - R(R'R)^{-1}R'$ is a projection matrix.

2. (i) Use the notation of PS II.4. Investigate the size and power properties of the J-test of the null hypothesis $Ez_i\varepsilon_i = 0$ in a Monte Carlo study where $S_n = (\sum_{i=1}^n z_iz_i'/n)^{-1}$ is used as the weighting matrix. The reduced form is $x_i = z_i'\pi + u_i$ for $\pi = (\eta, ..., \eta)'$. The vectors $(z_i, \varepsilon_i, u_i) \in \mathbb{R}^{K+2}$ are *iid*, distributed as $N(0, \Delta)$, where the $(K+2)$-dimensional square matrix Δ is given by

$$\Delta = \begin{pmatrix} I_K & c & 0 \\ c' & 1 & \rho \\ 0' & \rho & 1 \end{pmatrix},$$

where I_K is the K-dimensional identity matrix, c is a K-vector, and ρ a scalar. Assume the true β equals 0. Generate $R = 1000$ data samples for each of the following parameter combinations: $n = 100$, $K = (2, 10)$, $\rho = .3$, $\eta = (.05, 1)$ and CASE 1: $c = (\delta, ..., \delta)$ or CASE 2: $c = (\delta, 0, ..., 0)$, for $\delta = ... - .04, -.02, 0, .02, .04, .06...$(if $|\delta|$ is too big, Δ will eventually be indefinite). For each parameter combination, calculate the value of the J-statistic and calculate the percentage of the R cases where it exceeds the asymptotic critical value of a test with nominal level 5%. Report the rejection probabilities and interpret the results. How do K, ρ, and δ influence the results? What is the interpretation of c and π?

(ii) Assume the J test from (i) is used as a pretest in a two-stage testing procedure. In the second stage a t-test is used to test $H_0 : \beta = 0$ versus $H_0 : \beta \neq 0$ where all instruments are used if the pretest did not reject. If the pretest rejected, nothing is done in the second stage. Assume the pretest and second stage nominal sizes are both 5%. For Cases 1 and 2 in (i) with $\delta = .01$ and .02 (and choices of K, ρ, η as before) simulate the null rejection probability in the second stage, conditional on the pretest not rejecting.

3. The data "CARD.DAT" is taken from David Card "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" in *Aspects of Labour Market Behavior* (1995). There are 2215 observations with 19 variables. The attached sheet describes the variables. We want to estimate a wage equation

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 Exper^2 + \beta_4 South + \beta_5 Black + e,$$

where $Educ$ =Education(Years), $Exper$ =Experience(Years), and $South$ and $Black$ are regional and racial dummy variables.

(i) Estimate the model by OLS. Report estimates and standard errors.

(ii) Give reasons why $Educ$ may be endogenous. Treat $Educ$ as endogenous, and the remaining variables as exogenous. Estimate the model by two-stage least squares (2SLS), using the instrument $near4$, a dummy variable indicating that the observation lives near a four year college. Report estimates and standard errors and discuss the assumption that $near4$ is a valid instrument.

(iii) Re-estimate by 2SLS (report estimates and standard errors) adding three additional instruments: $near2$ (a dummy indicating that the observation lives near a 2-year college), $fatheduc$ (the education, in years, of the father) and $motheduc$ (the education, in years, of the mother). Again, discuss the assumption that these variables are valid instruments.

(iv) Re-estimate the model by efficient GMM. I suggest that you use the 2SLS estimates as the first-step to get the weight matrix, and then calculate the GMM estimator from this weight matrix (see Hayashi, p.213). Report the estimates and standard errors.

(v) Calculate the J statistic of overidentification.

(vi) Discuss your findings. Also, how reasonable is the assumption that $Exper$ is exogenous?

4. "OLS is BLUE": In the model $y = X\beta + \varepsilon$ with $X \in R^{n \times k}$ nonstochastic and full column rank and $E[\varepsilon] = 0$, $E[\varepsilon\varepsilon'] = \sigma^2 I_n$ (for some unknown positive number σ^2) show that the OLS estimator $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$ is the minimum variance linear unbiased estimator. That is any other unbiased estimator $\widetilde{\beta}$ that is given as a linear combination of y has non-smaller covariance in the positive definite sense.