Selective Instrumental Variable Regression

Alberto Abadie, Jiaying Gu and Shu Shen

February 11, 2015

Incomplete Draft, Do Not Circulate

Abstract

In this paper, we propose consistent data-driven selective instrumental variable (IV) regression methods that improve the efficiency of IV regression when the first stage correlation between the instrument and the endogenous variable is heterogeneous. The improvement in efficiency is important because in applications inferential procedures robust to weak instruments often generate uninformative confidence intervals. Exploring heterogeneity in the first stage correlation, our procedures either reduce the length of the confidence intervals for the causal effect parameters or they reduce the probability of producing unbounded confidence intervals. We show that the proposed methods are consistent and that the efficiency gain can be quite large in some situations with substantial heterogeneity in first stage effects. As a side product of the paper, we also show that a naive selective IV regression procedure used in some applied literature is inconsistent due to pre-testing bias and tends to over-reject null hypotheses. Monte Carlo experiments show that the proposed methods have good small sample performance. We apply the methods to study the return to compulsory schooling using census data in the United States. We demonstrate that the proposed methods substantially tighten the confidence interval of the return to compulsory education parameter, often from unbounded intervals generated by classic AR tests to informative intervals.

1 Introduction

Weak instruments, that is, instruments that correlate weakly with the endogenous variable, are pervasive in empirical economics. Weak instruments not only bias two stage least squares (2SLS) towards ordinary least squares (OLS), but also nullify classical inferential procedures. A rich weak instrumental variable (IV) literature has focused on developing inferential methods that are robust to both weak and non-weak instruments. For example, Staiger and Stock (1997) and Dufour (1997) propose to use the Anderson and Rubin (1949) test, (AR) thereafter, and Moreira (2003) develops a Conditional Likelihood Ratio (CLR) test that is more powerful than the AR test when the number of instrument exceeds the number of endogenous covariates.[1] As might be expected, robust inferential procedures such as AR and CLR tests often generate uninformative confidence intervals in the presence of weak instruments. In this article we take a different angle on IV regression with potentially weak instruments. We notice that, in many applied problems, the first stage correlation between the instrument and the endogenous covariate is heterogeneous. In such situations, it may be worthwhile to take advantage of the first stage heterogeneity and develop methods that can improve the efficiency of IV regression by using only observations for which the instruments correlate with the endogenous variables. As far as we know, the consistent data-driven selective IV regression approach we propose in this article is the first attempt to exploit first stage heterogeneity to improve efficiency of IV regression.

Applied researchers have long recognized that restricting IV regression to subsamples with a strong first stage may help alleviate problems caused by weak instruments. For example, in the literature on return to compulsory schooling (e.g., Oreopoulos, 2006; Lleras-Muney, 2005) researchers have used variations in compulsory schooling laws to instrument for education attainment, and IV regressions are typically run on white students only because

[1] See Andrews and Stock (2005) for an excellent survey on IV regression with potentially weak instruments.

data suggest that black students were weakly affected by changes in compulsory schooling laws.[2] Similarly, Cervellati, Jung, Sunde, and Vischer (2014) find that the instrument used in an influential article by Acemoglu, Johnson, Robinson, and Yared (2008) on income and democracy is weak for a sample of non-colonies but much stronger for a sample of former colonies. Motivated by this finding Cervellati, Jung, Sunde, and Vischer (2014) restrict the sample to former colonies and find a negative and significant effect of income on democracy. However, as we formally show later, such naive selection method based on subsample first stage correlation is invalid and tend to generate overly large t-statistics. Intuitively, because the naive method uses subsamples with large t-statistics on the instrument in first stage regressions, it picks out not only subsamples with (true) strong first stage effects but also subsamples with large correlations between the first stage error term and the instrument. Because the error terms in the first stage and the second stage are correlated, selecting subsamples on the basis of the first stage t-statistic results in violations of the exclusion restriction and leads to over rejection in significant tests.

In this article, we prove that the naive selective IV regression method used in the applied literature is inconsistent and show that it may cause substantial over-rejection in finite samples. Then we consider a simultaneous equation model with heterogeneous first stage effects and propose consistent data-driven selective IV regression methods that are potentially more efficient than the full sample method. Like the naive method, the improvement in efficiency comes from strengthening the first stage correlation through multiple testing of first stage effects. But unlike in the naive method, the proposed method preserves the validity of the exclusion restriction and produces correct statistical inference. Monte Carlo experiments suggest that the proposed method has very good behavior in finite samples.

The remainder of this article is organized as follows. Section 2 sets up a simulaneous

[2]The footnote 44 (page 209) of Lleras-Muney (2005) justifies the exclusion of blacks, "Lleras-Muney (2002) shows, for example, that the laws affected whites but not blacks."

equation model where the first stage correlation between the instrument and the endogenous variable is heterogeneous. We first discuss the inconsistency of the naive selective IV regression approach used in the applied literature and then propose various split-sample selective IV regression methods based on different first stage testing techniques. In Section 2 we discuss the situation where the instrument is strong for a nontrivial subpopulation of the data but irrelavant for the rest. We show that when the number of groups (for first stage selection) is fixed, the asymptotic lengths of IV confidence intervals follow a discrete distribution that, under certain conditions, stochastically dominates the constant asymptotic lengths of standard (full sample) confidence intervals. This means that split-sample selective IV regression methods can be more efficient than the standard IV regression. We then propose in Section 2.3 and 2.4 selective IV regression methods that generate deterministic asympotic length of confidence intervals. This is important because randomness in asymptotic lengths implied that the performance of split-sample methods relies on luck of the draw for which data being used for the first stage testing even at the limit with $N \to \infty$. Section 2.3 disucsses the asymptotic behavior of the split-sample IV regression with growing number of groups. Section 2.4 presents a repeated split-sample selective IV regression approach. We show that both the single split-sample with growing numbr of groups and the repeated split-sample methods can be more efficient than the standard IV regression methods with the latter out-performing the former.

Section 3 and 4 discuss the benchmark model with weak or semi-weak instrumental variables. There, the performance of the selective IV regression methods depends on the statistical size adopted in the multiple testing procedure for the selection of groups with non-zero first stage coefficients. We then propose an adaptive procedure that carries out the selective IV regression using an estimated optimal size in the first stage multiple testing. We show that once we allow the size to be adaptive and optimal, the asymptotic properties of our selective IV regression estimators become equivalent regardless of what kind of multiple

testing technique is used for first stage selection. Section 5 elaborates on the efficiency gain of the proposed methods with simulations. Section 6 applies the methods proposed in this study to the compulsory schooling data of Stephens and Yang (2014) and Oreopoulos (2006).

2 IV regression With Strong Instrument

In this section, we consider a simultaneous equation model with one endogenous covariate, W, and one instrument Z. The first stage correlation between W and Z is assumed to be heterogeneous, and for a non-trivial proportion of groups of individuals there is no correlation between W and Z.

For each individual i in group g, assume that

$$Y_{ig} = \beta W_{ig} + X_{ig}^{0\prime}\theta + u_{ig},$$
$$W_{ig} = \rho_g Z_{ig} + X_{ig}^{0\prime}\gamma + v_{ig},$$

where X_{ig}^0 is a $d \times 1$ random variable and W_{ig}, Z_{ig} are scalar random variables. The dataset is generally unbalanced. Let n_g denote the sample size in group g, G denote the total number of groups, and $N = \sum_{g=1}^{G} n_g$ denote the size of the full sample.

In this section we consider a benchmark two-type model with a strong instrumental variable.

Assumption 1 (Strong IV)**.** *Assume that*

$$\rho_g = \begin{cases} 0 & \textit{for some } G_0 \textit{ groups}\ , \\ a & \textit{for the rest } G_1 \textit{ groups}\ . \end{cases}$$

with $G_1 + G_0 = G$.

Let $\tilde{p} = G_1/G$ denote the proportion of *groups* where the instrument Z_{ig} has a non-zero

correlation with the endogenous covariate W_{ig}. Without loss of generality, assume that a is positive.

We consider asymptotic properties of our selective IV regression methods with $n_1, ..., n_G \to \infty$. The number of group G can be fixed or going to infinity. First we consider the case where G is a fixed positive integer. Let $p_g = \frac{n_g}{N}$ denote the sample proportion of data in group g and assume that it is fixed as $n_1, ..., n_G \to \infty$. Let $G_+ = \{g : \rho_g = 1\}$ denote the set of groups with non-zero first stage correlation and G_+^c its compliement. Let p be the sample proportion of *individuals* in groups with non-zero first stage correlation, $p = \sum_{g \in G_+} p_g$. It is also fixed as the sample size increases. Then the simultaneous equation defined above can be reduced to

$$Y_{ig} = \beta W_{ig} + X_{ig}^{0\prime}\theta + u_{ig},$$
$$W_{ig} = \rho Z_{ig} + X_{ig}^{0\prime}\gamma + \epsilon_{ig},$$

where $\rho = \sum_{g=1}^{G} \rho_g n_g = ap$, and $\epsilon_{ig} = v_{ig} + (\rho_g - \rho)Z_{ig}$.

The simultaneous equation model with heterogeneous first stage is a natural specification in a variety of economic applications. For example in the return to compulsory schooling literature, economists compile information from multiple natural experiments (e.g., state laws that shift minimum school dropping age) to create an instrument Z (e.g., the minimum school dropping age an individual faced at the age of 14). This instrument is used to estimate the effect of an endogeneous variable W (e.g. individual years of education) on the outcome Y (e.g., individual wage). Effective policies make the instrument correlated with the endogenous variable while ineffective policies undermine this correlation.

