Consequences and detection of invalid exogeneity conditions

Niemczyk, J.

Chapter 5

Sequential test procedures for instrumental variable selection

5.1 Introduction

The method of moments estimation technique exploits a set of moment conditions for estimation of the underlying parameters of a model. In the GMM framework, Hansen's J test is used to check whether the moment restrictions utilized in the estimation are satisfied indeed. As we mentioned earlier, only overidentifying restrictions can be tested since in general sample moments will be satisfied automatically in the exactly identified case. In the previous Chapter on a linear model, we analyzed different versions of Hansen tests, together with Sargan tests and other statistics based on Generalized Empirical Likelihood. In this Chapter we will propose several sequential procedures for selecting moment conditions from a set of potential ones that will form the basis for the ultimate estimation. In a linear regression model for example, from a set of potential variables we want to choose a subset that is uncorrelated with the error term.

Andrews (1999) proposed several search procedures for choosing correct moment conditions. For example, he showed that, by running the J test on all the possible combinations of moment restrictions and choosing the one with the maximum number of them not rejected by the J test, one can consistently determine which moment restrictions are satisfied indeed, provided that at least $p + 1$ of them are valid (p is the number of

parameters to estimate).

Apart from being valid, it is also very important for the instrumental variables to be relevant, i.e. sufficiently correlated with the regressors. Including more valid instruments produces asymptotically more efficient estimators but can also increase finite sample bias when the instruments are only weakly correlated with the explanatory variables. Donald and Newey (2001) propose to search, among the set of valid instruments, for those subsets which improve the performance of a given IV estimator. This performance is measured by the minimum of the approximate mean squared error of an estimator which can be obtained using refined asymptotic theory.

To achieve a better quality of asymptotic approximation to the finite sample behavior of GMM estimators Hall and Peixe (2003) and Hall, Inoue, Jana, and Shin (2007) propose a method for selecting relevant instruments based on certain canonical correlations. Hall and Peixe (2003) produce simulation evidence that the methods of Andrews (1999) can select irrelevant instruments, and that, on the other hand, their methods for selecting relevant instruments can also choose invalid ones. This could happen because if an *invalid* instrument is highly correlated with an endogenous variable, then we would most probably select it, based on some relevance criteria that neglects validity of instruments. In the same way we may select an irrelevant instrument when we only consider validity criteria. Based on the available Monte Carlo (MC) evidence, the best strategy seems to be to join the forces of the two methods, using first the relevance method and then the validation method. Inoue (2006) also proposes a bootstrap *relevance* search procedure from a set of instruments known to be valid. In this Chapter we will examine the performance of validity tests when the set of instruments may contain both strong and weak instruments.

Kapetanios (2006) recognizes that due to nonstandard discrete minimization problems all the above methods are computationally unfeasible. That is because they require searching over almost all possible subsets of available instruments. If the number of instruments is large the number of such subsets simply becomes too large. The author proposes to use algorithms which provide a theoretically valid and computationally tractable solution to the minimization problem and do not require to search over all possible subsets.

In this Chapter we propose and examine three sequential procedures (indicated by A,

B and C) for finding valid instruments. The procedures are sequential in a sense that at a stage s they consider only the sets of variables that consists of the already accepted instruments from the previous stage $s-1$ plus one more available variable. Procedure A will apply the *incremental* version of an overidentifying restriction test instead of its *standard* one. Hall (2005) showed that this incremental version leads to 'local power' improvement. Following the proof in Andrews (1999), we show that the procedures put forward below are consistent too, that is to say, when the sample size goes to infinity, they will yield the set of all available valid instruments. The procedures are computationally much less expensive than those proposed in Andrews (1999). Only in the first stage, where they search for the initial 'smallest' set of the valid instruments, they could be assisted by the algorithm of Kapetanios (2006). However, this is not required here for our empirical example.

In a Monte Carlo study we compare the performance of our procedures with two proposed by Andrews (1999). In an empirical illustration we apply two of our procedures to the Angrist and Krueger (1991) data, where depending on the implementation, we have 30 or 180 overidentifying restrictions, hence making the other procedures computationally unfeasible. Bound, Jaeger, and Baker (1995) showed that the analysis of Angrist and Krueger (1991) is affected by the weak identification problem. They also suggest that the 'instruments' used in Angrist and Krueger (1991) could also be invalid due to some correlation of the quarter of birth dummies with some omitted factors which can affect earnings.

In that empirical example, the Hansen or the Sargan tests on the full set of instruments do not reject the overidentifying restrictions. However, our sequential analysis applying incremental versions of those tests identifies moment restrictions that are not valid. That suggests that indeed the entire set of instruments (due to its particular structure generated by quarter of birth dummies) could be invalid, but because it is also weak we do not have sufficient power to reject it.

In the last illustration we will utilize the findings from the Chapters 2 and 3 for the Angrist and Krueger data in order to produce an alternative inference based on making varying assumptions on the degree of simultaneity, and, when external instruments have

been used, on their degree of invalidity. We will demonstrate that this unfeasible inference allows a useful sensitivity analysis. From this perspective, we will see that for these data inference based on the unfeasible bias corrected OLS is more attractive than the unfeasible bias corrected IV inference.

In Section 5.2, we start with technical considerations relating to the model and we state some results on the procedures to be operationalized. In Section 5.3 we describe three sequential procedures and sketch the proof of their consistency. Sections 5.4-5.6 reports on the Monte Carlo simulation and on the empirical study. Section 5.7 concludes.

5.2 Notation, assumptions and the test statistic

For some stationary data vector $\mathcal{X}_i$, $i = 1, \ldots, n$, from $\mathcal{X} = [\mathcal{X}_1, \ldots, \mathcal{X}_n]'$, where n is the sample size, we denote a particular $l \times 1$ vector function of the data by $g_i(\theta) \equiv g(\mathcal{X}_i, \theta)$ and the corresponding sample moment function by $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$, where θ is a $p \times 1$ vector of parameters and $l \geq p + 1$, θ does not necessarily characterize fully the distribution of $\mathcal{X}_i$. We assume that under true but unknown data generating process $\mathrm{E}g_i(\theta) = g(\theta)$ and that $\bar{g}_n(\theta) \xrightarrow{p} g(\theta)$ for every $\theta \in \Theta \subseteq \mathbb{R}^p$. We aim to estimate a unique θ_0 for which $g(\theta_0) = 0$, however this might not be possible if some moment conditions in the moment function are misspecified. That is $g(\theta) \neq 0$ for any $\theta \in \Theta$. We assume that there is a $(p+s) \times 1$ subvector $g^*(\theta)$ of $g(\theta)$ for which $g^*(\theta_0) = 0$, $s > 0$, and our aim is to design an iterative procedure that establishes this subvector $g^*(\theta_0)$, or rather its sample equivalent $\bar{g}^*_n(\hat{\theta})$.

In order to distinguish test statistics on different subsets of moment conditions, we need to generalize the notation used in the previous Chapter. Let the $l \times 1$ vector c_s be a selector vector of zeros and ones with the total number of ones equal to $p + s$, hence $l \geq p+s$. Vector c_s indicates which elements of $g(\theta)$ we are taking into account: by $g(c_s, \theta)$ we will denote the subvector of $g(\theta)$ obtained by deleting the entries of $g(\theta)$ corresponding to zeros in c_s. Define

$$\mathcal{C}_s \equiv \{c_s \in \mathfrak{R}^l : c_s = (c^1, \ldots, c^l)',\ c^i \in \{0, 1\},\ \text{for } i = 1, \ldots, l \text{ and } c'c = p + s\}$$

$$\equiv \{c_s \in \{0,1\}^l : c'c = p + s\}$$

to be the set of such *selection vectors*, $s = 1, \ldots, l - p$. Let

$$\mathcal{Z}^0 \equiv \{c_s \in \mathcal{C}_s, s = 1, ..., l - p : \exists! \theta_0 \in \Theta, \quad g(c_s, \theta_0) = 0\}$$

denote the set of all selection vectors that select valid moment conditions. In this definition we assume that there exists a unique parameter value that satisfies the moment conditions.

For notational ease we do not indicate the dependence of the vectors or sets on p. Below, s will indicate both the number of overidentifying restrictions and the current stage in a procedure. This should not cause any confusion since the procedures will be designed in such a way that those quantities will correspond.

For a given $c_s \in \mathcal{C}_s$, by $\mathcal{C}_{s+1}(c_s)$ we denote the set of vectors that can be obtained from the vector c_s by replacing one of its elements corresponding to 0 by 1, hence

$$\mathcal{C}_{s+1}(c_s) \equiv \{c_{s+1} \in \mathcal{C}_{s+1} : c'_{s+1} c_s = p + s\}.$$

Similarly to the definition of $g(c_s, \theta)$, let $\bar{g}_n(c_s, \theta)$ denote the $(p + s) \times 1$ sub-vector generated from $\bar{g}_n(\theta)$ by removing the elements corresponding to zero of the vector c_s. Other vectors and matrices when depending on a selection vector are also obtained by removing appropriate elements or columns.

Now, we can re-define the test statistic (4.6) for overidentifying restrictions. Let $W(\mathcal{X}, c_s) = O_p(1)$ be a $(p + s) \times (p + s)$ positive definite weighting matrix. Let

$$\tilde{\theta}_s \equiv \underset{\theta \in \Theta}{\operatorname{argmin}} \; \bar{g}_n(c_s, \theta)' W(\mathcal{X}, c_s) \bar{g}_n(c_s, \theta). \tag{5.1}$$

For a $c_s \in \mathcal{Z}^0$ we assume

$$\sqrt{n}(\bar{g}_n(c_s, \tilde{\theta}_s) - g(c_s, \theta_0)) \xrightarrow{d} N(0, \Omega(c_s, \theta_0)).$$

Then the asymptotically optimal choice for $W(\mathcal{X}, c_s)$ is $\hat{\Omega}^{-1}(c_s, \tilde{\theta}_s)$, where

$$\underset{n\to\infty}{\text{plim}}\hat{\Omega}(c_s, \tilde{\theta}_s) = \Omega(c_s, \theta_0),$$

and where $\tilde{\theta}_s$ is the initial consistent estimator of θ_0, (5.1), that resulted from using some preliminary $W(\mathcal{X}, c_s)$ (say I_{p+s}). For the reasons we mentioned in the previous Chapter, we will use

$$\hat{\Omega}(c_s, \tilde{\theta}_s) = \frac{1}{n}\sum_{i=1}^{n} g_i(c_s, \tilde{\theta}_s) g_i(c_s, \tilde{\theta}_s)' - \bar{g}_n(c_s, \tilde{\theta}_s)\bar{g}_n(c_s, \tilde{\theta}_s)' \tag{5.2}$$

instead of

$$\hat{\Omega}(c_s, \tilde{\theta}_s) = \frac{1}{n}\sum_{i=1}^{n} g_i(c_s, \tilde{\theta}_s) g_i(c_s, \tilde{\theta}_s)'.$$

For any c_s and using expression (5.2) we re-compute (5.1) to obtain the efficient 'two step' GMM estimator

$$\hat{\theta}_s \equiv \underset{\theta\in\Theta}{\text{argmin}}\ \bar{g}_n(c_s, \theta)'\hat{\Omega}^{-1}(c_s, \tilde{\theta}_s)\bar{g}_n(c_s, \theta) \tag{5.3}$$

and Hansen test statistic

$$J_n(c_s) = n\bar{g}_n(c_s, \hat{\theta}_s)'\hat{\Omega}^{-1}(c_s, \hat{\theta}_s)\bar{g}_n(c_s, \hat{\theta}_s). \tag{5.4}$$

For example for a linear model we have $\mathcal{X}_i = (y_i, x_i', z_i')$, where z_i is an $l \times 1$ vector of alleged instruments and x_i is a $p \times 1$ vector of regressors (x_i and z_i can share the same elements, z_i may contain invalid instruments), $i = 1, \ldots, n$, whereas $Z' = [z_1, \ldots, z_n]$ and $X' = [x_1, \ldots, x_n]$, we will have

$$\bar{g}_n(c_s, \hat{\theta}_s) = \frac{1}{n}\sum_{i=1}^{n} z_i(c_s)(y_i - x_i'\hat{\theta}_s) = \frac{1}{n}Z(c_s)'(y - X\hat{\theta}_s), \tag{5.5}$$

$$\begin{aligned}\hat{\Omega}(c_s, \tilde{\theta}) \;=\; & \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\tilde{\theta}_s)^2 z_i(c_s) z_i(c_s)' \\ & -\frac{1}{n^2}Z(c_s)'(y - X\tilde{\theta}_s)(y - X\tilde{\theta}_s)'Z(c_s)\end{aligned} \tag{5.6}$$

or for the Sargan test case

$$\begin{aligned}\dot{\Omega}(c_s,\tilde{\theta}) &= \frac{(y-X\tilde{\theta}_s)'(y-X\tilde{\theta}_s)}{n}\frac{1}{n}Z(c_s)'Z(c_s) \\ &\quad -\frac{1}{n^2}Z(c_s)'(y-X\tilde{\theta}_s)(y-X\tilde{\theta}_s)'Z(c_s).\end{aligned} \tag{5.7}$$

Below, we collect the assumptions we made above and state some further ones on the selection vectors that will lead to consistency of the sequential procedures to be developed.

Assumption 5.1 *(a) θ_0 is the same for all the selection vectors $c_i \in \mathcal{Z}^0$.*

(b) There exists at least one selection vector $c_s \in \mathcal{Z}^0$, for some $s = 1,\ldots,l-p$, (that is $\mathcal{Z}^0 \neq \emptyset$)

Assumption 5.1(a) is required for the identification reasons and says that any set of variables that satisfies the moment conditions will uniquely identify the parameter θ_0. Assumption 5.1(b) states that the model needs to be overidentified, hence, requires us to have more valid moment conditions than parameters to estimate. We need this because the J_n statistic can only test overidentifying restrictions. In case $l = p$ we obtain $J_n = 0$.

Self-evidently, for any $c_i \in \mathcal{Z}^0$ also $c_{i-1} \in \mathcal{Z}^0$ for all c_{i-1} such that $c_{i-1}'c_i = p + i - 1$. That is, for some set of *valid* instruments, all it's subsets are also valid sets of instruments. See the Remark 5.2 below.

Below we adopt further assumptions from Hall (2005) for the regularity of $J_n(c_s)$. More general assumptions can be found in Andrews (1999) or Andrews (1997).

Assumption 5.2 *(a) $J_n(c_s) \xrightarrow{d} \chi^2(s)$, for any $c_s \in \mathcal{Z}^0$,*

(b) $J_n(c_{s+1}) - J_n(c_s) \xrightarrow{d} \chi^2(1)$, for $c_{s+1} \in \mathcal{C}_{s+1}(c_s)$ and both c_s and c_{s+1} belong to $\mathcal{Z}^0$, and

(c) $\underset{n\to\infty}{\text{plim}}\frac{J_n(c_s)}{n} = \mu(c_s,\theta^*) > 0$, for any $c_s \notin \mathcal{Z}^0$.

Assumption 5.2 essentially allows us to distinguish (asymptotically) between valid moment conditions for which $J_n(c_s) = O_p(1)$ and invalid ones for which $J_n(c_s) = O_p(n)$. Let $\gamma_{n,s}$ be a $1-\alpha_n$ quantile of the $\chi^2(s)$ distribution. It is used as the critical value for

the $J_n(c_s)$ test or its incremental version. For the consistency result it is assumed that $\gamma_{n,s} \to +\infty$ and $\gamma_{n,s} = o(n)$. This will hold if the significance level α_n satisfies $\alpha_n \to 0$ and $\log \alpha_n = o(n)$, for example $\alpha_n = K \exp\{-\sqrt{n}\}$, for some $K > 0$. As $J_n(c_s) = O_p(1)$ for $c_s \in \mathcal{Z}^0$, the statistic will not exceed $\gamma_{n,s}$ (when n goes to infinity), but for $c_s \notin \mathcal{Z}^0$ it will exceed it as $J_n(c_s) = O_p(n)$. Hence, critical value $\gamma_{n,s}$ will separate valid instruments from invalid ones.

5.3 Selection of valid instruments

In this section we propose three sequential procedures for detecting all the valid moment conditions from a set of $l > p$ potential ones. All three can be shortly described as follows: in stage $s = 1$, compute the J test on all the possible sets of $p + 1$ moment conditions and then consider only those sets that were not rejected by the test. Next, if $l > p + 1$, compute all possible J tests based on one extra moment restriction in addition to the $p+1$ ones accepted earlier. Continue until a stage s where either $s = l - p$ or the test rejects all such possible combinations of $p + s$ moment conditions. Finally, make a selection out of the set of all combinations of $p + s$ moment restrictions that were not rejected. The differences between the three alternative procedures are described in the next subsection.

In the Monte Carlo simulation we will also study Andrews' procedure of BIC type. This procedure finds the optimal selection vector, c^*_{ABIC}, by minimizing over all $c_s \in \mathcal{C}_s$, $s = 1, \ldots, l-p$, the criterion function $J_n(c_s) - s \log(n)$. We will call this procedure $ABIC$. Another Andrews' procedure that we will examine is the upward searching procedure, called AU here. It finds the optimal selection vector c^*_{AU} by computing in the first step $s = 1$ the $\binom{l}{p+1}$ possible $J_n(c_1)$ statistics. If the minimum of them does not exceed a critical value $\gamma_{n,1}$ the procedure continues to the next step, $s = 2, 3, \ldots$, and computes the $\binom{l}{p+s}$ possible $J_n(c_s)$ statistics until a stage is reached where all the combinations of instruments are rejected. Then the procedure accepts the combination c^*_{AU} from the previous stage that produces the smallest value of the J statistic. The procedure accepts all the moment conditions as the solution if in the final stage none of them were rejected.

Next, we introduce the three procedures that we propose.

5.3.1 Procedure A

The first procedure exploits the incremental version of the J statistic.

$s = 1$ Find the set of moment conditions, indicated by c_1^*, that produces the smallest $J_n(c_1^*)$ statistic, that is, $c_1^* = \underset{c_1 \in \mathcal{C}_1}{\operatorname{argmin}} \{J_n(c_1)\}$. If $l = p + 1$ then accept the moment conditions selected by c_1^* and stop. If $l > p + 1$ and $J_n(c_1^*)$ does not exceed the critical value $\gamma_{n,1}$ then go to the next stage ($s = 2$), otherwise accept the moment conditions selected by c_1^* and stop.

$s > 1$ For the set of accepted moment conditions selected from the stage $s - 1$ by c_{s-1}^*, compute all the $l - p - s + 1$ possible J statistics that use the accepted moments plus one of the remaining. That is, for all $c_s \in \mathcal{C}_s(c_{s-1}^*)$ compute the $J_n(c_s)$ statistics. Then find $J_n(c_s^*)$, where $c_s^* \equiv \underset{c_s \in \mathcal{C}_s(c_{s-1}^*)}{\operatorname{argmin}} \{J_n(c_s)\}$. If $J_n(c_s^*) - J_n(c_{s-1}^*) < \gamma_{n,1}$ then, if $s = l - p$ (e.g. there are no more moment restrictions to search from), accept c_s^* as the solution, or if $s < l - p$, take $g(c_s^*, \theta_0)$ to the next stage and repeat this point. If $J_n(c_s^*) - J_n(c_{s-1}^*) \geq \gamma_{n,1}$, accept the previous result c_{s-1}^* and stop the search.

At the first stage of this procedure, we allow it when $l \geq p + 1$ to accept for the final solution a set of moment restrictions that was actually rejected by the J test. Since we assume that there are at least $p + 1$ valid restrictions, we are not allowing to commit a type I error (we do anyhow accept $p + 1$ orthogonality conditions), but may commit a type II error (if there are invalid moment restrictions in the accepted set).

Further stages ($s = 2, \ldots$) of procedure A can be summarized as follows. To the accepted set of moment conditions, add one extra and check if the extended set of moment restrictions is accepted by the incremental J test. If no extra restriction can produce a small enough value of the incremental J test then we take the moment restrictions that were already accepted set as the solution. If there are such restrictions, then to the set of already accepted restrictions add the one that produces the smallest value of the incremental J test, and continue the investigation.

The next procedure is very similar to the one above. Instead of considering the incremental version of the J test when $s > 1$, it uses the standard one. That is, procedure B applies the $J_n(c_s^*)$ statistic instead of $J_n(c_s^*) - J_n(c_{s-1}^*)$.

5.3.2 Procedure B

Do exactly as in procedure A, but instead of the incremental statistic $J_n(c_s^*) - J_n(c_{s-1}^*)$ and the critical value $\gamma_{n,1}$ use $J_n(c_s^*)$ with the critical value $\gamma_{n,s}$.

The two alternative approaches can give different results. How good they are depends on the actual properties of the instruments in question and the size and power properties of the two tests in finite sample.

5.3.3 Procedure C

This procedure is an extension of procedure B. The difference is that it will carry to the next stage all the sets of moment conditions that were accepted by the J test, instead of just one.

$s = 1$ for all $\begin{pmatrix} l \\ p+1 \end{pmatrix}$ selector vectors c_1 compute the $J_n(c_1)$ statistic. If $l > p+1$, then determine the set $\mathbb{C}_1^* \equiv \{c_1 \in \mathcal{C}_1 : J_n(c_1) < \gamma_{n,1}\}$ of *all* the sets of moment conditions for which $J_n(c_1) < \gamma_{n,1}$. If $\mathbb{C}_1^* \neq \emptyset$ then move to $s = 2$. Otherwise, if $l = p+1$ or $\mathbb{C}_1^* = \emptyset$, accept the moment conditions selected by $c_1^* = \underset{c_1 \in \mathcal{C}_1}{\operatorname{argmin}} \{J_n(c_1)\}$, and stop.

$s > 1$ From the sets of instruments accepted at the previous stage $s-1$, $c_{s-1} \in \mathbb{C}_{s-1}^*$, construct the set $\mathbb{C}_s$ of such selector vectors that can be obtained from any $c_{s-1} \in \mathbb{C}_{s-1}^*$ by changing one of the zero elements of c_{s-1} into 1. Next, compute all the $J_n(c_s)$ for $c_s \in \mathbb{C}_s$ and determine $\mathbb{C}_s^* \equiv \{c \in \mathbb{C}_s : J_n(c_s) < \gamma_{n,s}\}$. If $\mathbb{C}_s^*$ is not empty then, if $p + s = l$ accept all the moment conditions, otherwise repeat this point increasing s. If the set $\mathbb{C}_s^*$ is empty then stop and accept the moment conditions from the previous stage $s-1$ selected by c_{s-1} as the solution, where $c_{s-1}^* \in \mathbb{C}_{s-1}^*$ corresponds to the smallest $J_n(c_{s-1})$ in that stage.

Remark 5.2 *The possible advantage of the procedure C over B is that it considers more combinations of the instruments which could possibly lead to improvement of the final result. The disadvantage is that it will be computationally more involved. In some cases it might run the same number of computations as the upward searching procedure of Andrews (for example when all the instruments are valid).*

Remark 5.3 *The procedures A and B are computationally expensive only at the first stage, where they need to run* $\binom{l}{p+1}$ *operations whereas after that they run at most* $\frac{(l-p)(l-p-1)}{2} = \sum_{i=1}^{l-p-1} i$. *Whenever* $\binom{l}{p+1}$ *is too big we could resort to an algorithm proposed by Kapetanios (2006) in order to find a* c_1 *that is not rejected by the J test. Even when it is not the one that minimizes the* $J_n(c_s)$ *it should still lead to an asymptotically correct result, by Assumption 5.1.*

The approximate search time of some of the procedures, assuming it can search through 1 million different selection vectors per second, is the following

(l,p)	$ABIC$	A,B
$(20,5)$	1.04 sec	0.015 sec
$(30,5)$	18 min	0.1425 sec
$(40,5)$	13 days	0.6580 sec
$(40,20)$	7 days	38 hours
$(50,20)$	33 years	1.5 years
$(50,15)$	36 years	26 days
$(50,10)$	36 years	3 hours
$(150,9)$	$> 10^{30}$ years	3 years
$(150,8)$	$> 10^{30}$ years	60 days
$(150,6)$	$> 10^{30}$ years	4 hours
$(150,5)$	$> 10^{30}$ years	10 min

In the 'worse' case, when it reaches the maximum possible stage, procedure AU will be as computationally involved as procedure $ABIC$. Procedure C should be less computationally involved than procedure AU, but in principle could be as much time consuming as AU, for example when all the moment conditions are valid.

Theorem 5.1 *Under Assumption 5.1 and 5.2 procedures A, B and C are consistent.*

Proof. Following Andrews (1999), by definition of the set $\mathcal{Z}^0$, for any $c_s \in \mathcal{Z}^0$ we have $P(J_n(c_s) < \gamma_{n,s}) \to 1$, because $J_n(c_s) = O_p(1)$ by Assumption 5.2(a). For $c_s \notin \mathcal{Z}^0$ we have

$\underset{n\to\infty}{\text{plim}} J_n(c_s)/\gamma_{n,s} = \infty$, because $\underset{n\to\infty}{\text{plim}} J_n(c_s)/n > 0$ by Assumption 5.2(c). Thus the only selection vectors that do not exceed asymptotically the critical value are in the set $\mathcal{Z}^0$. Let $c_s^* \in \mathcal{Z}^0$ be the (unique) selection vector which selects all the $(p+s)$ valid instruments. By Assumption 5.1, there exists $c_1 \in \mathcal{Z}^0$ which will not be rejected, $P(J_n(c_1) < \gamma_{n,1}) \to 1$, and so consistency of the first step of the procedures follows. Next, at each stage $1 < i \leq s$ there are some $c_i, c_{i-1} \in \mathcal{Z}^0$ ($c_{i-1}' c_i = p + i - 1$) for which $P(J_n(c_i) - J_n(c_{i-1}) < \gamma_{n,1}) \to 1$, and that assures consistent progress from one step to another. When we reach the stage $s+1$, where there is no $c_{s+1} \in \mathcal{Z}^0$, we will have $\underset{n\to\infty}{\text{plim}}(J_n(c_{s+1}) - J_n(c_s))/\gamma_{n,1} = \infty$ for any c_{s+1}, but $P(J_n(c_s) < \gamma_{n,1}) \to 1$, which can only happen for $c_s = c_s^*$. If $(c_s^*)' c_s^* = l$, that is, if all the instruments are valid, then for any c_i, c_{i-1} $P(J_n(c_i) - J_n(c_{i-1}) < \gamma_{n,1}) \to 1$, so as a result we will select all the valid instruments as n goes to infinity.

This proves the consistency of Procedure A. Consistency of the other procedures is proven similarly. ■

5.4 A Monte Carlo Study

Here, we compare the methods A, B, C and the Andrews' methods $ABIC$ and AU. We examine them in a simple static linear model with one regressor and $l = 4$ instruments, $z_t' = (z_{t1}, \ldots, z_{tl})$. For a $(n \times l)$ matrix Z we denote by Z_i its i'th column and its t'th row we denote by z_t'.

We generate the data according to the following linear model

$$y_t = x_t\theta + u_t \tag{5.8}$$

$$x_t = \bar{z}_t'\pi + v_t, \tag{5.9}$$

where y_t, x_t are scalar endogenous variables, $t = 1, \ldots, n$, π is a $(l \times 1)$ vector of reduced form parameters. The instruments $\bar{Z} = [\bar{z}_1, \ldots, \bar{z}_T]'$ are predetermined, $\mathrm{E}(u_t|\bar{z}_t) = 0$ with $\mathrm{Var}(\bar{z}_t) = I_l$. We take

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim \mathrm{IIN}\big(0, \Sigma\big), \quad \Sigma = \begin{pmatrix} \sigma_u^2 & \rho_{uv}\sigma_u\sigma_v \\ \rho_{uv}\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}, \tag{5.10}$$

and create one invalid instrument (say the first one, Z_1) by generating Z_1 according to

$$Z_1 = \sqrt{1 - \rho_{Z_1u}^2}\bar{Z}_1 + \rho_{Z_1u}u. \tag{5.11}$$

We have $\mathrm{Var}(z_{t1}) = 1$. For the normalization we take $\sigma_u^2 = 1$, $\theta = 1$. As in the previous chapter, we will choose values for π and σ_v^2 via population versions of the concentration parameter

$$\mu_p^2 = n\pi'\Sigma_{\bar{Z}'\bar{Z}}\pi/\sigma_v^2 = \frac{n}{\sigma_v^2}\sum_{i=1}^{l}\pi_i^2. \tag{5.12}$$

and the signal to noise ratio of (5.8)

$$\eta^2 = \frac{\mathrm{Var}(x_t\theta)}{\mathrm{Var}(u_t)} = \mathrm{Var}(x_t) = \sigma_x^2.$$

From (5.9) we then have

$$\sigma_x^2 = \pi' \Sigma_{\bar{Z}'\bar{Z}} \pi + \sigma_v^2 = \sum_{i=1}^{l} \pi_i^2 + \sigma_v^2, \tag{5.13}$$

and combining (5.12) with (5.13) we get

$$\sigma_v^2 = \frac{\sigma_x^2}{\mu_p^2/n + 1}. \tag{5.14}$$

Taking $\pi_i = \pi_0$ in (5.12) for $i = 1, \ldots, l$, we get $\mu_p^2 \sigma_v^2 = nl\pi_0^2$. Hence, for a given sample size n, μ_p^2, and η^2 we can calculate σ_v^2 from (5.14) and π_0 from $\pi_0^2 = \mu_p^2 \sigma_v^2/(nl)$.

Borrowing from Eryuruk, Hall, and Jana (2008), we will also try *declining coefficients* $\pi_i = k(l)(1 - \frac{i}{l+1})$ for $i = 1, 2, 3, 4$, and the constant $k(l)$ is such that (5.12) is satisfied.

We are going to utilize the Sargan test instead of Hansen since in our example we have conditional homoscedasticity. We will also use the residual bootstrap described in the previous chapter which worked best for the Sargan test.