The following assumption states the regularity conditions we impose on our simultaneous equation model.

Assumption 2. *1. I.I.D. Data: the data* $\{Z_{ig}, X_{ig}^0, W_{ig}, Y_{ig}\}_{i=1}^N$ *are i.i.d.* $X_{ig} = (Z_{ig}\ X_{ig}^0)$.

2. *Finite Moments:* $E[X'_{ig}X_{ig}] = M$, $E[Z^2_{ig}] = k$, $E[Z^4_{ig}] = k'$. *M is positive definite, k and k' are bounded from above and away from zero.*

3. *Exclusion Restriction:* $E[u_{ig}|X_{ig}] = E[v_{ig}|X_{ig}] = 0$;

4. *Rank Condition:* $a \neq 0$, $\tilde{p} > 0$. *WLOG, assume that* $a > 0$.

5. *Homoskedasticity:* $E[u^2_{ig}|X_{ig}] = \sigma^2_u$, $E[v^2_{ig}|X_{ig}] = \sigma^2_v$,

6. *Fixed Group Proportion: WLOG, assume that* $0 < p_g < 1$ *is fixed as* $n_g \to \infty$, *for all* $g = 1, \ldots, G$.

Assumption 2.1-2.4 are standard assumptions for IV regression under strong IV. Assumption 2.5 is used to obtain closed-form solution of AR confidence intervals. It does not have impact on any consistency (or inconsistency) results stated in the article but simplifies the efficiency comparison between standard and proposed selective IV regression methods. Assumption 2.6 is not restrictive under the fixed G regime and is made to simplify notations used in the article. Without Assumption 2.6, p_g can be replaced with a deterministic sequence $p_{g,N}$, and p a deterministic sequence $p_N = \sum_{g \in G_+} p_{g,N}$. All results derived under Assumption 2.6 will still hold if one assumes $p_N \to p$ as $N \to \infty$. Assumption 2.6 is not relevant when the number of group G is allowed to increase with sample size and will be lifted up then.

Write the above model for a sample size N in the matrix form, we have

$$Y = \beta W + X^0\theta + u,$$
$$W = \rho Z + X^0\gamma + \epsilon,$$

where W, Y, u, ϵ are $N \times 1$ and X^0 is $N \times d$. The $i + \sum_{l=1}^{g-1} n_l$ elements of W, Y, u, ϵ and X^0 are respectively $W_{ig}, Y_{ig}, u_{ig}, \epsilon_{ig}$ and X^0_{ig}. Let $W_g = (W_{i'g} \ ... \ W_{i''g})'$ with $i' = 1 + \sum_{l=1}^{g-1} n_l$

and $i'' = \sum_{l=1}^{g} n_l$. Then $W = (W_1' \ ... \ W_G')'$. Then define Y_g, u_g and v_g correspondingly. Following Andrews and Stock (2005), assume WLOG that $Z_g' X_g^0 = 0$ for all $g = 1, ...G$.[3]

Because $Z'X^0 = \sum_{g=1}^{G} Z_g' X_g^0 = 0$, the 2SLS estimator reduces to

$$\hat{\beta} = (Z'W)^{-1} Z'Y.$$

Under Assumption 2, $Z'W/N$ converges to a non-zero constant in the limit and $\hat{\beta}$ is consistent. Asymptotic variance of $\hat{\beta}$ can be estimated by

$$\hat{AVAR}(\hat{\beta}) = \sigma_u^2 (Z'W)^{-1} (Z'Z) (Z'W)'^{-1} = \frac{\hat{\sigma}_u^2}{W' P_Z W},$$

where $P_Z = Z(Z'Z)^{-1}Z'$. Given σ_u, estimation precision of 2SLS depends on the value of the scalar $W' P_Z W$.

If the instrument is potentially weak the researcher wants to use inferential methods that are robust to both non-weak and weak instruments. In this paper, we look at the AR test, which is an optimal test given our just identified model set-up.

The AR test statistic for $H_0 : \beta = \beta_0$ is

$$\begin{aligned} AR(\beta_0) &= \frac{(Y - \beta_0 W)' P_Z (Y - \beta_0 W)}{(Y - \beta_0 W)' M_X (Y - \beta_0 W)/(N - d - 1)} \\ &\Rightarrow \chi^2(1) \ \textit{under the null,} \end{aligned}$$

where $M_X = I - X(X'X)^{-1}X'$. If the error term u_{ig} is additionally normally distributed, $AR(\beta_0) \sim F(1, N-d-1)$ under the null. In practice, critical values are typically calculated according to the $F(1, N-d-1)$ distribution. Let c be the $(1-\alpha)$ quantile of the $F(1, N-d-1)$ distribution, then c is the critical value for the AR test with significance level α. Additionally,

[3]Let $\tilde{Z}_g$ be the original instrument. Define $Z_g = \tilde{Z}_g - X_g^0 (X_g^{0'} X_g^0)^{-1} X_g^{0'} \tilde{Z}_g$. Then Z_g is independent of u_g, v_g and $Z_g' X_g^0 = 0$.

$\sqrt{c} = \mathsf{t}_{N-d-1,\alpha/2} \to \mathsf{Z}_{\alpha/2}$, where $\mathsf{t}_{N-d-1,\alpha/2}$ is the $(1-\alpha/2)$ quantile of Student-t distribution with degree of freedom $N-d-1$ and $\mathsf{Z}_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of standard normal distribution. Then the AR confidence interval is

$$CI^{AR} = \{\beta_0 : (Y - \beta_0 W)' H_X (Y - \beta_0 W) \leq 0\}$$

where $H_X = P_Z - \frac{c}{N-d-1} M_X$.

It is well known that the AR confidence interval has four different forms: 1) the real line, 2) $(-\infty, x_1) \cup (x_2, \infty)$, 3) (x_1, x_2) and 4) $\emptyset$. The confidence set is bounded if and only if $W' H_X W > 0$, while the length of the confidence interval equals to $2\sqrt{\left(\frac{Y' H_X W}{W' H_X W}\right)^2 - \frac{Y' H_X Y}{W' H_X W}}$ given boundedness.[4] In this paper, we define improvement of inference as reducing the length of classic 2SLS or the AR confidence interval when the IV is strong. When the IV is weak, improvement is defined as an increase in the probability of having a bounded confidence interval, or a reduction of the length of the interval provided it is bounded.

2.1 Full Sample, Infeasible and Naive Selective IV Regression Methods

When the instrumental variable is not weak, both the classic 2SLS and the AR method produce confidence intervals with correct coverage. Let $|CI^{2SLS}|$ denote the length of the 2SLS confidence interval; $|CI^{2SLS}| = 2\mathsf{t}_{N-d-1,\alpha/2}\sqrt{\frac{\hat{\sigma}_u^2}{W' P_Z W}}$, where $\hat{\sigma}_u$ is a consistent estimator of σ_u. Let $|CI^{AR}|$ denote the length of the AR confidence interval, $|CI^{AR}| = 2\sqrt{\left(\frac{Y' H_X W}{W' H_X W}\right)^2 - \frac{Y' H_X Y}{W' H_X W}}$ as is discussed earlier. Recall that $\mathsf{Z}_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. For the standard (full sample) IV regression, the following lemma holds.

[4] There is also a zero-measure event where the confidence set is bounded and empty, i.e. when $W' H_X W = Y' H_X W = 0$ and $Y' H_X Y > 0$.

Lemma 1. *Under Assumption 1 and 2, when $N \to \infty$, the probability that the AR confidence interval is bounded goes to one, and the lengths of the $(1-\alpha)\times 100$ percent 2SLS and AR confidence intervals for β are asymptotically equivalent with*

$$\sqrt{N}|CI^{2SLS}(CI^{AR})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{k}}.$$

The derivation is provided in the appendix. Since the correlation between the instrument and the endogenous variable differs among groups, it is natural to consider ways of improving the IV regression utilizing the first stage heterogeneity. First we consider the infeasible situation where the set G_+ is known and IV regression is performed only on individuals with $g \in G_+$. Compared with standard IV regression, the infeasible selective IV regression has stronger correlation between the instrument and the endogenous regressor but at the same time works with a smaller sample size. Let $CI^{2SLS,INFSEL}$ and $CI^{AR,INFSEL}$ be the infeasible 2SLS and AR confidence interval constructed using only data with $g \in G_+$, where the superscript $INFSEL$ is used to denote infeasible selection. It is easy to show that the length of a $(1-\alpha)\times 100$ percent 2SLS and AR confidence intervals for β are asymptotically equivalent with,

$$\sqrt{N}|CI^{2SLS,INFSEL}(CI^{AR,INFSEL})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{pk}}. \tag{1}$$

So the infeasible selective IV regression is always more efficient than the full sample IV regression as long as $p < 1$, or there is a non-trivial proportion of the individuals with zero first stage correlation. The efficiency improvement can be substantial when p is small.