For the significance level α_n instead of $\alpha_n = K \exp\{-\sqrt{n}\}$ suggested earlier, we take $\alpha_n = K/\ln n$ which is justified in Andrews (1997) using the law of the iterated logarithm. K is chosen in such a way that $\alpha_{50} = 0.05$, then for example $\alpha_{200} = 0.0369$, $\alpha_{500} = 0.0315$.

The graphs display, (in the first three 'columns' of each graph) acceptance frequencies of different sets of instruments. We group under INV: the sets of instruments that contain the invalid one, under AVL: the sets of all the valid instruments and under VAL: the sets of valid instruments, but not all of them. The right-hand graph presents [10%, 50%, 90%] quantiles of the estimators obtained from the sets of instruments resulting from the five different sequential procedures. The upper panels present results involving the constant coefficients and the lower descending coefficient cases.

In the simulation we fix $\rho_{xu} = 0.2$, we take $\mu_p^2/l = 20$ for the 'strong' instruments case and $\mu_p^2/l = 1$ for the 'weak' ones. We do not present bootstrap results since they did not give much different results from the the 'asymptotic' ones.

Figures 5.1 to 5.4 present different findings for different cases: $n = 50, 500$, strong or weak instruments, constant or descending coefficients. Note that for the descending coef-

ficients cases the first instrument (the only one which we make invalid) is the 'strongest' relative to the others. Figure 5.5 presents the $n = 1000$ and strong instruments case.

Looking at the acceptance frequencies, we notice, that the procedure A is 'almost' always between the $ABIC$ and B, C or AU (those three procedures, for small $n = 50$, give very similar results, whereas C and AU in all the cases give almost identical results). With respect to separating valid from invalid instruments the procedure $ABIC$ is always better than the rest.

For, n=50, procedures B, C and AU produce very similar 'intervals' and are the narrowest whereas estimators based on $ABIC$ have the widest 'confidence intervals'. Apart from that, the 'confidence intervals' are rather similar, except the large samples/strong instruments/descending coefficients case where the procedures A and B less frequently (about 80% for strong invalidity) to the rest of the procedures (above 90% for strong invalidity) capture the valid instruments. That is most probably due to the fact that there the invalid instrument gets the 'highest' weight in the generating scheme making it relatively most preferable by the procedures A and B which in turn heavily rely on the initial set of instruments they choose.

The consistency of the procedures is best seen comparing $n = 50$ and 500 for strong and constant coefficients. However, when the invalidity of the instrument is small ($\rho_{Z_1 u} = 0.1$), sample size $n = 1000$ is not yet large enough to notice the consistency (only the procedures A and $ABIC$ show some). For the weak instrument cases, the consistency is not apparent in the cases examined.

For the weak instrument cases the acceptance frequencies of the invalid instruments are quite bad for all the procedures. $ABIC$ is there still the most preferable and the procedure A is the second best.

However, the 'distribution' of the estimates resulting from the resulting procedures is optimistic. For small invalidity the resulting distribution of the estimates over the chosen instruments is almost centered around the true value. All of the 'confidence intervals' cover the true parameter. For the weak instruments, however, the median of the distribution diverges.

In the Appendix we give an intuition why the procedures AU and C give almost

identical results. We also discus there some potential traps during the moment selection search.

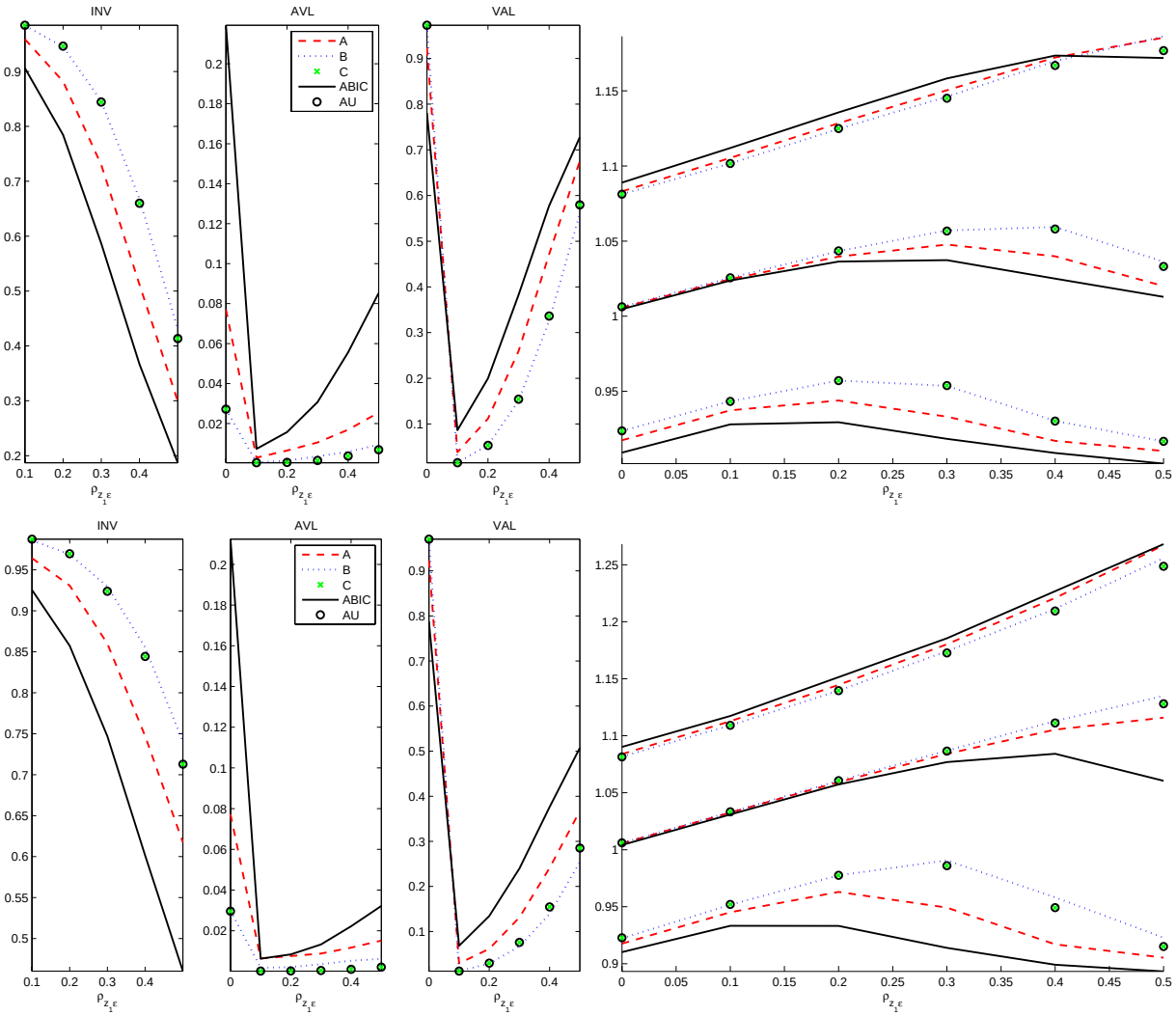

Figure 5.1: Acceptance frequencies for the strong instruments case at $n = 50$ for different sets of instruments by different procedures grouped by: INV - containing the invalid one; AVL - containing all the valid instruments; VAL - containing valid instruments. For the left hand panels the three upper diagrams present results involving the constant coefficients and the three lower diagrams for the descending coefficients. The right-hand graphs present [10%, 50%, 90%] quantiles of the estimators obtained from the accepted sets of instruments.

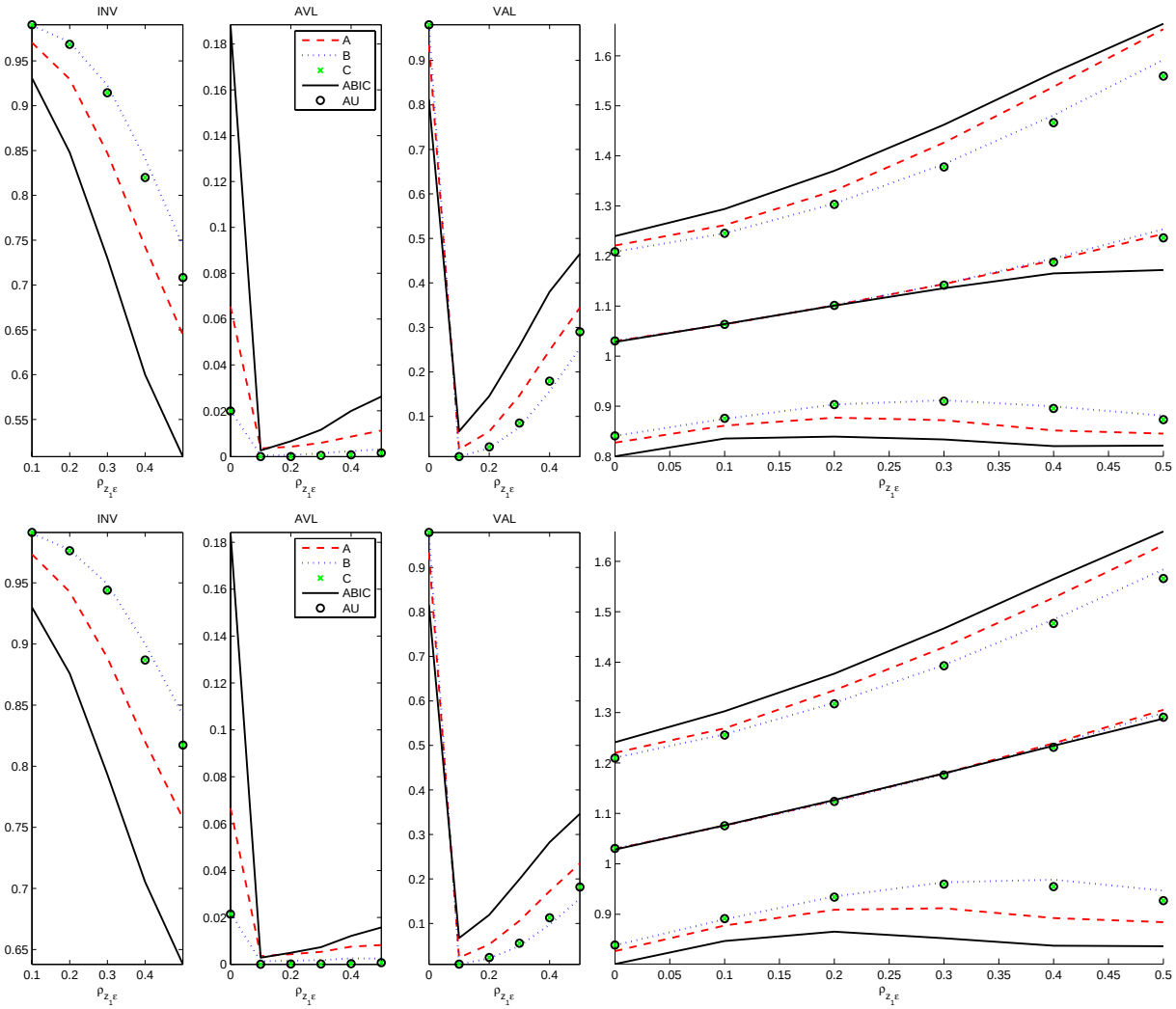

Figure 5.2: Same description as for Figure 5.1. $n = 50$; weak instruments case; upper (lower) panels - constant (descending) coefficients;

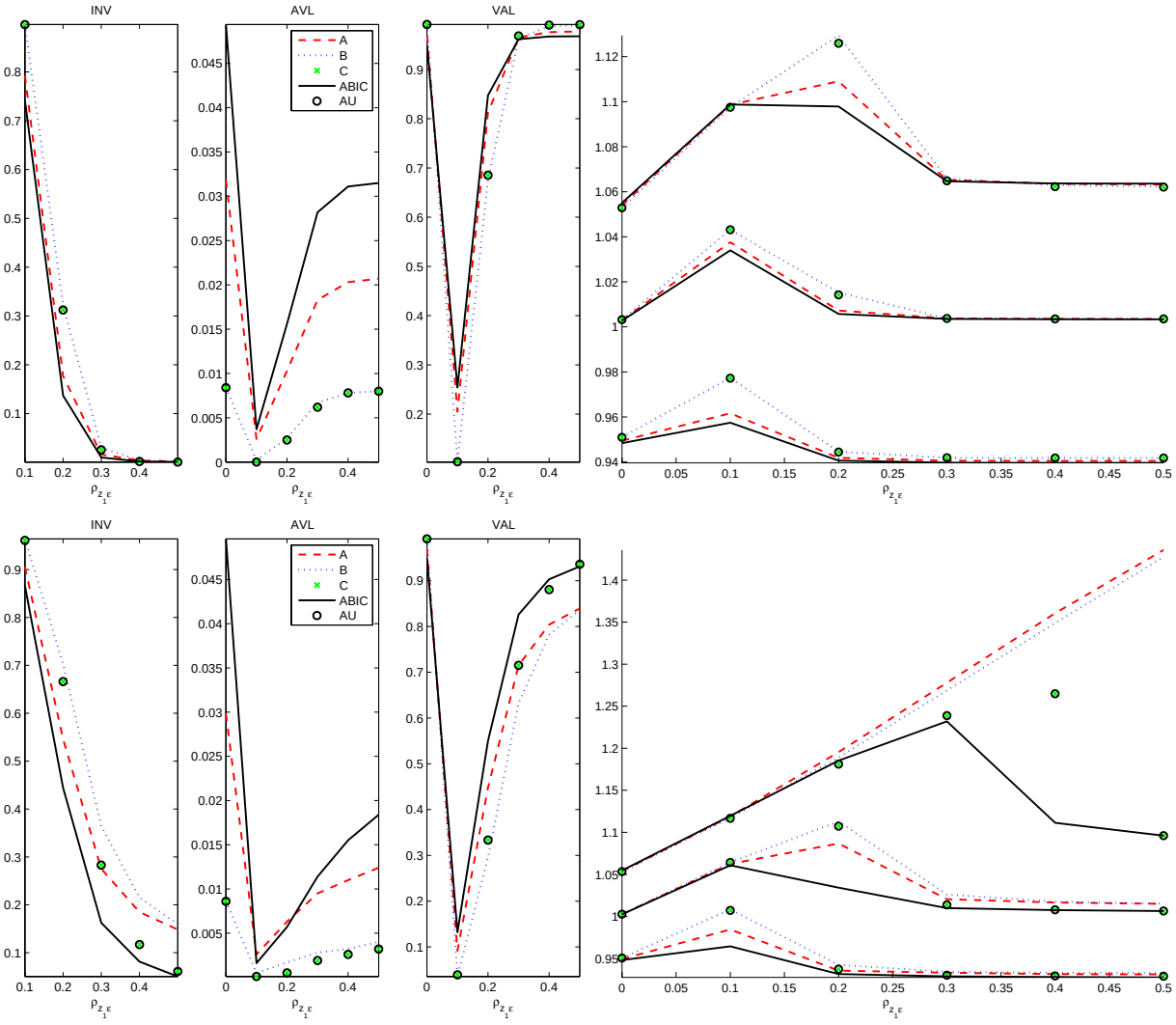

Figure 5.3: Same description as for Figure 5.1. $n = 500$, strong instruments case; upper (lower) panels - constant (descending) coefficients;

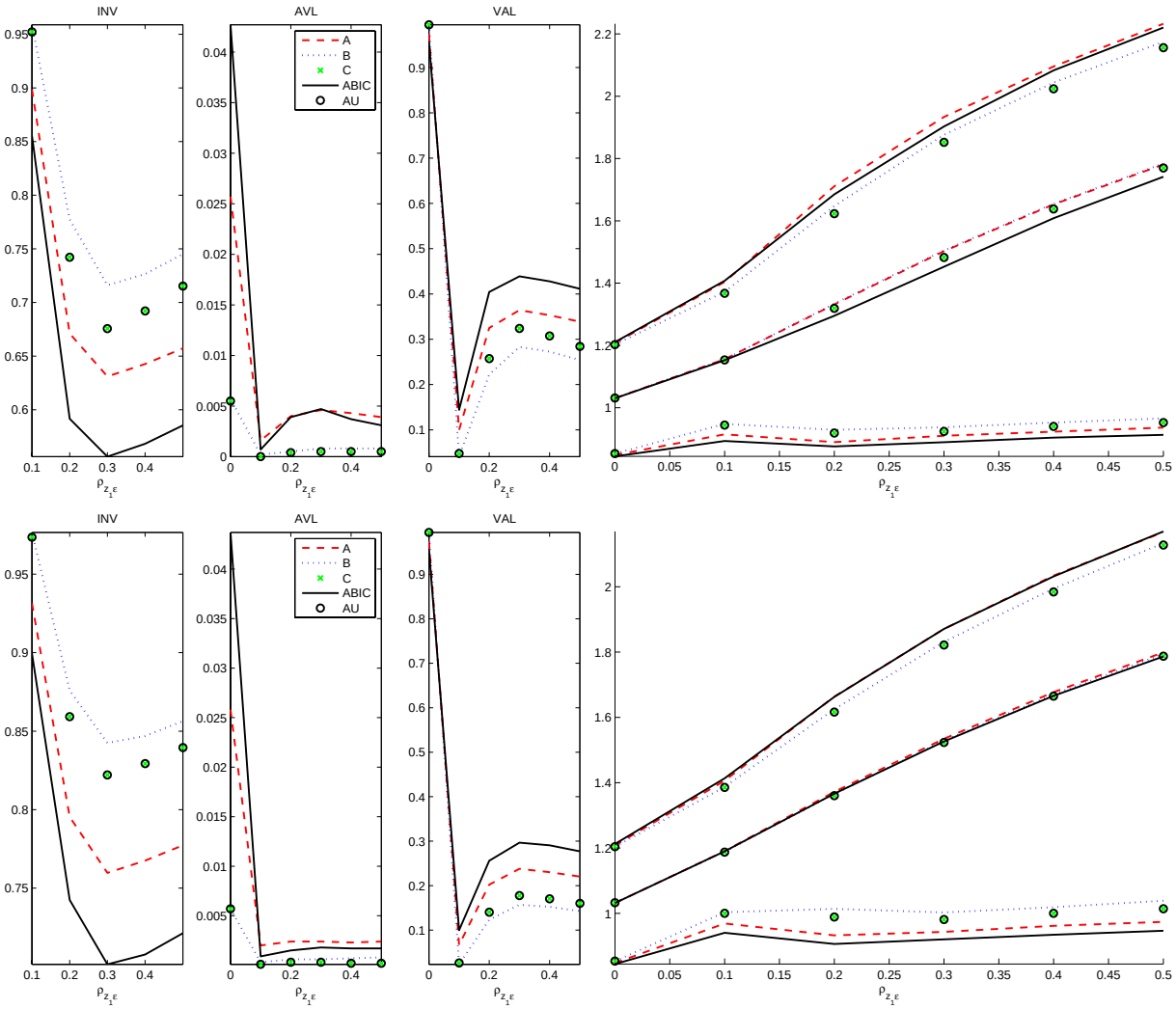

Figure 5.4: Same description as for Figure 5.1. $n = 500$, weak instruments case; upper (lower) panels - constant (descending) coefficients;

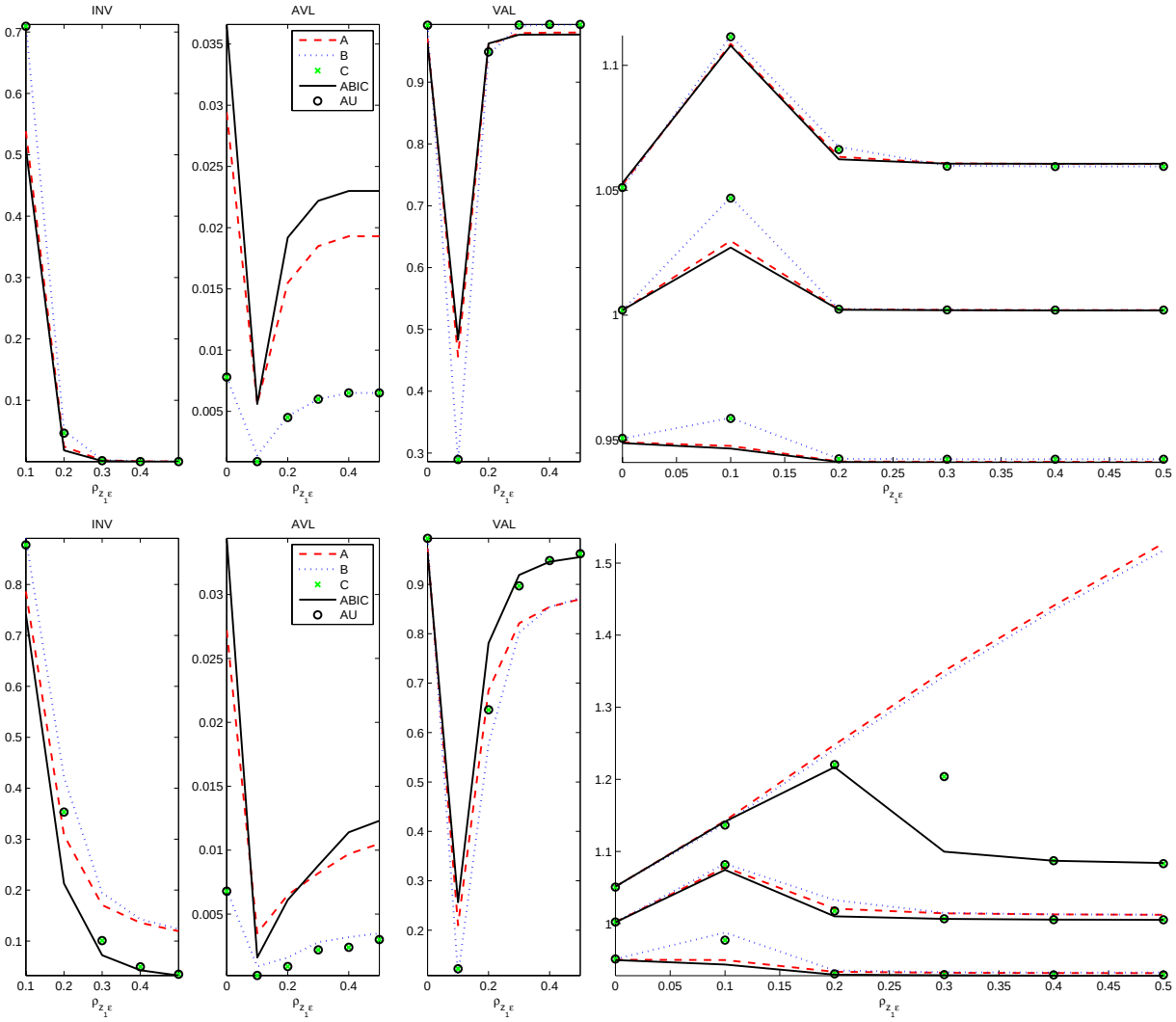

Figure 5.5: Same description as for Figure 5.1. $n = 1000$, strong instruments case; upper (lower) panels - constant (descending) coefficients;

5.5 An Empirical Example

We will re-analyze the data of Angrist and Krueger (1991) exploiting the various techniques developed in this and earlier chapters. Angrist and Krueger investigate the effect of years of education E on the logarithm of earnings, $\ln\{W\}$. They found that OLS and IV estimators provide remarkably similar estimates, and that the existing differences (however small) could be explained due to the (omitted variables or measurement error) downward bias of OLS. Bound, Jaeger, and Baker (1995)) found that the instruments used in the IV estimation are extremely weak and hence jeopardize the conclusions of Angrist and Krueger which were based on the 'classic' IV inference techniques.

We consider here the cohort of men born between 1930-1939. The sample size $n = 329509$. In the first implementations we use the ten year of birth dummy variables, Y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, 10)$, as additional regressors and in the second implementation also 'state' dummy variables, S_{il} $(l = 1, \ldots, 50)$. As potential instruments we use the interaction terms of the dummy variables with the quarter of birth (QOB) dummy variables, Q_{ik} $(k = 1, \ldots, 3)$. Angrist and Krueger (1991) argue that because men born at the beginning of a year are older than those born at the end of the same year, they are eligible to leave school earlier due to the school compulsory attendance law. This law creates a natural experiment in which the quarter of birth is correlated with the school attendance (but, as recognized later, only very weakly, see Bound, Jaeger, and Baker (1995)), whereas it is unlikely to be correlated with any unobserved omitted earnings determinants. For a more detailed discussion on the instrument validity of Q_{ik} see Angrist and Krueger (1991).

The first model (M_1) is

$$\ln\{W_i\} = X_i'\beta + \rho E_i + \sum_{j=1}^{10} Y_{ij}\xi_j + u_i \tag{5.15}$$

$$E_i = X_i'\pi + \sum_{j=1}^{10} Y_{ij}\delta_j + \sum_{j=1}^{10}\sum_{k=1}^{3} Y_{ij}Q_{ik}\theta_{jk} + v_i, \tag{5.16}$$

where X_i contains variables such as race dummies, a dummy for residence in SMSA ("standard metropolitan statistical area"), a marital status dummy and 8 region of residence dummies. Equation (5.16) is the reduced form for E_i. Its regressors are assumed

to be uncorrelated with u_i and v_i, whereas u_i and v_i may be correlated. Below we will apply our sequential procedures to test the validity of all the regressors in (5.16) as instruments for E_i in (5.15). We also consider implementations where β and π are zero (we include all 10 year of birth dummies hence the constant is excluded). For observation i and $j = 1, \ldots, 10$, Y_{ij} is the year dummy variable which, for birth year y, is equal to one if $j = y - 1929$ and zero otherwise. Q_{ik} is the quarter of birth dummy variable which, for quarter $k = 1, 2, 3$, is equal to one if individual i was born in quarter k and zero otherwise. We do not include Age and Age^2 variables in X_i because they were not found significant in the analysis of Angrist and Krueger (1991). They also lead to almost perfect collinearity because the Year dummy variables Y_{ij} are already included in the regression.

Let $Y_i' = [Y_{i1}, \ldots, Y_{i10}]$, $Q_i' = [Q_{i1}, Q_{i2}, Q_{i3}]$ then we can rewrite (5.15) and (5.16) as

$$\begin{aligned} \ln\{W_i\} &= \rho E_i + Y_i'\xi + u_i \qquad (5.17) \\ E_i &= Y_i'\delta + (Q_i' \otimes Y_i')\theta + v_i, \qquad (5.18) \end{aligned}$$

where $\theta' = [\theta_{1\cdot1}, \ldots, \theta_{1\cdot10}, \theta_{2\cdot1}, \ldots, \theta_{2\cdot10}, \theta_{3\cdot1}, \ldots, \theta_{3\cdot10}]$. If we include 'State' dummy variables $S_i' = [S_{i1}, \ldots, S_{i50}]$ in the model (we have 51 'states' in the data), we get the second implementation (M_2)

$$\begin{aligned} \ln\{W_i\} &= \rho E_i + Y_i'\xi + S_i'\eta + \breve{u}_i \qquad (5.19) \\ E_i &= Y_i'\delta + (Q_i' \otimes Y_i')\theta + S_i'\psi + (Q_i' \otimes S_i')\lambda + \breve{v}_i, \qquad (5.20) \end{aligned}$$

where $\lambda' = [\lambda_{1\cdot1}, \ldots, \lambda_{1\cdot50}, \lambda_{2\cdot1}, \ldots, \lambda_{2\cdot50}, \lambda_{3\cdot1}, \ldots, \lambda_{3\cdot50}]$. The selection procedures which we are going to apply should determine which regressors of the reduced forms (5.18) and (5.20) are in fact correlated with the error term and hence should not be included in the reduced form equation.

We apply the selection procedures A and B of section 5.3.1 and 5.3.2 only, since the other procedures are not feasible even for the model excluding S_i, where we have 30 potential (excluded) instruments (hence almost $2^{30} \approx 10^9$ subsets of instruments to search from). For the second implementation (including S_i) we have 180 potential instruments.

We apply the procedures employing (adapted versions of) the Sargan and Hansen tests (that is (4.6) with the weighting matrix (4.16) or (4.11)) with critical values based on the asymptotic null distribution.

The Hansen test we apply with the following modification: From the $n \times (1+k+l)$ matrix of all available data $D = [y, X, Z]$, it can be seen that the simple IV estimator and the Sargan tests (adapted or standard) which exploit a subset of the instruments can be calculated from the elements of the $(1+k+l) \times (1+k+l)$ matrix $D'D$ (see 4.14, 4.15 or 4.16). Hence, when applying the procedures to the data D we do not need to operate on the whole data matrix D but just on the much smaller $D'D$. Taking different subsets of the 'instruments' requires simply considering appropriate submatrices of $D'D$. However, when applying the Hansen test we need to calculate the weighting matrix, which involves

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\tilde{\theta})^2 z_i(c_s)z_i(c_s)', \tag{5.21}$$

see (4.11). This cannot be derived from $D'D$. Due to the 'residual factor' $(y_i - x_i'\tilde{\theta})$, it needs to be calculated from the original matrix D. In this example, when dealing with the second model, computation of a single Hansen test takes at least half a minute. With 180 instruments to consider (if we would need to reach the last stage) that would require calculating Hansen tests about 30000 times, which would take at least 250 hours!

Therefore we propose and use the following operational implementation for the procedures A and B:

- In the first stage use the Sargan test to find the initial set of instruments c_1^*.