Given this encouraging result, one might be tempted to do IV regression only using groups selected by the multiple test $H_{0,g} : \rho_g = 0$ against the alternative that $H_{a,g} : \rho_g > 0$, $g = 1, ..., G$. As we discussed in the introduction, applied researchers do use this strategy to get stronger correlation between the instrument and the endogenous covariate. Let t_g be the

t-statistic for group g and $c_{g,FS}$ the critical value for unadjusted pointwise t-test with significance level α_{FS}. The next lemma shows that such a naive selective IV regression procedure generates data samples that violate the exclusion restriction and is hence inconsistent.[5]

Lemma 2. *The naive selective IV regression approach violates the exclusion restriction (in the sense that* $E[Z_g' u_g | t_g > c_{g,FS}] \neq 0$*) given non-zero correlation between* u_{ig} *and* v_{ig}*. Moreover, if one further assumes that*

$$u_{ig} = \eta v_{ig} + e_{ig}$$

with $\eta \neq 0$ *and* $E[e_{ig}|v_{ig}, Z_{ig}] = 0$*, then with a* $1-(1-\alpha_{FS})^{G_0}$ *chance, the exclusion restriction is violated at the rate* $O_p\left(\frac{1}{\sqrt{N}}\right)$*.*

The proof is given in the appendix. Intuitively, the exclusion restriction is violated because the first stage selection is based on the value of $Z_g' v_g$ both for the G_+ groups and the G_+^c groups. Therefore, the expected value of $Z_g' v_g$ is not zero for the selected groups. When the simultaneous model is endogenous, u_{ig} and v_{ig} are correlated. The violation of the exclusion restriction is then clear.

Moreover, when researchers use the naive selective IV regression method, $\left(1-(1-\alpha_{FS})^{G_0}\right) \times 100$ percent of the times, the exclusion restriction is violated at rate $O_p\left(\frac{1}{\sqrt{N}}\right)$. This is because given G for the G_+ groups, naive selection brings a violation at rate $o_p\left(\frac{1}{\sqrt{N}}\right)$, or $E[Z_g' v_g | t_g > c_{g,FS}] = o_p\left(\frac{1}{\sqrt{n_g}}\right) = o_p\left(\frac{1}{\sqrt{N}}\right)$ but for the G_+^c groups, naive selection brings a violation at rate $O_p\left(\frac{1}{\sqrt{N}}\right)$. The chance of at least one G_+^c group being selected for the IV regression equals to $1-(1-\alpha_{FS})^{G_0}$. With $\alpha_{FS} = 5$ %, this probability is, for example, 0.226

[5] One could also choose to use multiple testing procedures that control for family-wise error rates or false discovery rates instead of pointwise error rate for the first stage selection. But the selected groups still contain those with large sample correlation between the instrument and the first stage error term hence violating the exclusion restriction.

when $G_0 = 5$ and 0.923 when $G_0 = 50$.

As shown by Guggenberger (2012)[6], when the exclusion restriction is violated at a rate as large as $O_p\left(\frac{1}{\sqrt{N}}\right)$, IV regression methods, including all weak instrument robust methods, are inconsistent with inflated asymptotic size. To avoid this near exogeneity situation, we propose consistent selective instrumental variable methods first by employing the split-sample strategy, a standard trick for gaurenteeing consistency in papers with pre-testing bias or similar problems (e.g. Dufour, 1997 and Abadie, Chingos, and West, 2014). We then propose a novel repeated split-sample IV regression approach and show that it is more efficient than the single split-sample IV regression approach for our simultaneous equation model.

2.2 Split-sample Selective IV Regression with Fixed Number of Groups

For some $q \in (0,1)$, let D_{ig} be a dummy variable that equals to 1 for the first $n_{g,D} = \lceil n_g q \rceil$ individuals in group g, where $\lceil n_g q \rceil$ is used to denote the smallest integer larger than $n_g q$. Then $D_{ig} = 0$ for the rest $n_{g,1-D} = \lfloor n_g(1-q) \rfloor$ individuals in group g. For each group, we use individuals with $D_{ig} = 0$ for multiple testing with null hypothesis $H_{0,g} : \rho_g = 0$, $g = 1, ..., G$. Let SEL_g be the decision rule that equals to one if the null $H_{0,g}$ is rejected and zero otherwise. Let the set $\hat{G}_+ = \{g : SEL_g = 1\}$ include all the groups that the null hypothesis is rejected. $\hat{G}_+ = \hat{G}_f \cup \hat{G}_t$, where $\hat{G}_t = \{g : g \in G_+, g \in \hat{G}_+\}$ are the correctly rejected groups and $\hat{G}_f = \{g : g \in G_+^c, g \in \hat{G}_+\}$ the false rejections. The split-sample selective IV regression approach that we are going to propose collects the multiple testing results from the first stage selection and use all individuals with $D_{ig} = 1$ and $g \in \hat{G}_+$ to conduct the IV regression with either 2SLS or AR test. For the first stage selection, we consider multiple testing procedures that control the unadjusted pointwise error rate (PWER), the familywise error

[6]Conley, Hansen, and Rossi (2012) and Berkowitz, Caner, and Fang (2012) also develop inference methods that are robust to weak instrument under near/plausible exogeneity.

rate (FWER) or the false discovery rates (FDR). For multiple testing controlling FWER we focus on the Bonferroni procedure, while for multiple testing controlling FDR (see van der Laan and Dudoit, 2007, for a review) we discuss two dominating approaches, the pioneer procedure proposed by Benjamini and Hochberg (1995) (BH) and an improvement proposed by Benjamini and Hochberg (2000) and Benjamini, Krieger, and Yekutieli (2006) (adaptive BH).

Let t_g be the one-sided Student-t test statistic of $H_{0,g} : \rho_g = 0$ constructed using all individuals in group g with $D_{ig} = 0$. And let p_g be the associated p-value for the one-sided t test and let $\mathsf{p}_{(1)} < \cdots < \mathsf{p}_{(G)}$ denote the ordered p-values and define $\mathsf{p}_{(0)} \equiv 0$. The four different multiple testing procedures are described in below.

1. If unadjusted pointwise error rate is controlled, $SEL_g = 1(t_g > \mathsf{t}_{\lfloor n_g(1-q)\rfloor - d - 1, \alpha_{FS,PW}})$ where $\alpha_{FS,PW}$ is the unadjusted size of the individual t-test.

2. If familywise error rate is controlled using the Bonferroni procedure, $SEL_g = 1(t_g > \mathsf{t}_{\lfloor n_g(1-q)\rfloor - d - 1, \frac{1}{G}\alpha_{FS,FW}})$ where $\alpha_{FS,FW}$ is the unadjusted size of the individual t-test.

3. If false discovery rate is controlled based on the BH procedure, $SEL_g = 1(\mathsf{p}_g \leq \mathsf{p}_{(\bar{g}_{BH})})$ where $\bar{g}_{BH} = \max\{0 \leq g \leq G : \mathsf{p}_{(g)} \leq \alpha_{FS,FDR}\frac{g}{G}\}$ where $\alpha_{FS,FDR}$ is the case discovery rate being controlled.

4. If false discovery rate is controlled based on the adaptive BH procedure, $SEL_g = 1(\mathsf{p}_g \leq \mathsf{p}_{(\bar{g}'_{BH})})$ with $\bar{g}'_{BH} = \max\{0 \leq g \leq G : \mathsf{p}_{(g)} \leq \frac{\alpha_{FS,FDR}}{1-\hat{p}}\frac{g}{G}\}$ where $\hat{p}$ is a consistent estimator of $\tilde{p}$.

The adaptive BH procedure is an improvement upon the original BH procedure because the original BH procedure targets the false discovery rate at $\alpha_{FS,FDR}$ but only effectively controls the rate $\alpha_{FS,FDR}(1 - \tilde{p})$ when $\tilde{p} > 0$. Therefore, the original BH procedure, in fact, gets increasingly conservative as the null proportion gets smaller. To improve, the

alternative proportion $\tilde{p}$ is first consistently estimated by $\hat{p}$, then the original BH procedure is applied using the adapted target $\frac{\alpha_{FS,FDR}}{1-\hat{p}}$. There are various proposals in the literature for estimating the null proportion, notably the Bayesian approach (Storey, 2002, 2003), the empirical Bayes approach (Efron, Tibshirani, Storey, and Tusher, 2001) and the empirical characteristic function approach (Cai and Jin, 2007).

The following lemma states the findings in Genovese and Wasserman (2002), who showed that, under very mild assumptions, both the BH procedure and the adaptive BH procedure are asymptotically equivalent to a thresholding p-value procedure, for which the cutoff value can be found by solving explicit functions described in Lemma 3. They also show that the adaptive BH procedure is optimal in the sense that it minimizes the false nondiscovery rate while controlling false discovery rate under pre-specified level.[7]

Lemma 3. *Denote the cumulative distribution function for p_g as $F(u) = (1-\tilde{p})u + \tilde{p}F_1(u)$ where $F_1(u)$ is the cumulative distribution of the p-values under the alternative and under the null, it follows a uniform distribution. We further denote the density function associated with $F_1(u)$ be $f_1(u)$.*

1. *Provided that $f_1(u)$ is monotonically decreasing in u and $f_1(0) > 1/u^*$,[8] the BH procedure controlling false discovery rate at level $\alpha_{FS,FDR}$ is equivalent to an unadjusted*

[7]The adaptive BH procedure is p-value based. Sun and Cai (2007) shows that it is better to use procedures based on likelihood ratio statistics for two-sided tests. Since for the benchmark two-type model we described, we perform multiple one-sided tests, p-value approach and likelihood ratio approach is equivalent. See Gu and Shen (2014) for the equivalence result for composite null case. The same conclusion holds for simple null hypothesis as we have here.