- For that set calculate the Hansen test $J(c_1^*)$. If c_1^* is not rejected by the Hansen test, then, for the resulting $\tilde{\theta}(c_1^*)$ and the entire matrix of potential instruments Z, calculate
$$\mathbf{Q}^* \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\tilde{\theta}(c_1^*))^2 z_i z_i'. \tag{5.22}$$

- Then, in the various stages of the selection procedures, instead of calculating the

'official' weighting matrix for the Hansen test

$$\hat{\Omega}_a(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\tilde{\theta})^2 z_i(c_s)z_i(c_s)' - \frac{1}{n^2}Z(c_s)'u(\tilde{\theta})u(\tilde{\theta})'Z(c_s), \tag{5.23}$$

for $\tilde{\theta} = \tilde{\theta}(c_s)$, use the 'simplified' matrix

$$\hat{\Omega}_a(\tilde{\theta}) = \mathbf{Q}^*(c_s) - \frac{1}{n^2}Z(c_s)'u(\tilde{\theta})u(\tilde{\theta})'Z(c_s), \tag{5.24}$$

where $\mathbf{Q}^*(c_s)$ is obtained by keeping only the rows and columns of $\mathbf{Q}^*$ that are indicated by c_s.

- Now, re-run the first stage of the procedure when using this 'modified' Hansen test. The resulting c_1^* may or may not be the same as the one found initially by the Sargan test. We may again calculate $\mathbf{Q}^*$ for this new c_1^* and resulting $\tilde{\theta}(c_1^*)$. We may repeat this until c_1^* 'stabilizes'. For example in the model M_2 without the extra explanatory variables the Sargan test first found instruments labeled [126, 142] and subsequently the Hansen test found [2, 36] which was found to be 'stable'.

- Next stages could now use $\mathbf{Q}^*$ all the time or we could recalculate it each time we move up with the number of instruments, which in our example would require a maximum of 180 re-calculations (this is feasible). For simplicity, we could recalculate $\mathbf{Q}^*$ every several steps. We did that: for M_1 every two steps and every ten steps for M_2.

Obviously this modification is asymptotically valid since $\tilde{\theta}(c_s^*)$ is consistent for $c_s^* \in \mathcal{Z}^0$.

Results

Table 5.1 presents results for the implementations M_1 (without State dummies) and M_2 (with State dummies) with extra explanatory variables (β and π unrestricted) or without ($\beta = 0$, $\pi = 0$) referred to in the table as *unrestricted* and *restricted* respectively. It shows:

- $\hat{\rho}_{OLS}$: OLS estimates for the coefficient of education ρ (with standard and heteroscedastic robust standard errors underneath it);

- $B-P$: Breusch-Pagan heteroscedasticity test (with p-values underneath). In the regression of the squared OLS residuals on all the explanatory variables from the wage equation and the constant (hence, we exclude one year dummy), it is the F statistic on the significance of all the explanatory variables apart from the constant;

- F, F_h: the standard and heteroscedasticity robust F test on the strength/weakness of the instruments. In the most general reduced form equation (5.16) or (5.20) it is the F-test on the significance of the regressors which have been 'excluded' from the wage equation variables (the interaction variables involving Q_{ik});

- $\hat{\rho}_{GMM}, \hat{\rho}_{IV}$: the GMM (when applied with the Hansen test) and IV (when applied with the Sargan test) estimates resulting from procedures A and B. The % indicates at which significance level a procedure has been applied. For both the procedures A and B we used the asymptotic critical values at 5 or 1 percent significance levels based on the asymptotic chi-squared distribution. This way we find a set of instruments c_s for which any $c_{s+1} \in \mathcal{C}_{s+1}(c_s)$ is rejected at 5% or 1% significance level.

- CIK, CIK_h: the 95% confidence intervals obtained from the K-statistic (homoscedastic and heteroscedastic robust versions). The K-statistic is presented in the appendix below, see also Kleibergen (2002) and Kleibergen (2007).

- CIC is the 95% confidence interval obtained in the standard way: $[\hat{\rho} \pm 1.96SE(\hat{\rho})]$ (heteroscedasticity robust version when applied with GMM and homoscedastic when applied with IV)

- %: in percentage terms, how much shorter CIC is in comparison to CIK (when applied with IV) or CIK_h (when applied with GMM);

Table 5.2 shows the indices of the 'instruments' that were excluded by procedures A and B for different implementations. Procedure B did not exclude any instruments at

any significance levels for both the Hansen and the Sargan tests (indicated here by H and S respectively, followed by the nominal significance level). This Table also presents, with respect to the variable E, the DWH tests (and the corresponding p-values) for the 'instruments' that are the outcomes of the selection procedures. It checks whether the instrumental variables estimation for the wage equation is indeed necessary, i.e. whether E is endogenous or not. It is the t-test, in the wage equation, on the significance of the included residuals from the regression of education on all the instruments. See for example Davidson and MacKinnon (2004, p. 338).

The indices of the excluded 'instruments' (which are the interaction terms between State or Year dummy variables with the QOB dummy variable) should be interpreted as follows: Y_y^q means Season q ($q = 1, 2, 3$) Year $y + 1929$ ($y = 1, \ldots, 10$), S_s^q means Season q State s ($s = 1, \ldots, 50$). For implementation M_1, '$\times$' means that no instrument was excluded. For implementation M_2 and a given testing method, the second line displays instruments (between brackets) that were found common for both the restricted and unrestricted cases.

From Table 5.2 we note that the DWH test for the M_1 implementation, using procedure B, suggests that IV estimation is not needed (the p-values are close to 0.3 for the whole set of 30 instruments). Only when the instrument Y_3^1 is dropped then the p-values become close to 0.05 (0.0699 for the unrestricted case and 0.054 for the restricted). For M_2, the DWH test has quite small p-values both for the whole set of 180 instruments and for the reduced sets. We see there that the reduced sets of instruments found by the procedure A have smaller p-values than the ones on the whole set (from procedure B). That suggests that IV estimation is needed. That is also confirmed by the confidence intervals (CI's) given in Table 5.1. For the first implementation (M_1) they cover the OLS coefficient estimators when all the instruments are considered, and when the Y_3^1 instrument is dropped then CI's contain (or almost contain) the OLS estimate, which is very close to the lower bound of the CI. For the second implementation none of the CI's contain the OLS estimate.

We also note from Table 5.1 that the first stage F statistics for testing the (full set of) additional exogenous variables ($H_0 : \theta = 0$) for M_1 and ($H_0 : \theta = 0$, $\lambda = 0$) for M_2

are less than 5. That shows that we are dealing here with very weak instruments indeed. The 'weak instruments' robust K-statistic's CI's are then most appropriate to use. We see that the heteroscedastic robust and homoscedastic versions are very similar, but the CI's obtained in the classic way are different: its lower bound is almost the same as CIK's but the upper bound is much lower than those for CIK. That can also be seen from the '%' entry showing how much shorter CIC is in comparison to CIK: they are roughly 16-19% shorter than the *weak instruments robust* ones for implementation M_1 and 30% for implementation M_2.

The F statistics (not reported) on the instruments that were selected by procedure A gave very similar results to those on the whole set of variables (from procedure B). That is they were less than 3 for M_2 and less than 5 for M_1.

For implementations M_1 and M_2, Figures 5.6 and 5.7 depict, in the upper panels, the 'evolution' (over different stages of the selection procedures) of the 95% confidence intervals for ρ obtained using the K-statistic (homoscedastic version). Here, procedure A has been used with the Hansen (solid lines) and the Sargan tests (dashed lines). Middle lines show the evolution of the resulting estimates. In the left-hand panels the *extra exogenous variables* case is presented and in the right-hand panels we have the *no extra variables* case. In the lower panels, we show the corresponding p-values for the procedures A and B, for the Hansen (solid lines) and the Sargan (dashed lines) tests. The p-value lines for procedure B are those which are closest to one and the p-value lines for procedure A are those which are declining fast.

The circles in the CI panels, and the vertical line in the p-value panels, indicate where the procedures stopped at the 5% level. The horizontal line in the CI panels show the OLS estimate.

Note that the CI's would be the same for procedure B, since the A and B procedures evolve in the same way (in a stage s procedure B minimizes $J(c_s)$ and procedure A minimizes $J(c_s) - J(c^*_{s-1})$, hence the minimized c^*_s are the same for A and B). But, due to the difference in the p-values of the statistic considered procedure A can finish earlier.

We see that for procedure A the p-values were steadily decreasing whereas the p-values for procedure B were almost equal to one throughout the search, but dropped when the

Table 5.1: Estimators

	M_1		M_2	
	unrestricted	*restricted*	*unrestricted*	*restricted*
$\hat{\rho}_{OLS}$	0.063246	0.071081	0.062793	0.067339
$SE(\hat{\rho})$	(0.00033926)	(0.00033901)	(0.00034379)	(0.00034643)
$SE_R(\hat{\rho})$	(0.00037705)	(0.00038146)	(0.0003815)	(0.00038831)
$B-P$	35.525	15.058	11.436	5.1295
$DF(p-value)$	21 (0)	10 (0)	71 (0)	60 (0)
F	4.7474	4.9071	2.4276	2.5823
F_h	4.6245	4.8013	2.2626	2.4114
		A		
$\hat{\rho}_{GMM}-5\%$	0.095424	0.104270	0.086098	0.097922
CIK	(0.0603, 0.1416)	(0.0704, 0.1494)	(0.0692, 0.1240)	(0.0857, 0.1362)
CIK_h	(0.0609, 0.1432)	(0.0711, 0.1510)	(0.0699, 0.1260)	(0.0832, 0.1352)
CIC	(0.0609, 0.1300)	(0.0706, 0.1379)	(0.0671, 0.1051)	(0.0789, 0.1169)
%	0.1604	0.1577	0.3226	0.2692
$\hat{\rho}_{GMM}-1\%$	0.082131	0.090695	0.083545	0.092937
CIK	(0.0456, 0.1236)	(0.0547, 0.1328)	(0.0658, 0.1243)	(0.0811, 0.1355)
CIK_h	(0.0474, 0.1262)	(0.0566, 0.1354)	(0.0661, 0.1278)	(0.0779, 0.1354)
CIC	(0.0499, 0.1144)	(0.0590, 0.1224)	(0.0646, 0.1025)	(0.0742, 0.1117)
%	0.1815	0.1954	0.3857	0.3478
$\hat{\rho}_{IV}-5\%$	0.094700	0.103552	0.091920	0.103032
CIK	(0.0603, 0.1416)	(0.0704, 0.1494)	(0.0770, 0.1340)	(0.0903, 0.1409)
CIK_h	(0.0609, 0.1432)	(0.0711, 0.1510)	(0.0763, 0.1344)	(0.0879, 0.1399)
CIC	(0.0602, 0.1292)	(0.0700, 0.1371)	(0.0724, 0.1114)	(0.0843, 0.1218)
%	0.1513	0.1506	0.3158	0.2589
$\hat{\rho}_{IV}-1\%$	0.080552	0.103552	0.087103	0.098926
CIK	(0.0456, 0.1236)	(0.0704, 0.1494)	(0.0710, 0.1269)	(0.0866, 0.1415)
CIK_h	(0.0474, 0.1262)	(0.0711, 0.1510)	(0.0714, 0.1299)	(0.0840, 0.1427)
CIC	(0.0484, 0.1127)	(0.0700, 0.1371)	(0.0683, 0.1059)	(0.0804, 0.1175)
%	0.1756	0.1506	0.3274	0.3242
		B		
$\hat{\rho}_{GMM}-5,1\%$	0.082131	0.090695	0.083699	0.091872
CIC	(0.0499, 0.1144)	(0.0590, 0.1224)	(0.0648, 0.1026)	(0.0733, 0.1105)
CIK_h	(0.0474, 0.1262)	(0.0566, 0.1354)	(0.0666, 0.1314)	(0.0766, 0.1383)
%	0.1815	0.1954	0.4167	0.3971
$\hat{\rho}_{IV}-5,1\%$	0.080552	0.089115	0.083147	0.092818
CIC	(0.0484, 0.1127)	(0.0575, 0.1207)	(0.0645, 0.1018)	(0.0746, 0.1110)
CIK	(0.0456, 0.1236)	(0.0547, 0.1328)	(0.0658, 0.1264)	(0.0787, 0.1356)
%	0.1756	0.1908	0.3845	0.3603

Table 5.2: Rejected Instruments and the DWH test

	M_1		M_2	
	unrestricted	*restricted*	*unrestricted*	*restricted*
		A		
$H-5\%$	Y_3^1	Y_3^1	Y_3^1 $\left(Y_7^2 S_{29}^1 S_{48}^1 S_5^2 S_{20}^3 S_{22}^3 S_{44}^3\right)$	$Y_3^3 S_9^1 S_{24}^1 S_{19}^3$
$DWH,\ p-val$	(3.2853, 0.0699)	(3.7085, 0.0541)	(5.8866, 0.0153)	(11.7613, 0.0006)
$H-1\%$	$\times$	$\times$	$S_5^2 S_{22}^3$ $\left(S_{44}^3\right)$	$Y_3^3 S_{20}^3$
$DWH,\ p-val$	(1.1249, 0.2889)	(1.2645, 0.2608)	(4.6894, 0.0303)	(8.8341, 0.0030)
$S-5\%$	Y_3^1	Y_3^1	$Y_3^1 S_{43}^1 S_{22}^3$ $\left(S_{48}^1 S_5^2 S_{25}^2 S_{20}^3 S_{21}^3 S_{44}^3\right)$	$Y_3^3 S_9^1 S_{24}^1 S_{29}^1 S_{19}^3$
$DWH,\ p-val$	(3.2853, 0.0699)	(3.7085, 0.0541)	(8.7691, 0.0031)	(14.3993, 0.0006)
$S-1\%$	$\times$	Y_3^1	$Y_3^1 S_{48}^1 S_{22}^3$ $\left(S_5^2 S_{25}^2 S_{44}^3\right)$	Y_3^3
$DWH,\ p-val$	(1.1249, 0.2889)	(3.7085, 0.0541)	(6.4981, 0.0108)	(11.4342, 0.0007)
		B		
$DWH,\ p-val$	(1.1249, 0.2889)	(1.2645, 0.2608)	(4.6519, 0.0310)	(7.6364, 0.0057)

p-values of procedure A came close to the critical level, suggesting that this set is indeed suspicious. Due to the weakness of the whole set of instruments, and an overwhelming excess of 'valid instruments' (and low power due to the great number of degrees of freedom in the test), procedure B does not reject the instruments which procedure A does.

Note that since the p-values of procedure B are all above 5% we do not have to re-run it in order to get results at the 1% level. Both would give the same result, viz. all the 'instruments' are accepted. Also note that we do not need to re-run procedure A from scratch to get the results at the 1% significance level. We just need to continue the search from what procedure A has found at the 5% level (appropriately initializing the matrix $\mathbf{Q}^*$).