[8]This assumption is equivalent to Genovese and Wasserman (2002)'s original assumption on the concavity of $F(u)$. It is better to cast assumption on the density of the p-values directly, since we can estimate the sample density of the p-value, hence the assumption is easier to check (See Cao, Sun, and Kosorok, 2013). The assumption basically implies that u^* is the unique solution satisfies $u^*/((1-\tilde{p})u^* + \tilde{p}) = \alpha_{FS,FDR}$ and u^* is bounded away from zero.

pointwise testing procedure with cutoff p-value equal to

$$u^* = \frac{\tilde{p}\alpha_{FS,FDR}}{1 - \alpha_{FS,FDR}(1-\tilde{p})} \tag{2}$$

2. *Provided that $f_1(u)$ is monotonically decreasing in u and $f_1(0) > 1/c^*$,[9], the adaptive BH procedure controlling false discovery rate at level $\alpha_{FS,FDR}$ is equivalent to an unadjusted pointwise testing procedure with cutoff p-value equal to*

$$c^* = \min\left\{1, \frac{\tilde{p}\alpha_{FS,FDR}}{(1-\tilde{p})(1-\alpha_{FS,FDR})}\right\} \tag{3}$$

It is well-known that multiple testing controlling FWER with level $\alpha_{FS,FW}$ using the Bonforroni procedure is equivalent to pointwise testing controlling significance level $\alpha_{FS,FW}/G$. Lemma 3 shows that the two FDR controlling multiple testing methods we study are also asymptotically equivalent to pointwise testing methods controlling size at some other cut-off significance levels, i.e. u^* and c^*. It can be shown that $u^* \leq c^*$, hence the adaptive BH procedure uses a larger threshold than the BH approach and is more liberal. Let

$$\alpha_{FS} = \begin{cases} \alpha_{FS,PW} & \textit{if PWER is controlled} \\ \alpha_{FS,FW}/G & \textit{if FWER is controlled by the Bonforroni procedure} \\ u^* & \textit{if FDR is controlled by the BH procedure} \\ c^* & \textit{if FDR is controlled by the adaptive BH procedure} \end{cases}$$

summarizes the significance level or asymptotic equivalent significance levels.

Let $CI^{2SLS,SS}$ and $CI^{AR,SS}$ be the 2SLS and AR confidence intervals constructed using data with $D_{ig} \times SEL_g = 1$. The following lemma presents asympototic property of the

[9]Similar to the assumptions for the BH procedure, here the assumption is to secure that c^* is the unique solution that satisfies $(1-\tilde{p})c^*/((1-\tilde{p})c^* + \tilde{p}) = \alpha_{FS,FDR}$ and c^* is bounded away from zero.

split-sample selective IV regression methods.

Lemma 4. *Under Assumption 1, 2, when $n_g(1-q) \to \infty$ for all $g = 1, ..., G$, the split-sample IV regression based on pointwise, familywise, BH and adaptive BH first stage selection methods satisfy the following properties:*

1. *the probability that the split-sample selective AR confidence interval is bounded goes to one,*

2. *with fixed q the lengths of $(1-\alpha) \times 100$ percent split-sample selective 2SLS and AR confidence intervals for β satisfy the following property*

$$\sqrt{N}|CI^{2SLS,SS}(CI^{2SLS,AR})| \Rightarrow \mathsf{L},$$

where L is a discrete random variable with

$$P\left(\mathsf{L} = 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{kq}}\sqrt{p + \sum_{g \in G_+^c} p_g SEL_g}\right) = \alpha_{FS}^{\sum_{g \in G_+^c} SEL_g}(1-\alpha_{FS})^{G_0 - \sum_{g \in G_+^c} SEL_g}$$

Note that $\sum_{g \in G_+^c} p_g = 1 - p$, then if $q \to 1$ as $n_g \to \infty$ for all $g = 1, ..., G$,

$$2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{pk}} \leq \mathsf{L} \leq 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{k}},$$

or that the distribution of L stochastically dominates the (deterministic) asymptotic length of the full sample IV regression method and is stochastically dominated by the (deterministic) asymptotic length of the infeasible selective IV regression method.

The proof is straightforward. Note that with strong IV, Type II error of the t-tests goes to zero in the limit. But unless $\alpha_{FS} \to 0$ Type I error does not go to zero. The asymptotic lengths of the split-sample selective IV regression method therefore depends on how many

G_+^c groups are falsely selected. The calculation of the distribution of L then follows similar argument as in Lemma 1. If all groups have balanced size, or $n_g = n$ for all $g = 1, ..., G$, the distribution of L can be simplified with

$$P\left(\mathsf{L} = 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{kq}}\sqrt{p + \frac{j}{G}}\right) = \binom{G_0}{j}\alpha_{FS}^j(1-\alpha_{FS})^{G_0-j} \text{ for } j = 0, 1, ..., G_0.$$

To understand the intuition for the stochastic dominance relationship described in the lemma, we note that given $0 < q < 1$, the best case scenario for split-sample selective IV regression with fixed G is when no groups with zero first stage correlation is selected and the worse case scenario is when all groups with zero first stage correlation are (falsely) selected. When q approaches 1, in the the limit there is no loss of efficiency from splitting sample for selection and IV regression and the best case scenario mimics the infeasible selective IV method and has a $(1-\alpha_{FS})^{G_0}$ occurrence probability. On the other hand, the worse case scenario mimics the full sample selective IV method and has a $\alpha_{FS}^{G_0}$ chance of happening. The stochastic dominance relationship among the three IV methods is hence clear. An easy calculation shows that the smaller α_{FS}, the larger probability that the split-sample method achieves the infeasible case. For example, when α_{FS} equals to 1 % and $G_0 = 20$, then the probability of reaching the infeasible bound is in fact roughly 82 %.

The stochastic dominance relationship implies that the split-sample selective IV regression method improves the efficiency of 2SLS and AR regression given $q \to 1$, $n_g(1-q) \to \infty$ for all $g = 1, ..., G$. However, the fact that the asymptotic lengths of the resulted confidence intervals are stochastic is quite inattractive. Because stochastic asymptotic lengths mean that the lengths depend on the random seed drawn for sample splitting even when the sample size goes to infinity. Therefore, in the next two sections, we consider other consistent selective IV regression methods that generate deterministic asymptotic lengths of confidence intervals.

So far, we have been discussing the asymptotic property of the split-sample selective IV method using fixed and predetermined α_{FS}. But one can see that the asymptotic property of the split-sample selective IV method in fact improves if a smaller α_{FS} is used. And if one let α_{FS} approach zero as the sample size increases, the split-sample selective IV method approaches the infeasible IV method. This is because in this section we assume that the instrument is strong. Therefore all groups with non-null first stage correlation will be selected in the limit (i.e., power goes to 1), then a smaller α_{FS} means a smaller type I error probability which leads to improvement of the asymptotic property of the split-sample selective IV method. However, in practice, with a finite sample size, too small a α_{FS} will result in unsatisfactory power of testing, which means that only a small proportion of groups with non-null first stage effects will be selected and the finite sample performance of the split-sample selective IV method may be unsatisfactory. We will leave the problem as is for now. After we introduce two estimation methods that provide deterministic asymptotic lengths of IV confidence intervals, we will look at IV regression with a weak instrument. There we will discuss the tradeoff between the size and the power of the first stage selection and propose size-adaptive methods where the optimal size for first stage selection can be estimated.

2.3 Split-sample Selective IV Regression with Growing Number of Groups

One way to mitigate the influence of random seed used for split-sampling is to get finer groups and increase the total number of groups, or G. In this section, we discuss the asymptotic behavior of the split-sample method allowing G to grow as sample size. Note that when $G \to \infty$ as $n_g \to \infty$, for all $g = 1, ..., G$, the proportion of groups with non-zero first stage correlation, originally denoted using $\tilde{p} = G_1/G$, becomes a deterministic sequence with $\tilde{p}_G = G_1/G$. The sample proportion of individuals in G_+ groups, originally denoted using p

under the simplifying assumption in Assumption 2.6 also becomes a deterministic sequence with $p_{N,G} = \sum_{g \in G_+} \frac{n_g}{N}$.

Assumption 3. *Assume that both sequences converges with*

$$\tilde{p} = \lim_{G \to \infty} \tilde{p}_G, \qquad p = \lim_{N \to \infty, G \to \infty} p_{N,G}.$$

It is easy to see that this change does not affect the asymptotic behavior of standard (full sample) and the infeasible selective IV regression confidence intervals and that the naive selective IV regression is still inconsistent. Meanwhile, we will show that, under the new assumption that $G \to \infty$ the split-sample selective 2SLS and AR confidence intervals, unlike the fixed G case discussed in Section 2.2, will have deterministic asymptotic lengths.

Note that in this section under $G \to \infty$, we consider first stage selection method based on multiple testing controlling unadjusted pointwise rate, and two BH methods controlling FDR. Procedures based on FWER is excluded because unless we make an assumption about the rate of convergence between n_g and G, the first stage selection based on multiple testing controlling FWER no longer has power function going to one in the limit.

The next lemma summarizes the asymptotic property of split-sample selective IV methods under strong instrument Assumption 1. Intuitively, because G goes to infinity together with the sample size, the proportion of individuals falsely selected for IV regression is fixed in the limit by law of large number. Therefore, the split-sample selective IV methods now have confidence intervals with deterministic asymptotic lengths.

Lemma 5. *Under Assumption 1-3, when $n_g(1-q) \to \infty$ for all $g = 1, ..., G$ and $G \to \infty$, split-sample IV regression based on pointwise, BH and adaptive BH first stage selection methods satisfy the following properties:*

1. *the probability that the AR confidence interval is bounded goes to one;*

2. *with fixed q the lengths of $(1-\alpha)\times 100$ percent 2SLS and AR confidence intervals for β are asymptotically equivalent with,*

$$\sqrt{N}|CI^{2SLS,SS}(CI^{AR,SS})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{kq}}\sqrt{\alpha_{FS}(1-p)+p}.$$

The proof is provided in the appendix. Next we compare the lengths of the split-sample $(1-\alpha)\times 100$ percent confidence intervals with the standard (full sample) and infeasible selective IV confidence intervals. We have that

$$\begin{aligned}
\text{Full sample:} \quad & \sqrt{N}|CI^{2SLS}(CI^{AR})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{k}} \\
\text{Infeasible:} \quad & \sqrt{N}|CI^{2SLS,INFSEL}(CI^{AR,INFSEL})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{k}}\sqrt{p} \\
\text{Split-sample Selective:} \quad & \sqrt{N}|CI^{2SLS,SS}(CI^{AR,SS})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{k}}\sqrt{\frac{p+\alpha_{FS}(1-p)}{q}}
\end{aligned}$$

The following proposition then summarizes the efficiency improvement of split-sample IV regression methods when the number of group grows with the sample size.

Proposition 1. *Under Assumption 1-3, when $n_g(1-q)\to\infty$ for all $g=1,...,G$, $G\to\infty$, the split-sample selective IV regression method generates shorter 2SLS and AR confidence intervals than the full sample IV regression when $(1-\alpha_{FS})p < q-\alpha_{FS}$.*

Further, if $q\to 1$ then the split-sample selective IV regression methods based on unadjusted pointwise testing and the BH method are always more efficient than the standard methods as long as $p<1$. Meanwhile, the split-sample selective IV regression method based on the adaptive BH method is more efficient if $\tilde{p} < 1-\alpha_{FS,FDR}$.

If in addition we let the controlled size in first stage selection α_{FS} go to zero as sample size goes to infinity, then lengths of all split-sample selective IV regression methods are asymptotically equivalent to that of the infeasible selective IV regression.

Results in the proposition are very intuitive. If q approaches 1, then in the limit there is no efficiency loss from reduced sample size used for IV regression. Proposed selective IV regression methods improve inference only when the proportion of effective policies increases after first stage selection. When unadjusted pointwise tests are used, improvement is always made if $p < 1$. When false discovery controlling tests are used, improvement takes place when $1-\tilde{p}$ is larger than the controlled false discovery rate. As is discussed in Lemma 3, the BH procedure actually controls FDR at $\alpha_{FS,FDR}(1-\tilde{p})$ while the adaptive BH procedure controls FDR exactly at level $\alpha_{FS,FDR}$. This explains why the split-sample selective IV regression method based on the BH procedure can improve efficiency as long as $p < 1$ while that based on the adaptive BH procedure only improves efficiency if $\tilde{p} < 1 - \alpha_{FS,FDR}$.

For all three split-sample selective IV regression methods discussed above, the asymptotic length of the confidence interval decreases with α_{FS}. As is already discussed in Section 2.2 after Lemma 4, this asymptotic property does not mean that in finite sample applications, researchers shall use extremely small α_{FS} values. A small α_{FS} hurts, in finite sample, the power of the first stage testing and hence the finite sample performance the selective IV regression procedures. The tradeoff between the size and the power in first stage testing will be discussed in the next section where we allow our instrument to be weak and our tests on first stage effect to face local alternatives.

2.4 Repeated Split-sample Selective IV Regression

In the last section, we consider the asymptotic property of split-sample selective IV regression when the number of group G grows with sample size. Intuitively, this is suggesting researchers to get finer groups when they have a larger dataset. This approach has two advantages, asymptotically, over the split-sample selective IV regression method with fixed G discussed in Section 2.2. First, with larger G, the asymptotic lengths of IV confidence intervals depend less on the random seed drawn in the sample-splitting stage. When $G \to \infty$ as $n_g \to \infty$

for all $g = 1, ..., G$, the asymptotic lengths of IV confidence intervals are deterministic and shorter than the asymptotic lengths of the standard (full sample) IV confidence intervals as long as $p < 1$. Second, with larger G, the model for first stage is more robust as all the finer groups are allowed to have different levels of first stage heterogeneity. As an example, one can think of the researcher first divides a random sample of U.S. residents by geographic regions, assuming first stage heterogeneity across regions. Researcher can also divide the random sample by states and even individual characteristics (e.g. gender, race) when he/she gets more data, that is to allow group number G grow with sample size N. This new method allows first stage heterogeneity across states and individual characteristics and is therefore more robust.

However, the strategy of getting finer groups with larger dataset may not be efficient if researchers know that the first stage correlation is homogeneous among, say, all states within the same geographic region. In this section, we propose a selective IV regression method that utilizes researchers' prior information that the first stage correlation is homogeneous among certain large groups. For example, sometimes it may be reasonable to assume that a state-level policy has the same effect on all residents in that state. In that sense, G is again fixed in this subsection.

Now, divide the dataset in each group g randomly into R disjoint subsamples. Let $M_{ig} = r$ if individual i in group g belongs to subsample r. Use superscript r to denote which subsample an individual belongs to. For each group g and subsample r, let t_g^r be the t-statistic of the hypothesis test for no first stage correlation using data belonging to that group and subsample. Since the first stage effect is only indexed by g, but not by r, we have imposed the homogeneity assumption among all R divisions for a particular group g to have the same first stage effect ρ_g. As we will see, this set of restrictions provides the repeated split sample method with extra efficiency. As is in Section 2.2, the first stage selection can be performed by unadjusted pointwise testing, familywise testing, or false discovery rate

testing. Further, define a weighting variable w_{ig} of all individuals of group g as

$$w_{ig} = \frac{1}{R-1}\sum_{r=1}^{R} SEL_g^r 1(M_{ig} \neq r),$$

where SEL_g^r are analogously defined as SEL_g in Section 2.2, except the tests are now performed for all subsamples divided by group and subsample configuration. By construction, w_{ig} is independent of $(X_{ig}, Y_{ig}, u_{ig}, v_{ig})$. This is a novel weighting strategy, with which we do not have efficiency loss from throwing away data as we utilize the whole sample for IV regression. The independence construction between the weights and the data also avoids the local violation of exclusion restriction, yet improves the IV regression efficiency by giving zero-signal groups less weights. Lemma 6 shows that the weight $w_{ig} \xrightarrow{p} 1(g \in G_+) + \alpha_{FS} 1(g \in G_+^c)$ if $R \to \infty$ and $n_g/R \to \infty$ for all $i = 1, ..., N$ and $g = 1, ..., G$. Definition for α_{FS} is the same as Section 2.2.

Let $\hat{Q}$ be a $N \times N$ diagonal weighting matrix with each diagonal element taking quantity w_{ig}. In this subsection, assume WLOG that $Z'\hat{Q}X^0 = 0$.[10] The repeated split-sample 2SLS estimator is then

$$\hat{\beta}^{2SLS,RSS} = (Z'\hat{Q}W)^{-1}Z'\hat{Q}Y.$$

The repeated split-sample AR confidence interval is

$$CI^{AR,RSS} = \{\beta_0 : (Y - \beta_0 W)' H_X^{\hat{Q}} (Y - \beta_0 W) \leq 0\}$$

with $H_X^{\hat{Q}} = P_Z^{\hat{Q}} - \frac{\tilde{c}}{trace(\hat{Q})-d-1} M_X^{\hat{Q}}$ where $P_Z^{\hat{Q}} = \hat{Q}Z(Z'\hat{Q}\hat{Q}Z)^{-1}Z'\hat{Q}$, $M_X^{\hat{Q}} = \hat{Q} - \hat{Q}X(X'\hat{Q}X)^{-1}X'\hat{Q}$ and $\tilde{c}$ is the $(1-\alpha)$ quantile of the $F_{1,trace(\hat{Q})-d-1}$ distribution. Note that $P_Z^{\hat{Q}} Z \neq Z$ but

[10] Let $\tilde{Z}$ be the original instrument. Define $Z = \tilde{Z} - X^0(X^{0'}\hat{Q}X^0)^{-1}X^{0'}\hat{Q}\tilde{Z}$. Then Z is mean independent of u, v and $Z'\hat{Q}X^0 = 0$.

$M_X^{\hat{Q}} X = 0$. The interval is bounded if $W' H_X^{\hat{Q}} W \geq 0$. Given boundedness, its length equals to $2\sqrt{\left(\frac{Y' H_X^{\hat{Q}} W}{W' H_X^{\hat{Q}} W}\right)^2 - \frac{Y' H_X^{\hat{Q}} Y}{W' H_X^{\hat{Q}} W}}$.

The following Lemma shows the asymptotic property of repeated split-sample IV regression methods under strong instrument assumption.