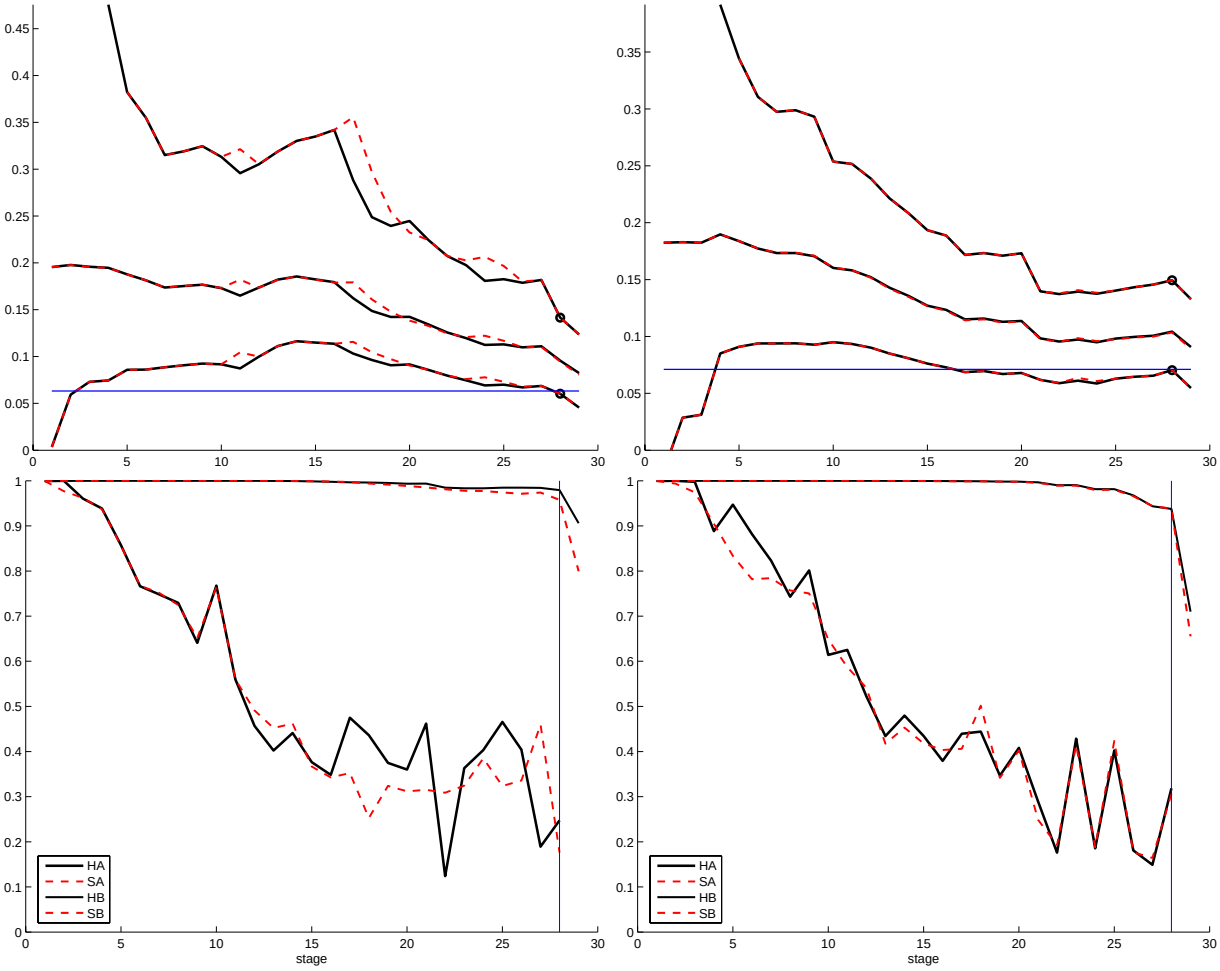

Figure 5.6: 95% CIK's and p-values over the consecutive selection stages for the M_1 implementation. Left hand graphs show the extra exogenous variables case, right hand graphs show the "no-extra" exogenous variables case.

For implementation M_1 the evolution of the estimates and CI's are almost equal for the Hansen and Sargan tests, but for implementation M_2 the two evolutions are quite distinct, though they conform at the end of the path. We also see that the CI's based on the Hansen implementations cover the OLS estimate (the horizontal line) for more than half way of the search and only later move away from it. The Sargan implementations on the other hand produce CI's that almost always are above the OLS estimate. Despite that, the p-values are similar for both tests.

We can assign an overall p-value to the final results using the bootstrap in the following way. Obtain bootstrapped data as in Section 4.3.3. Then, run procedure A such that it reaches the final stage and record the largest value of the (incremental) statistic over the

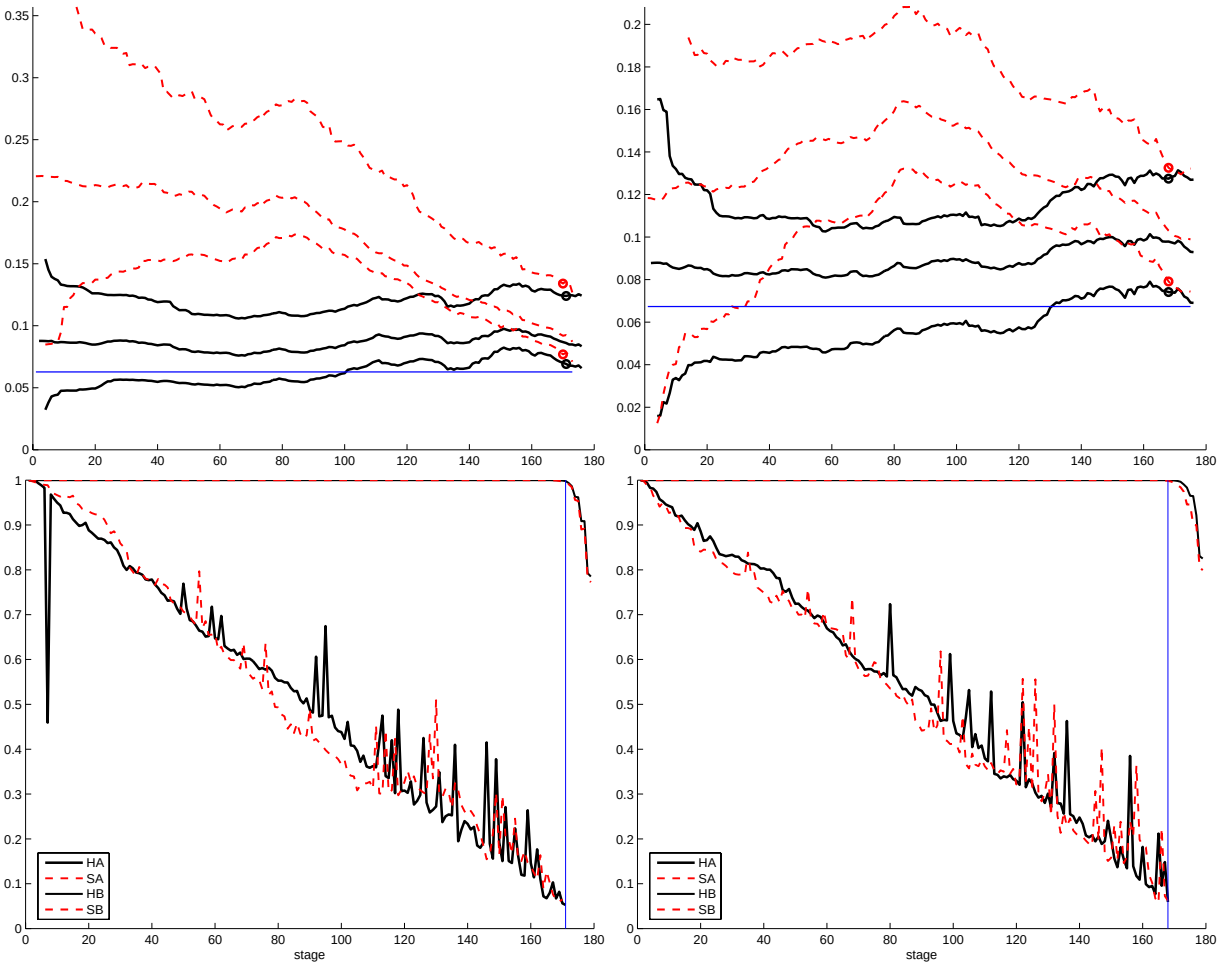

Figure 5.7: 95% CIK's and p-values over the consecutive selection stages for the M_2 implementation. Left hand graphs show the extra exogenous variables case, right hand graphs show the "no-extra" exogenous variables case

stages, T_n^b. Equivalently, record the smallest 'p-value' of the statistic, $p_n^b = 1 - \chi_1^2(T_n^b)$. Then, for the original data obtain the value of the statistic associated with the rejected instruments, T_n, (or the value $p_n = 1 - \chi_1^2(T_n)$). Finally, compare this value with those from the bootstrap distribution. The bootstrap estimate of the asymptotic p-value for the rejected instruments will be

$$\frac{1}{B}\sum_{b=1}^{B} I(T_n^b > T_n) = \frac{1}{B}\sum_{b=1}^{B} I(p_n^b < p_n).$$

For example, based on $B = 2000$ repetitions and considering only the Sargan test, the p-value associated with the M_1-*unrestricted* case, where the instrument Y_3^1 was rejected, is found to be 0.1280. For the M_2-*unrestricted* case, the same instrument was rejected with the bootstrap p-value estimated to be 0.1690, while the instruments reported in Table 5.2 under $S - 1\%$ entry have the p-value of 0.2135.

These p-values are not small enough to reject the validity of instruments under the usual 95% significance level. On the other hand, they are neither exceptionally large to accept confidently the validity of the instruments.

Our analysis reveals that some of the instrumental variables used for the returns on education equation are possibly not exogenous. That might suggest that the whole set of instruments could be invalid due to the specific nature of those variables (interactions of the year dummies with the quarter of birth). As mentioned in the introduction, that could be caused by, for example, some correlation between the quarter of birth with a 'positive attitude' characteristic. Also, an astrologist could argue, the quarter of birth is correlated with a zodiac sign of a person which in turn carries some cosmic predispositions of a person that affect the earnings (and the schooling), causing endogeneity of the QOB. More recently, Buckles and Hungerman (2008)[1] give further evidence why QOB is an

[1]I allow myself to quote the comment on the very interesting article in *The Economist* 'Cause and defect' from 15th of August 2009 made by Prof. David Jaeger:
"After 15 years of the weak instrument literature, it's hard to believe that the Economist can claim that Angrist and Krueger's instrument is valid. See Bound, Jaeger, and Baker (1995) for the beginning of this critique of Angrist and Krueger's instruments, but there is a huge literature that follows.
Quarter of birth meets neither the relevance (at least in AK's preferred specification) nor the exclusion restriction assumption required for IV. See Buckles and Hungerman's excellent paper that, one hopes, is the death knell for any paper that claims quarter of birth is unrelated to outcomes except through the compulsory schooling/school age starting law mechanism. There's also a very good paper by

invalid instrument.

5.6 Alternative inference and a sensitivity analysis

In this section, we will consider the Angrist and Krueger (1991) data again, but now against the background of our findings in the earlier chapters on the effects of possibly invalid instruments. We will utilize the asymptotic expressions for the inconsistent OLS and IV estimator distributions in order to depict 'bias corrected' versions of the estimators and the implied approximating confidence intervals around them. This alternative inference is based on making varying assumptions on the degree of simultaneity, and, when external instruments have been used, on their possible degree of invalidity. Hence, because it is based on unobserved nuisance parameters one might classify such inference as unfeasible. We will demonstrate nonetheless that it allows a useful sensitivity analysis. We will see that for these particular data, this unfeasible bias corrected OLS inference is more attractive than the even more unfeasible (depending on assumptions regarding both simultaneity and instrument invalidity) bias corrected IV inference.

The analysis

We will analyze the implementation 'M_1-unrestricted' from the previous example, only. Written in SEM form, we have

$$y = x\beta + Z\gamma + \varepsilon$$

$$x = Z\Pi + W\Gamma + v,$$

Barua and Lang (2009) that shows that the compulsory schooling/school age starting law mechanism doesn't meet the monotonicity requirements for a local average treatment effect interpretation of the IV estimate of the effects of school entry on outcomes.

Despite all this, Angrist and Pischke in their otherwise excellent *Mostly Harmless Econometrics* once again belabor the AK results and present them as if they are sensible estimates of the returns to education.

Why is this? How can a result which has for many reasons and by many authors been shown to have problems persist in being held up as a shining example of the usefulness of the IV technique? Part of the answer is that more than any other IV story (and every good IV paper has a story), AK's story is really good. Incredibly clever. Easy to see in graphs. Believable. So, we want to think that AK's instrument is sensible and good. Because if the AK story doesn't hold, then lots of other IV stories are probably invalid. But AK's story doesn't hold... and with it goes much of the natural experiment movement."

where y is the log of wage, x is education level (hence, β is now what we called ρ before), Z are the exogenous explanatory variables and W are the further 'instruments' (where W_{i3} is possibly invalid). Regressing off Z in both equations we get

$$M_Z y = M_Z x\beta + M_Z \varepsilon$$

$$M_Z x = M_Z W\Gamma + M_Z v.$$

For the sake of simplicity, below we will write y for $M_Z y$, x for $M_Z x$, and so on. Then we have the simplified two equations

$$y = x\beta + \varepsilon \tag{5.25}$$

$$x = W\Gamma + v,$$

where now

$$\hat{\beta}_{ols} = x'y/(x'x) = \beta + x'\varepsilon/(x'x)$$

$$\hat{\beta}_{giv} = (x'P_W y)/(x'P_W x) = \beta + (x'P_W \varepsilon)/(x'P_W x).$$

If

$$x = \bar{x} + \varepsilon\xi$$

with $E(x|\bar{x}) = \bar{x}$, then

$$\beta^*_{ols} = \operatorname*{plim}_{n\to\infty} \quad \hat{\beta}_{ols} = \beta + \xi\sigma^2_\varepsilon/\sigma^2_x (= \beta + \ddot{\beta}_{ols}),$$

and hence, an unfeasible consistent bias corrected estimator for β is

$$\hat{\beta}^\sharp_{ols} \equiv \hat{\beta}_{ols} - \xi\sigma^2_\varepsilon/\sigma^2_x. \tag{5.26}$$

Since

$$\rho_{x\varepsilon} = \sigma_{x\varepsilon}/(\sigma_x\sigma_\varepsilon) = \xi\sigma^2_\varepsilon/(\sigma_x\sigma_\varepsilon)$$

we have

$$\xi = \rho_{x\varepsilon}\sigma_x/\sigma_\varepsilon, \tag{5.27}$$

giving

$$\hat{\beta}^{\sharp}_{ols} = \hat{\beta}_{ols} - \rho_{x\varepsilon}\sigma_{\varepsilon}/\sigma_x. \tag{5.28}$$

Next, we make an attempt to operationalize this bias corrected estimator.