Lemma 6. *Under Assumption 1 and 2, when $R \to \infty$ and $n_g/R \to \infty$ for all $g = 1, ..., G$,*

$$w_{ig} \xrightarrow{p} 1(g \in G_+) + \alpha_{FS} 1(g \in G_+^c).$$

Moreover, the repeated split-sample IV regression based on pointwise, BH and adaptive BH first stage selection methods satisfy the following properties:

1. *the probability that the AR confidence interval is bounded goes to one, and*
2. *the lengths of the $(1-\alpha) \times 100$ percent repeated split-sample selective 2SLS and AR confidence intervals for β are asymptotically equivalent with*

$$\sqrt{N} |CI^{2SLS,RSS}(CI^{AR,RSS})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2} \frac{\sigma_u}{ap\sqrt{k}} \times \sqrt{1 - (1-p)(1-\alpha_{FS}^2)}.$$

The proof is provided in the appendix. The lemma shows that the repeated split-sample IV regression methods are always more efficient than the full sample IV regression methods and less efficient than the infeasible selective IV regression methods as long as $0 < \alpha_{FS} < 1$ and $0 < p < 1$. If $\alpha_{FS} \to 0$, the repeated split-sample IV regression methods approaches the infeasible selective IV regression methods. Given α_{FS}, the improvement in efficiency is larger if p is smaller.

The repeated split-sample selective IV regression method is also more efficient than the single split-sample with growing G as long as $0 < p < 1$ and $0 < \alpha_{FS} < 1$. The argument is made through comparing the asymptotic lengths of confidence intervals while driving $q \to 1$

for the single split-sample method. Therefore, the improvement in efficiency lies in two dimensions. First, the repeated split-sample method do not have efficiency loss from losing $1-q$ proportion of data in IV regression. Second, the repeated split-sample method is more efficient than the single split-sample IV regression method even with $q \to 1$ as $n \to \infty$ for the latter approach. This may be due to the fact that different subsamples within the same coarse group are assumed to be homogeneous while the single split-sample IV regression allows heterogeneity among all finely defined groups.

3 IV Regression with Weak Instrument

3.1 Full Sample, Infeasible and Naive Selective IV Regression Methods

In this section, we discuss the same two-type model as is set up in Section 2 except now we assume that the IV is weak. This is of interest because under weak IV, the 2SLS estimator inference is no longer valid, yet the robust AR inference may produce uninformative result. We aim to discuss in this section under what circumstances the proposed method increases the probability of having information AR confidence intervals.

Assumption 4 (Weak IV)**.** *Assume that*

$$\rho_g = \begin{cases} 0 & \text{for some } G_0 \text{ groups}, \\ a/\sqrt{N} & \text{for the rest } G_1 \text{ groups}. \end{cases}$$

with $G_1 + G_0 = G$.

Under assumption 4, the classic inference for the 2SLS estimator fails because the estimators may be better approximated by a Cauchy distribution as the first stage correlation goes to zero. Staiger and Stock (1997) is the first to formalize the weak instrument problem

of 2SLS under the assumption that the first stage correlation goes to zero at rate $1/\sqrt{N}$. AR testing is still consistent under weak IV and generates confidence intervals with correct size. In this section we discuss the asymptotic properties of the full sample, the infeasible and the naive selective AR confidence intervals. These results are parallel to those in Section 2. Compared to the analysis in Section 2.1, the asymptotic power for tests for first stage selection no longer goes to one under weak IV. That is, some of the groups with non-zero first stage correlations fail to be selected for IV regression. On the other hand, the type I errors of first stage selection do not change, that is some of the groups with zero correlations can still be falsely selected out. Therefore, the naive selective AR method again violates the exclusion restriction. But because both type I error and type II error of first stage selection is bounded between zero and one, the exclusion restriction is violated with a $O_p(\frac{1}{\sqrt{N}})$ rate with probability 1, meaning that the naive selective IV regression method always have inflated size for hypotheses testing even in the limit.

The next lemma summarizes the asymptotic property of the full sample and the infeasible and naive selective IV regression methods.

Lemma 7. *Under Assumption 2 and 4, when $N \to \infty$,*

1. *the probability that the full sample AR confidence interval is bounded goes to*

$$\mathsf{P}^{AR} = 2\Phi\left(\frac{ap\sqrt{k} - \sqrt{c(\sigma_v^2 + a^2kp(1-p))}}{\sqrt{\sigma_v^2 + a^2p(1-p)k'/k}}\right);$$

2. *the probability that the infeasible selective AR confidence interval is bounded goes to*

$$\mathsf{P}^{AR,INFSEL} = 2\Phi\left(a\sqrt{pk}/\sigma_v - \sqrt{c}\right);$$

3. *the naive selective AR method violates the exclusion restriction at rate $O_p\left(\frac{1}{\sqrt{N}}\right)$.*

The infeasible selective AR confidence interval is more efficient in the sense that $\mathsf{P}^{AR,INFSEL} > \mathsf{P}^{AR}$ *as long as* $0 < p < 1$.

The proof is given in the appendix. The efficiency comparison at the end of the lemma is based on the fact that $k'/k = E[Z_{ig}^4]/k > E[Z_{ig}^2]^2/k = k$.

3.2 The Split-sample Selective IV Regression Method with Fixed Number of Groups

Next we discuss the asymptotic performance of split-sample selective AR test under fixed G. We are interested in the random variable $\mathsf{P}^{AR,SS}$ and derive its discrete probability distribution.

Lemma 8. *Under Assumption 2 and 4, when* $q \to 1$ *and* $n_g(1-q) \to \infty$ *for all* $g = 1, ..., G$, *the asymptotic probability that the* $(1-\alpha) \times 100$ *percent AR confidence intervals for* β *is bounded is a discrete random variable, denoted as* $\mathsf{P}^{AR,SS}$. *When* $\sum_g p_g SEL_g \neq 0$

$$P\left(\mathsf{P}^{AR,SS} = 2\Phi\left(\frac{\frac{a\sqrt{k}\sum_{g\in G_+} p_g SEL_g}{\sum_g p_g SEL_g} - \sqrt{c\left(\sigma_v^2 + a^2 k \frac{\left(\sum_{g\in G_+} p_g SEL_g\right)\left(\sum_{g\in G_+^c} p_g SEL_g\right)}{\left(\sum_g p_g SEL_g\right)^2}\right)}}{\sqrt{\sigma_v^2 + a^2 \frac{k'}{k} \frac{\left(\sum_{g\in G_+} p_g SEL_g\right)\left(\sum_{g\in G_+^c} p_g SEL_g\right)}{\left(\sum_g p_g SEL_g\right)^2}}}\right)\right)$$
$$= \alpha_{FS}^{\sum_{g\in G_+^c} SEL_g} (1-\alpha_{FS})^{G_0 - \sum_{g\in G_+^c} SEL_g} \Pi_{g\in G_+} S_{g,FS}^{SEL_g} (1 - S_{g,FS})^{1-SEL_g},$$

where $S_{g,FS} = \Phi\left(a\frac{\sqrt{k}}{\sigma_v}\sqrt{\frac{n_g}{N}} - \mathsf{Z}_{\alpha_{FS}}\right)$ *is the power of the first stage t-test for* ρ_g *given unadjusted size* α_{FS}. *When* $\sum_g p_g SEL_g = 0$, *no groups are selected for IV regression and*

$$P(\sum_g p_g SEL_g = 0) = (1-\alpha_{g,FS})^{G_0} \times \Pi_{g\in G_+} (1 - S_{g,FS}).$$

Note that $\sum_{g\in G_+} p_g = p$ and $\sum_{g\in G_+^c} p_g = 1-p$, then we have the following special cases

that

$$P\left(\mathsf{P}^{AR,SS}=\mathsf{P}^{AR}\right)=\alpha_{g,FS}^{G_0}\times\Pi_{g\in G_+}S_{g,FS},$$
$$P\left(\mathsf{P}^{AR,SS}=\mathsf{P}^{AR,INFSEL}\right)=\left(1-\alpha_{g,FS}\right)^{G_0}\times\Pi_{g\in G_+}S_{g,FS},$$
$$P\left(\mathsf{P}^{AR,SS}=\alpha\right)=\left(1-\left(1-\alpha_{g,FS}\right)^{G_0}\right)\times\Pi_{g\in G_+}\left(1-S_{g,FS}\right).$$

The proof is straightforward, one only need to substitute different values of ρ and different proportions of selected data, two factors depending on which G_+ and G_+^c groups are selected and the corresponding group size. If all groups have balanced size, or $n_g=n$ for all $g=1,...,G$, the distribution of $\mathsf{P}^{AR,SS}$ follows that

$$P\left(\mathsf{P}^{AR,SS}=\Phi\left(\frac{\frac{aj_1\sqrt{k}}{j_1+j_0}-\sqrt{c(\sigma_v^2+a^2k\frac{j_1j_0}{(j_1+j_0)^2})}}{\sqrt{\sigma_v^2+a^2\frac{k'}{k}\frac{j_1j_0}{(j_1+j_0)^2}}}\right)\right)=\binom{G_0}{j_0}\binom{G_1}{j_1}\alpha_{FS}^{j_0}(1-\alpha_{FS})^{G_0-j_0}S_{FS}^{j_1}(1-S_{FS})^{G_1-j_1}$$

where $j_0=1,...,G_0$ and $j_1=1,...,G_1$ and $S_{FS}=\Phi\left(a\frac{\sqrt{k}}{\sigma_v}\sqrt{\frac{1}{G}}-\mathsf{Z}_{\alpha_{FS}}\right)$. When $j_0=j_1=0$, no group is selected for AR test, this event occurs with probability $(1-\alpha_{FS})^{G_0}(1-S_{FS})^{G_1}$.

Lemma 8 shows that, unlike the result for the strong IV case in Lemma 4, the split-sample selective IV regression method no longer has guaranteed efficiency improvement over the standard (full sample) IV regression method. As discussed before, this is because, under the weak IV assumption given in Assumption 4, the probability that all G_+ groups be selected no longer goes to one in the limit. Of course, if the signal of the weak instrument is strong enough, or a large enough, the probability that the split-sample selective AR test be more efficient than the standard AR test can still be very close to one if S_{FS} is close to 1.