Since

$$\operatorname*{plim}_{n\to\infty} \quad \varepsilon'\varepsilon/n = \sigma^2_{\varepsilon} \quad \text{and} \quad \operatorname*{plim}_{n\to\infty} \quad x'x/n = \sigma^2_x, \tag{5.29}$$

we may define consistent estimators of σ^2_{ε} and σ^2_x as follows

$$\check{\sigma}^2_x = x'x/n \quad \text{and} \quad \check{\sigma}^2_{\varepsilon} = \check{\varepsilon}'\check{\varepsilon}/n, \tag{5.30}$$

where

$$\check{\varepsilon} \equiv y - x\check{\beta},$$

and where $\check{\beta}$ is obtained from substituting (5.30) in (5.28) :

$$\check{\beta}(\rho_{x\varepsilon}) \equiv \hat{\beta}_{ols} - \rho_{x\varepsilon}\check{\sigma}_{\varepsilon}/\check{\sigma}_x \equiv \hat{\beta}_{ols} - \ddot{\hat{\beta}}_{ols}. \tag{5.31}$$

Since

$$\check{\varepsilon} = \hat{\varepsilon} + x\rho_{x\varepsilon}\check{\sigma}_{\varepsilon}/\check{\sigma}_x,$$

and because $x'\hat{\varepsilon} = 0$ we obtain

$$\check{\sigma}^2_{\varepsilon} = \check{\varepsilon}'\check{\varepsilon}/n = \hat{\varepsilon}'\hat{\varepsilon}/n + \rho^2_{x\varepsilon}\check{\sigma}^2_{\varepsilon}(x'x/n)/\check{\sigma}^2_x = \hat{\varepsilon}'\hat{\varepsilon}/n + \rho^2_{x\varepsilon}\check{\sigma}^2_{\varepsilon},$$

so we can solve for $\check{\sigma}^2_{\varepsilon}$, giving:

$$\check{\sigma}^2_{\varepsilon} = (\hat{\varepsilon}'\hat{\varepsilon}/n)/(1 - \rho^2_{x\varepsilon}). \tag{5.32}$$

Assuming $\rho_{x\varepsilon}$ to be known, (5.31) is a consistent estimator for β, which is based entirely on OLS.

Using our conditional asymptotic result

$$\hat{\beta}_{ols} - \ddot{\beta} \overset{a}{\sim} N(\beta, \frac{1}{n}V_{ols}(\sigma^2_x, \sigma^2_{\varepsilon}, \rho_{x\varepsilon})),$$

and replacing the 'population moments' by their sample versions we use the approximation

$$\check{\beta}_{ols} = \hat{\beta}_{ols} - \rho_{x\varepsilon}\check{\sigma}_\varepsilon/\check{\sigma}_x \overset{a}{\approx} N(\beta, \frac{1}{n}V_{ols}(\check{\sigma}_x^2, \check{\sigma}_\varepsilon^2, \rho_{x\varepsilon})) \tag{5.33}$$

with

$$V_{ols}(\check{\sigma}_x^2, \check{\sigma}_\varepsilon^2, \rho_{x\varepsilon}) = \frac{\check{\sigma}_\varepsilon^2}{\check{\sigma}_x^2}(1-\rho_{x\varepsilon}^2)(1-2\rho_{x\varepsilon}^2+2\rho_{x\varepsilon}^4)$$

based on expression (2.57) that we found in Chapter 2 for the first illustration.

Figure 5.8 shows for a range of $\rho_{x\varepsilon}$ values the bias corrected $\check{\beta}_{ols}$ estimator together with the implied asymptotic 95% confidence intervals based on (5.33). We see that the bounds are extremely narrow, as we already saw for the case $\rho_{x\varepsilon} = 0$ in the previous chapter (Table 5.1) , where $\hat{\beta}_{ols} \approx 0.063$. We also saw that the IV estimator (excluding instrument Y_3^1) was about 0.0947. If this were indeed the true value of β that would, according to Figure 5.8, amount to having $\rho_{x\varepsilon} \approx -0.15$.

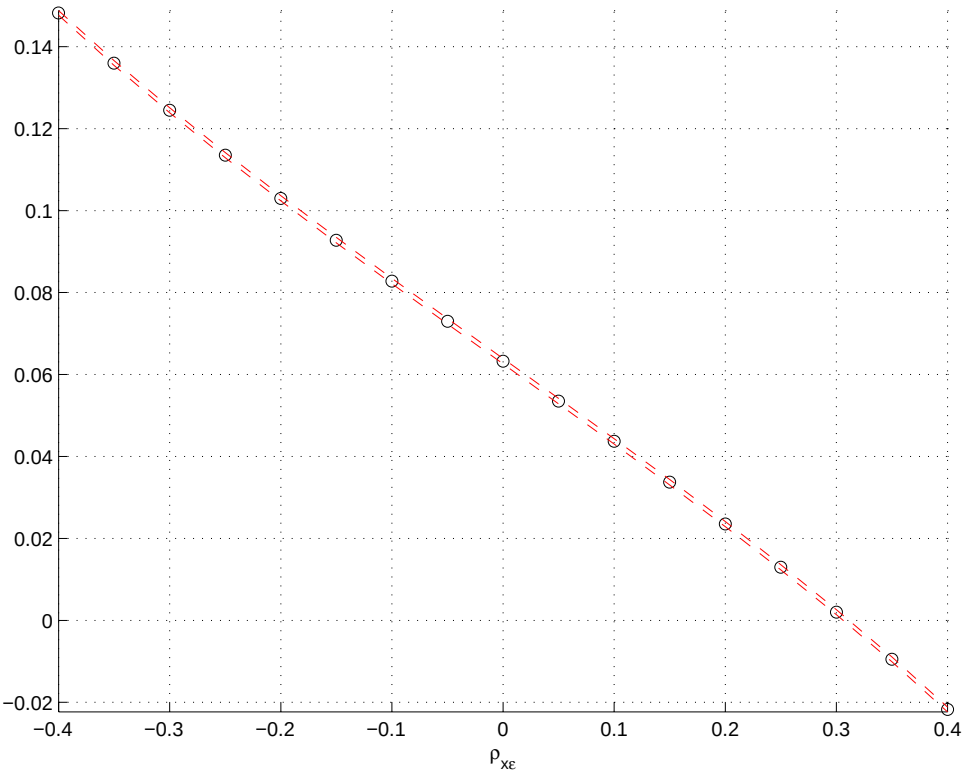

Figure 5.8: Bias corrected OLS (conditional on $\rho_{x\varepsilon}$) together with the asymptotic 95% confidence bounds.

We can perform similar derivations for the IV case. But first, to examine how weak

actually the instruments are, in Figure 5.9 we plot the sample correlations between x and the 30 individual instruments collected in W.

Analytically, see Fisher (1915), for two uncorrelated ($\rho = 0$) series we have the Fisher z-transform

$$Z = \frac{1}{2}\log\frac{1+R}{1-R} \overset{a}{\sim} N(0, \frac{1}{n-3}),$$

where R is the sample correlation and the asymptotic result follows from Hawkins (1989). Hence, given our sample size, we will have the approximate 95% interval for testing (using $Z \approx R$) whether the population correlation ρ is equal to zero is: $[-0.0034, 0.0034]$, and at 99%: $[-0.0045, 0.0045]$. In Figure 5.10 we plot Z and the 95% and 99% approximate bounds. That suggests that the population correlations in our data are not entirely zero (the correlation between x and the instruments W_1, W_3, W_6 and W_{22} lie outside the bounds), but surely this illustrates that the instrumental variables are extremely weak.

From Chapter 3 we have

$$\hat{\beta}_{giv} - \ddot{\beta}_{giv} \overset{a}{\sim} N(\beta, \frac{1}{n}V_{iv}(\Sigma_{W'x}, \Sigma_{W'W}, \sigma_x^2, \sigma_\varepsilon^2, \rho_{x\varepsilon}, \zeta)).$$

Again, replacing the population moments with the sample versions we use the approximation

$$\hat{\beta}_{giv} - \breve{\ddot{\beta}}_{giv} \overset{a}{\approx} N(\beta, \frac{1}{n}V_{iv}(\frac{1}{n}W'x, \frac{1}{n}W'W, \frac{1}{n}x'x, \breve{\sigma}_\varepsilon^2, \rho_{x\varepsilon}, \breve{\zeta})), \tag{5.34}$$

where we take $\breve{\sigma}_\varepsilon^2$ from (5.32)[2] and $\breve{\ddot{\beta}}_{iv}$, $\breve{\zeta}$ are obtained as follows.

Let

$$R_{W\varepsilon} = \Sigma_{W'W}^{-1/2}\sigma_\varepsilon^{-1}\Sigma_{W'\varepsilon} = \Sigma_{W'W}^{-1/2}\sigma_\varepsilon\zeta$$

be a vector of population correlations between W_i and ε_i, then

$$\zeta = \sigma_\varepsilon^{-1}\Sigma_{W'W}^{1/2}R_{W\varepsilon}.$$

[2]We could calculate $\breve{\sigma}_\varepsilon^2$ from the implied IV residual, similarly to how we obtained it from the OLS residuals (for a given $\rho_{x\varepsilon}$), but, since V_{iv} depends on $\rho_{x\varepsilon}$ anyway, we may as well take $\breve{\sigma}_\varepsilon^2$ from the OLS result. This is also easier to obtain and since $\breve{\beta}_{ols}$ is more precisely 'estimated' it should be also more accurate.

Now we define

$$\hat{\zeta}(R_{W\varepsilon}) = \check{\sigma}_{\varepsilon}^{-1}(\frac{1}{n}W'W)^{1/2}R_{W\varepsilon}.$$

Since

$$\ddot{\beta}_{giv} = \sigma_{\varepsilon}^2(\Sigma_{x'W}\Sigma_{W'W}^{-1}\zeta)/(\Sigma_{x'W}\Sigma_{W'W}^{-1}\Sigma_{W'x})$$

we use

$$\check{\ddot{\beta}}_{giv} = \check{\sigma}_{\varepsilon}^2(x'W(W'W)^{-1}\check{\zeta})/(x'P_Wx) = \check{\sigma}_{\varepsilon}(x'W(W'W)^{-1/2}R_{W\varepsilon})/(x'P_Wx).$$

In our analysis of the Angrist and Krueger (1991) data, from the previous example, the W_{i3} was rejected by the Sargan test. Hence, we take

$$R_{W\varepsilon} = (0, 0, \rho_{w_3\varepsilon}, 0, \ldots, 0)'.$$

Figure 5.11, for a range of $\rho_{w_3\varepsilon}$ values, shows $\check{\beta}_{giv}$ and the approximate asymptotic 95% confidence bounds based on (5.34). The 'cloud' effect for the IV lines is due to plotting, for a given $\rho_{w_3\varepsilon}$, several cases for different $\rho_{x\varepsilon} \in [-0.4, 0.4]$.

Since $\rho_{x\varepsilon}$ does not affect $\check{\beta}_{giv}$ and the confidence bounds much, Figure 5.12, for a range of $\rho_{x\varepsilon} = \rho_{w_3\varepsilon}$ values, shows $\check{\beta}_{ols}$, $\check{\beta}_{iv}$ and their approximate asymptotic 95% confidence bounds together. The 'cloud' effect for the IV lines now disappears because now arbitrarily (but also rather inconsequentially) $\rho_{x\varepsilon} = \rho_{w_3\varepsilon}$.

From the previous subsection we saw that the confidence intervals based on the K-statistic were about 20% wider than the 'classic' ones but with the 'lower' bounds almost the same. That is, due to the weak instrumentation confidence intervals around $\check{\beta}_{giv}$ should most likely be approximately 20% wider.

We saw that the bias corrected OLS estimator for a given $\rho_{x\varepsilon}$ is very precisely estimated. The same cannot be said about bias corrected IV. For example if we are willing to assume that $\rho_{x\varepsilon} \in [-0.1, 0.1]$, then from Figure 5.12 we note that it corresponds to having approximately $\check{\beta}_{ols} \in [0.05, 0.08]$. But, when $\rho_{w_3\varepsilon} \in [-0.1, 0.1]$ then this amounts to having $\check{\beta}_{giv} \in [0.03, 0.13]$, more than 3 times wider than the corresponding OLS interval. Whereas, $\check{\beta}_{ols} \in [0.03, 0.13]$ corresponds approximately with $\rho_{x\varepsilon} \in [-0.32, 0.17]$. This

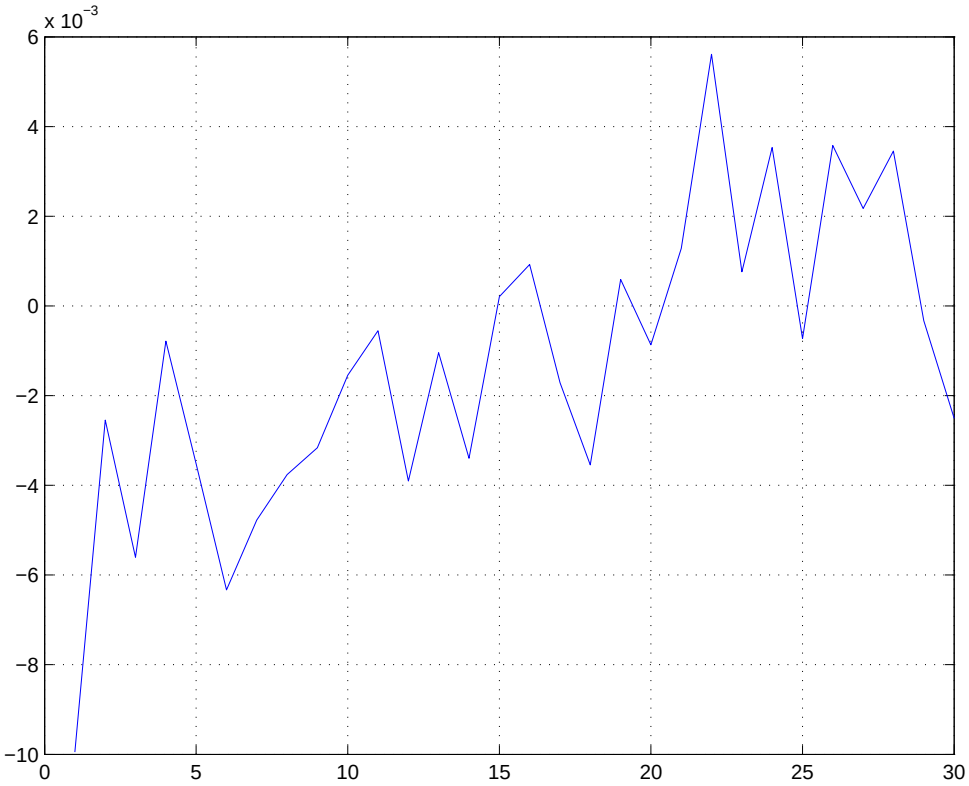

Figure 5.9: Sample correlations between x and individual instruments collected in W.

better precision surely makes the bias corrected OLS more attractive than bias corrected IV, in this extremely weak instruments case.

5.7 Conclusions

In this chapter we proposed three procedures for detecting invalid instruments from a possibly very large set of potential ones. Two of those procedures are computationally feasible due to their sequential nature. A Monte Carlo study reveals that all the procedures studied are vulnerable to weak instruments and that the sample size and instrument invalidity should be substantial in order for the procedures to show good power in detecting the invalid instruments. Nonetheless, for non serious invalidity of an instrument the resulting distribution of the estimates using the instruments selected (which also include the invalid ones) is almost centered around the true value for all the procedures.

In an empirical example we analyzed the Angrist and Krueger (1991) models where we

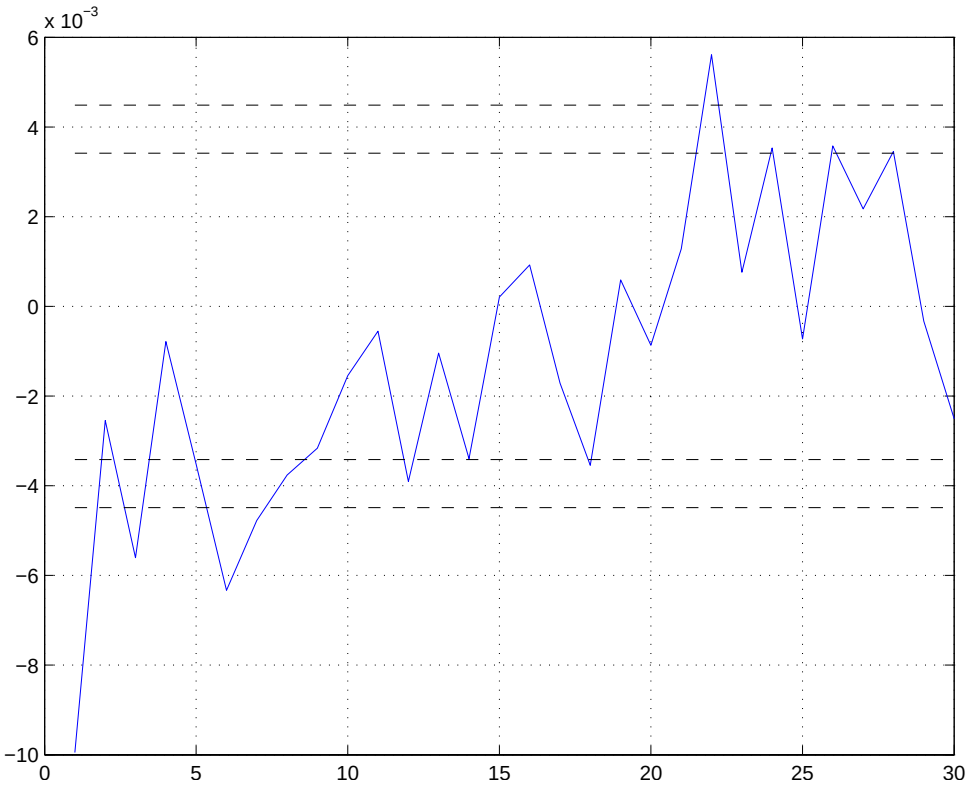

Figure 5.10: Transformed sample correlations (Z) between x and individual instruments collected in W together with the 2.5% and 0.5% one-sided critical values.

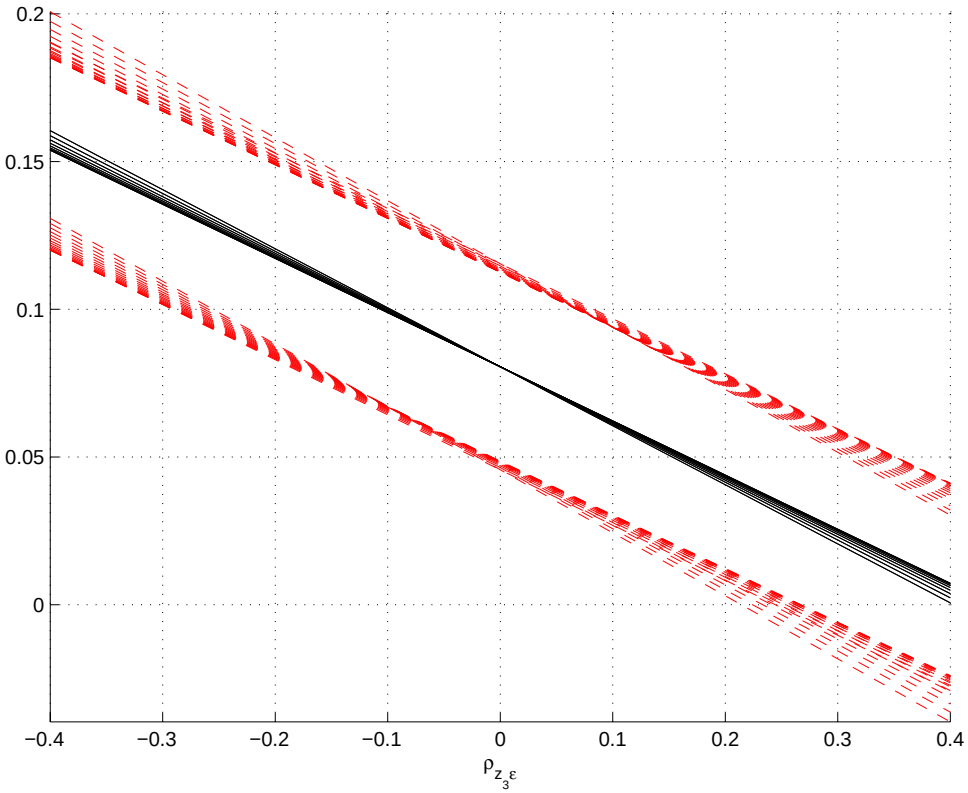

Figure 5.11: Bias corrected IV (solid lines) together with its asymptotic 95% confidence bounds (dashed).

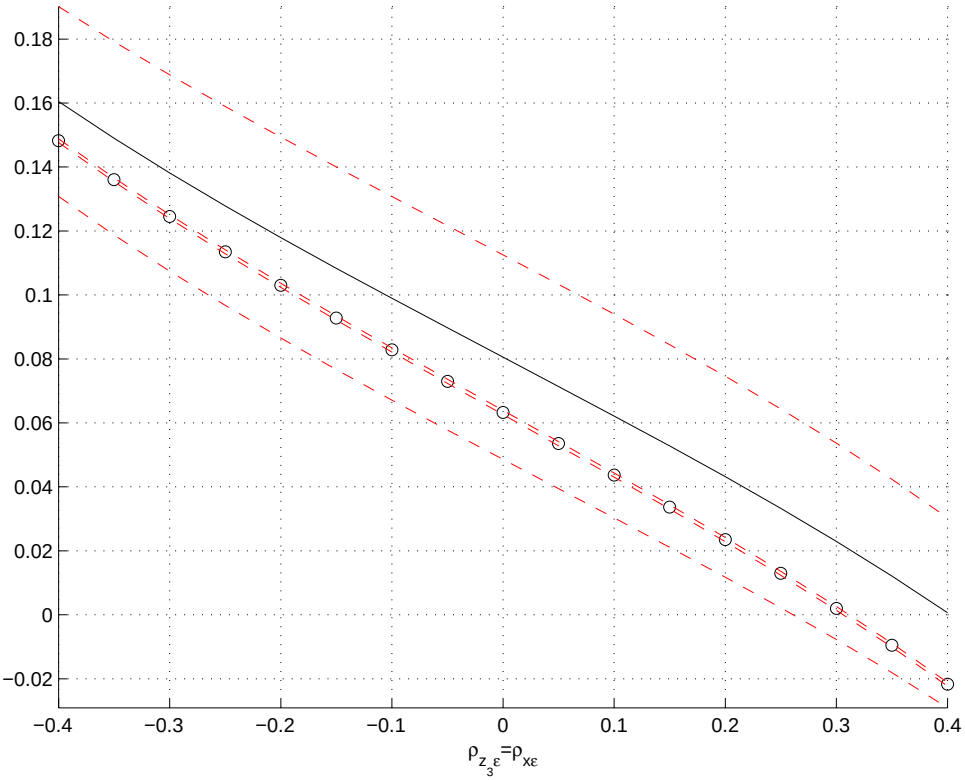

Figure 5.12: Bias corrected OLS (circles) and IV (solid line) together with their asymptotic 95% confidence bounds (dashed).

have an abundance of instruments to search from. That makes only two of our sequential procedures (A and B) computationally feasible. Procedure A, which utilizes the incremental version of the test statistics finds that some of the instruments are invalid. That suggests that the entire set of instruments might be in trouble due to the specific nature of the instruments. The invalidity of QOB as instrument could possibly be explained by that it is indeed correlated with some excluded characteristics that affect the earnings of the individual.

In the final section, applying the results from the earlier chapters to one of the model specifications, we find that making assumptions about the invalidity of the instruments produces less attractive inference based on bias corrected IV than similar inference based on making assumptions about simultaneity of the endogenous regressor using biased corrected OLS. Because of extreme weakness of the instruments, corrected OLS will beat corrected IV in terms of precision of that correction.

5.8 Appendix

Potential traps in the moment selection

For identification reasons we assumed that there is only one θ_0 satisfying the moment conditions $g(c_s, \theta_0) = 0$ for $c_s \in \mathcal{Z}^0$. Hence, for the linear model

$$y = X\theta + u, \tag{5.35}$$

where $\theta = \theta_0$ in the DGP, we have from (4.13)

$$\hat{\theta} = (X'Z\hat{\Omega}^{-1}X'Z)^{-1}X'Z\hat{\Omega}^{-1}Z'(X\theta_0 + u)$$

and so

$$\operatorname*{plim}_{n\to\infty}\hat{\theta} = \theta_0 + (\Sigma_{X'Z}\ddot{\Omega}^{-1}\Sigma_{Z'X})^{-1}\Sigma_{X'Z}\ddot{\Omega}^{-1}\Sigma_{Z'u} = \theta_0 + \ddot{\theta},$$

where $\ddot{\Omega}^{-1} = \operatorname*{plim}_{n\to\infty}\hat{\Omega}^{-1}$. If $\Sigma_{Z'u} \neq 0$ (but there exists a c_s for which $\Sigma_{Z(c_s)'u} = 0$) then $\ddot{\theta} \neq 0$.

If we rewrite model (5.35) as

$$y = X(\theta_0 + \ddot{\theta}) + u - X\ddot{\theta} = X\bar{\theta} + \bar{u}, \tag{5.36}$$

then

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}Z'\bar{u}=\Sigma_{Z'u}-\Sigma_{Z'X}\ddot{\theta}.$$

Hence, if there exist some $\ddot{\theta}$ for which $\Sigma_{Z'u}=\Sigma_{Z'X}\ddot{\theta}$ then the variables Z would 'become' 'valid' instruments for model (5.36) where $\bar{\theta}\neq\theta_0$ would be the 'true' parameter. Hence, if $\Sigma_{Z(\ddot{c}_s)'u}=\Sigma_{Z(\ddot{c}_s)'X}\ddot{\theta}$, for some s, $\ddot{\theta}$, and $\ddot{c}_s\notin\mathcal{Z}^0$ then the Hansen test statistic would not 'recognize' (asymptotically) $Z(\ddot{c}_s)$ as being invalid instruments.

In our example we have

$$\Sigma_{Z'u}=[\rho_{Z_1u},0,0,0]',\ \Sigma_{Z'X}=[\pi_1\sqrt{1-\rho^2_{Z_1u}},\pi_2,\pi_3,\pi_4]'$$

and if $\rho_{Z_1u},\pi_1,\pi_2,\pi_3,\pi_4$ are all different from 0 then we cannot find $\ddot{\theta}$ and $\ddot{c}_s\notin\mathcal{Z}^0$ for which $\Sigma_{Z(\ddot{c}_s)'u}=\Sigma_{Z(\ddot{c}_s)'X}\ddot{\theta}$. But, for example, if we had

$$\Sigma_{Z'u}=[\rho_{Z_1u},\rho_{Z_2u},0,0]',$$

then it would be indeed possible to find such a combination. So, we would need to exclude it from the simulation in order to satisfy Assumption 5.1(a).

In practice, we could use a Hausman test to detect cases like that, but we still would be left with the dilemma: which subsets of instruments are indeed valid.

K-statistic

For the model

$$\begin{aligned}y &= X\theta+\varepsilon\\ X &= Z\Pi+V\end{aligned}$$

Kleibergen's (2002) statistic is given by

$$K(\theta_0)=\frac{(y-X\theta_0)P_{P_ZD}(y-X\theta_0)}{s_{\varepsilon\varepsilon}},$$

with

$$\begin{aligned}D &= X-(y-X\theta_0)s_{\varepsilon v}/s_{\varepsilon\varepsilon}\\ s_{\varepsilon v} &= \frac{1}{n-k}(y-X\theta_0)'M_ZX\equiv\frac{1}{n-k}\tilde{\varepsilon}'\tilde{X}\\ s_{\varepsilon\varepsilon} &= \frac{1}{n-k}(y-X\theta_0)'M_Z(y-X\theta_0)\equiv\frac{1}{n-k}\tilde{\varepsilon}'\tilde{\varepsilon},\end{aligned}$$

where $\tilde{X} = M_Z X$ are the residuals from regressing X on Z. Kleibergen's (2007) heteroscedastic robust version is

$$K^h(\theta_0) = (y - X\theta_0)P_{\tilde{Z}}D(D'P_{\tilde{Z}}D)^{-1}D'P_{\tilde{Z}}(y - X\theta_0),$$

with

$$\begin{aligned}
D &= [D_1, \ldots, D_k], \quad D_i = X_i - S_{\varepsilon v_i}P_{\tilde{Z}}(y - X\theta_0) \\
P_{\tilde{Z}} &= Z(Z'S_{\varepsilon\varepsilon}Z)^{-1}Z' \\
S_{\varepsilon\varepsilon} &= diag(\tilde{\varepsilon}\tilde{\varepsilon}') = \begin{pmatrix} \tilde{\varepsilon}_1^2 & & \\ & \ddots & \\ & & \tilde{\varepsilon}_n^2 \end{pmatrix} \\
S_{\varepsilon v_i} &= diag(\tilde{\varepsilon}\tilde{X}_i') = \begin{pmatrix} \tilde{\varepsilon}_1\tilde{X}_{i1} & & \\ & \ddots & \\ & & \tilde{\varepsilon}_n\tilde{X}_{in} \end{pmatrix}.
\end{aligned}$$

Here, $\tilde{X}_i$ denotes the i-th column of $\tilde{X}$. Note that

$$Z'D_i = Z'X_i - Z'S_{\varepsilon v_i}Z(Z'S_{\varepsilon\varepsilon}Z)^{-1}Z'(y - X\theta_0)$$

and if $S_{\varepsilon v_i} = s_{\varepsilon v_i}$ and $S_{\varepsilon\varepsilon} = s_{\varepsilon\varepsilon}$ then $Z'D_i$ reduces to

$$Z'X_i - Z'(y - X\theta_0)s_{\varepsilon v_i}/s_{\varepsilon\varepsilon}$$

which shows that the $K(\theta_0)$ statistic is a special case of the 'heteroscedastic version' $K^h(\theta_0)$.

If we have additional exogenous variables in the model, i.e.

$$\begin{aligned}
y &= X\theta + W\delta + \varepsilon \\
X &= Z\Pi_Z + W\Pi_W + V,
\end{aligned}$$

then the form of the above statistics remains the same with y, X, Z replaced by the residuals $\hat{y} = M_W y, \hat{X} = M_W X, \hat{Z} = M_W Z$.