4 IV Regression With Semi-weak Instruments

In this section, we extend the above single split-sample IV regression method with weak IV so as to generate confidence intervals with deterministic asymptotic lengths. Parallel to the discussion in Section 2 we consider single split-sample selective IV regression with the number of group G grows together with the sample size N and the repeated split-sample IV regression method with the number of repetition R grows with the sample size N. In both cases, the growing G and R results in uninformative first stage selection with alternative signals weaker than the local $O(\frac{1}{\sqrt{n_g}})$ rate. Therefore, we consider in this section semi-weak IV assumptions to ensure that the first stage t-tests face local alternatives.

4.1 The Split-sample Selective IV Regression with Growing Number of Groups

In this section, we consider the split-sample selective IV regression method with $G \to \infty$ as $(n_1, ..., n_G) \to \infty$. Assume the first stage correlation bewteen the IV and the endogenous variable satisfy the following assumption.

Assumption 5 (Semi-Weak IV: SS)**.** *Assume that*

$$\rho_g = \begin{cases} 0 & \text{for some } G_0 \text{ groups}, \\ a/\sqrt{\frac{N}{G}} & \text{for the rest } G_1 \text{ groups}. \end{cases}$$

with $G_1 + G_0 = G$.

As $G \to \infty$, the strength of the instrument is stronger than the $1/\sqrt{N}$ rate assumed in Assumption 4. Under Assumption 5, both 2SLS and AR are consistent with the classic asymptotics with rate of convergence $\sqrt{G}$. First we summarize the asymptotic property of the full sample, the infeasible and naive selective IV regression methods under the semi-weak

IV assumption.

Lemma 9. *Under Assumption 2.1-2.5, 3 and 5, when $G \to \infty$ and $n_g \to \infty$ for all $g = 1, ..., G$, standard (full sample) and infeasible selective AR confidence intervals are both bounded with probability one in the limit. The asymptotic lengths of the $(1-\alpha)\times 100$ percent standard (full sample) and infeasible selective 2SLS and AR confidence intervals for β satisfy the properties defined in Lemma 1 and equation (1) except with the rate of convergence (originally $\sqrt{N}$) replaced by $\sqrt{G}$.*

Notice that $\rho_g = a/\sqrt{\frac{N}{G}}$ for all $g \in G_+$. The proof of Lemma **??** is exactly identical to those of Lemma 1 and equation 1 except with a replaced by $a/\sqrt{\frac{N}{G}}$. Next we summarize the asymptotic property of the split-sample selective IV regression method.

Lemma 10. *Under Assumption 2.1-2.5, 3 and 5, when $G \to \infty$ and $n_g \to \infty$ for all $g = 1, ..., G$, the split-sample selective IV regression based on pointwise, BH and adaptive BH first stage selection methods satisfy the following properties:*

1. *the probability that the AR confidence interval is bounded goes to one,*

2. *the lengths of the $(1-\alpha)\times 100$ percent repeated split-sample selective 2SLS and the AR confidence intervals are asymptotically the same with*

$$\sqrt{G}|CI^{2SLS,SS}(CI^{AR,SS})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{kq}S_{FS}} \times \sqrt{pS_{FS} + (1-p)\alpha_{FS}},$$

where $S_{FS} = \frac{1}{p}\lim_{N\to\infty,G\to\infty}\sum_{g\in G_+}\frac{n_g}{N}S_{g,FS}$ with $S_{g,FS} = \Phi\left(\frac{\sqrt{k}}{\sigma_v}\sqrt{\frac{n_g G}{N}}a - \mathsf{Z}_{\alpha_{FS}}\right)$ is the power of the first stage t-test given unadjusted size α_{FS}. If all groups have balanced sample size, or $n_g = n$, then $S_{FS} = \Phi\left(\frac{\sqrt{k}}{\sigma_v}a - \mathsf{Z}_{\alpha_{FS}}\right)$.

The proof of Lemma 10 is exactly the same as that for Lemma 5 but with a replaced by $a/\sqrt{\frac{N}{G}}$ and the proportion of data selected for IV regression replaced by $pS_{FS} + (1-p)\alpha_{FS}$ because the power of first stage t-test no longer goes to one in the limit.

Now we can compare the asymptotic lengths of the $(1-\alpha)\times 100$ percent IV confidence intervals generated by standard (full sample), selective and single split-sample selective IV regression. The next proposition summarizes the comparison.

Proposition 2. *Under Assumption 2.1-2.4 and 5, when $G \to \infty$ and $n_g \to \infty$ for all $g = 1, ..., G$, for fixed q, the split-sample selective IV regression method is more efficient than the standard (full sampel) method if $(S_{FS} - \alpha_{FS})p < qS_{FS}^2 - \alpha_{FS}$. When $q \to 1$, the split-sample selective IV regression method is more efficient if $(S_{FS} - \alpha_{FS})p < S_{FS}^2 - \alpha_{FS}$.*

Unlike the strong IV case discussed in Proposition 1, where a smaller α_{FS} always renders a shorter confidence interval for split-sample IV regression, here the conclusion is inconclusive as a smaller α_{FS} results in not only a smaller proportion of falsely rejected null groups but also a smaller number of non-discovered alternative groups. In fact, given p, an optimal α_{FS} can be solved so as to minimize the length of an AR confidence interval (given boundedness). In Section 4.3, we discuss such optimization idea in detail and compare the asymptotic property of AR confidence interval under optimally selected α_{FS}.

4.2 The Repeated Split-sample Selective IV Regression Method

In this section, we consider the repeated split-sample selective IV regression method introduced in Section 2.4. Similar to the discussion in last section, the repeated method is not applicable in the weak IV case under Assumption 4 because first stage t-tests are conducted for each subsample divided by g and subsample r. Under weak IV, each of these test has power equals to size. Therefore, we consider in this section the following semi-weak IV assumption.

Assumption 6 (Semi-Weak IV - RSS)**.** *Assume that*

$$\rho_g = \begin{cases} 0 & \text{for some } G_0 \text{ groups}, \\ a/\sqrt{\frac{N}{R}} & \text{for the rest } G_1 \text{ groups}. \end{cases}$$

with $G_1 + G_0 = G$.

As $R \to \infty$, the strength of the instrument is stronger than the $1/\sqrt{N}$ rate assumed in Assumption 4. Under Assumption 6, both 2SLS and AR are consistent with the classic asymptotics but rate of convergence becomes $\sqrt{R}$.As $R \to \infty$, the strength of the instrument is stronger than the $1/\sqrt{N}$ rate assumed in Assumption 4. Under Assumption 6, both 2SLS and AR are consistent with the classic asymptotics but rate of convergence becomes $\sqrt{R}$.[11]

Lemma 11. *Under Assumption 2 and 6, when* $R \to \infty$ *and* $n_g/R \to \infty$ *for all* $g = 1, ..., G$,

$$w_{ig} \xrightarrow{p} S_{g,FS} 1(g \in G_+) + \alpha_{FS} 1(g \in G_+^c).$$

Moreover, the repeated split-sample selective IV regression based on pointwise, familywise, BH and adaptive BH first stage selection methods satisfy the following properties:

1. *the probability that the AR confidence interval is bounded goes to one,*

2. *the lengths of the* $(1-\alpha) \times 100$ *percent repeated split-sample selective 2SLS and the*

[11]Similar to Lemma 9, under Assumption 2 and 6, when $n_g \to \infty$ for all $g = 1, ..., G$ and $R \to \infty$, the standard (full sample), infeasible and single split-sample selective AR confidence intervals are all bounded in the limit with probability one. The asymptotic lengths of the $(1-\alpha)\times 100$ percent the standard (full sample), infeasible and split-sample selective 2SLS and AR confidence intervals for β satisfy the properties defined in Lemma 1, equation (1), and Lemma 4 except with the rate of convergence (originally $\sqrt{N}$) replaced by $\sqrt{R}$.

AR confidence intervals are asymptotically equivalent with

$$\sqrt{R}|CI^{2SLS,RSS}(CI^{AR,RSS})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{k}\sum_{g\in G_+} p_g S_{g,FS}} \times \sqrt{\sum_{g\in G_+} p_g S_{g,FS}^2 + (1-p)\alpha_{FS}^2}.$$

with $S_{g,FS} = \Phi\left(a\frac{\sqrt{k}}{\sigma_v}\sqrt{\frac{n_g}{N}} - \mathsf{Z}_{\alpha_{FS}}\right)$ *being the power of the first stage t-test for* ρ_g *given unadjusted size* α_{FS}*. If the all groups have balanced sample size, or* $n_g = n$*, then* $S_{FS} = \Phi\left(\frac{\sqrt{k}}{\sigma_v}\frac{a}{\sqrt{G}} - \mathsf{Z}_{\alpha_{FS}}\right)$ *and the limit simplifies to*

$$\sqrt{R}|CI^{2SLS,RSS}(CI^{AR,RSS})| \xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{k}pS_{FS}} \times \sqrt{pS_{FS}^2 + (1-p)\alpha_{FS}^2}.$$

As power S_{FS} *approaches 1, the lengths of the confidence interval approach those in Lemma 6 except with a different normalizing constant.*

The proof is given in the appendix.

4.3 Optimal Adaptive Procedure for Balanced Groups

The take away message from the above discussions is that under semi-weak instrument, the asymptotic lengths of IV confidence intervals are deterministic and that there is a competition between size and power in the first stage selection. When we allow the multiple testing procedure used for selection to be more liberal, we have higher power of detecting true non-null cases at the price of making more Type I errors. It is then natural to look for the optimal size and power combination to get the shortest possible confidence interval length. As we will see, this optimal size and power depends on the signal strength $a\sqrt{k}/\sigma_v$ as well as the non-null proportion p. WLOG, let's assume $\sigma_v = 1$, for a given pair of $(a\sqrt{k}, p)$ and fixed q, to minimize the confidence interval length using single split sample IV regressive with the

point wise testing procedure, we are to solve the following optimization problem,

$$\min_{\alpha} \frac{pS_{FS} + \alpha(1-p)}{qS_{FS}^2},$$

with $S_{FS} = \Phi(a\sqrt{k} - \mathsf{Z}_\alpha)$. The first order condition is a complicated nonlinear equation, but it can be solved easily numerically. Figure 1 gives some intuition. For a given signal strength, for higher non-null proportion p, we would like to be more liberal of allowing more Type I error to be able to maintain the power of detecting the true non-null cases. For a given p, the stronger the signal is, the easier it is to reject the non-null case. Thus we can put a more stringent level of the test. In the extreme case, when $p = 1$, the optimal level is one. We would like to reject all cases because indeed they are all non-zero.

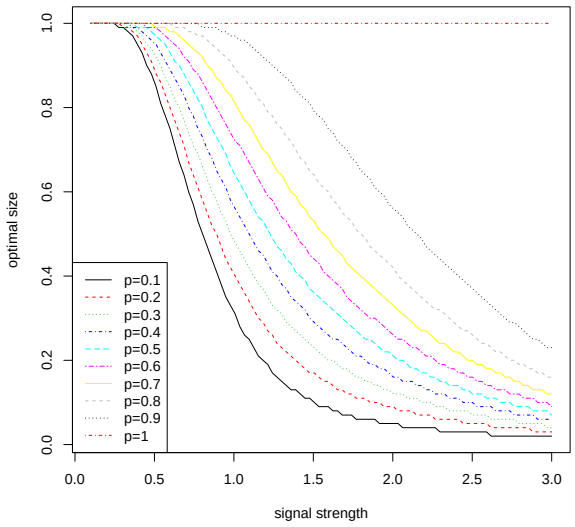

Figure 1: Optimal size to achieve the shortest confidence interval for point wise procedure. The curves correspond to the relationship between the size and the signal strength $a\sqrt{k}$ for a given non-null proportion p. Fix $q = 1$.

Similar optimization problem can be solved for the FDR procedures for the optimal α_{FDR} for a given pair of $(a\sqrt{k}, p)$. Not surprisingly, the best-achieved confidence interval length from all four testing procedures (i.e. unadjusted pointwise testing, familywise by Bonferroni,

and false discovery rate testing by BH and adaptive BH) are identical for a given pair of $(a\sqrt{k}, p)$. After all, all three procedures are to find the best cutoff for p values. Once we allow for the size (either point wise size or FDR size) to be determined adaptively to the specific problem, they all achieve the same optimal balance between size and power.

The next question is how far away these optimal adaptive confidence interval length compares to the infeasible Oracle confidence interval. Figure 2 illustrates the magnitude of the gap between the infeasible confidence interval length and the best achievable adaptive confidence interval length. The gap gets smaller as signal strength gets stronger. Or for given signal strength, the gap decreases as non-null proportion increases. It is intuitive to understand this result because the testing problem becomes easier as signal strength or the proportion of signals gets larger.

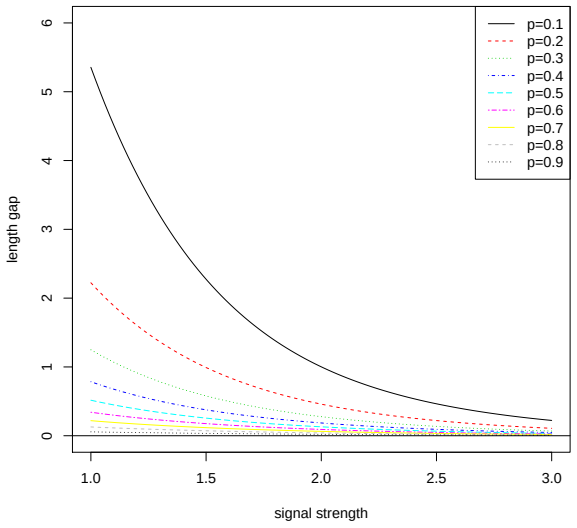

Figure 2: The difference of confidence interval length between the optimal that can be achieved by the adaptive procedure and the infeasible procedure. The curves traces the magnitude of the difference as signal strength $a\sqrt{k}$ increases for a given non-null proportion p. Fix $q = 1$

In practice, we need to know p, a, k as well as σ_v^2 to be able to choose the optimal size. k and σ_v^2 as the variance of the instrument and the first stage error term is easy to estimate

consistently. The proportion of non-null hypothesis can be estimated consistently by Cai and Jin (2007) under the current two type model of null and local alternative. Then we can back out a consistent estimator of a using the estimator of p from the pooled first stage OLS regression. Need some simulation support to see how the adaptive procedure works.

For repeated split sample IV regression method, the optimization problem to minimize the confidence intervals produced becomes

$$\min_{\alpha} \ \frac{pS_{FS}^2 + \alpha^2(1-p)}{S_{FS}^2},$$

with $S_{FS} = \Phi(a\sqrt{k} - \mathsf{Z}_\alpha)$. It can be shown that the objective function is monotonically increasing in α, hence the optimal solution is to choose α smallest possible. This implies that with repeated split sample method, being conservative in size does not hinder one from enhancing efficiency. This seems a bit counterintuitive because choosing small α hurts power. Let's take a closer look at the confidence interval.

We have from Lemma 11 under balanced group setting,

$$\begin{aligned}
\sqrt{R}|CI^{2SLS,RSS}(CI^{AR,RSS})| &\xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{k}pS_{FS}} \times \sqrt{pS_{FS}^2 + (1-p)\alpha_{FS}^2} \\
&= 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{k}\sqrt{p}} \times \sqrt{\frac{pS_{FS}^2 + (1-p)\alpha_{FS}^2}{pS_{FS}^2}} \\
&= 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{a\sqrt{k}\sqrt{p}} \times \sqrt{1 + \frac{1-p}{p}\Big(\frac{\alpha_{FS}}{S_{FS}}\Big)^2}
\end{aligned}$$

The closer the second factor to 1, the closer the confidence interval is to the infeasible CI length. Since the ratio α_{FS}/S_{FS} is monotonically increasing in $\alpha_{FS} \in [0,1]$ for all $a\sqrt{k} \in \mathbb{R}_+$, the optimal solution is to choose α_{FS} as small as possible. However, when the signal is very weak, then $\mathsf{Z}_{\alpha_{FS}}$ dominates the power function and we have α_{FS}/S_{FS} very close to 1 for all α_{FS}, then no matter how small α_{FS} is chosen, the confidence interval will converge to the

full sample CI. When the signal is not that weak, choosing a smaller α_{FS} pays off to have a smaller α_{FS}/S_{FS}, indeed, for a relatively big $a\sqrt{k}$, we have $\alpha_{FS}/S_{FS} \to 0$ as $\alpha_{FS} \to 0$, then we have the RSS CI converging to infeasible CI.

5 Monte Carlo

In this simulation, we examine the performance of the proposed split-sample and repeated split-sample selective AR test and compare it with the full sample AR test and the naive selective AR test. As is discussed above, the naive selective method violates the exclusion restriction and has inflated size while the full sample method may have very low power when effective first stage policies are mixed with ineffective polices. That is exactly what we will see in this section.

Suppose the instrument is constructed with 40 different natural experiments, so $G = 40$. n is the sample size in each group.For all data generating processes, $X_g^0, \tilde{Z}_g \sim i.i.d.\ N(0,1)$, for all $g = 1, ..., G$. The endogenous variable $W_g = \rho_g \tilde{Z}_g + X_g^0 + v_g$, the outcome variable $Y_g = \beta W_g + X_g^0 + u_g$, and the error terms u_g, v_g follow from standard normal distributions with correlation 0.8 for all $g = 1, ..., G$. The first stage coefficient ρ could be random while the second stage coefficient β is a parameter. We are interested in the estimation and inference of β. We report in tables the proportion of a total of 1000 simulations that reject the null hypothesis $H_0 : \beta = 0$. So far, only unadjusted pointwise testing method is used for the selection part of the proposed split-sample and repeated split-sample selective AR tests. The confidence level used is 0.05 for both the first stage selection and the second stage AR test.

We consider four data generating processes (DGP). For DGP 1 and 3, we set $\beta = 0$ while for DGP 2 and DGP 4, we set $\beta = 0.1$. As is explained in the theoretical section, the power reported in Table 1 depends on the strength of the first stage correlation ρ_g. Let $\rho_g = 0$ for the first 3/4 of the groups and $\rho_g = a$ for the rest 1/4. For DGP 1 and DGP 2, a is fixed

Table 1: Size and Power of AR Tests

	Full Sample AR				Naive Selective AR, PWER				split-sample Selective AR, PWER			
n	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
	DGP 1: $\rho_g = \{a, 0\}$, $\beta = 0$											
a=0	0.047	0.060	0.042	0.060	0.598	0.599	0.581	0.599	0.051	0.040	0.038	0.053
a=0.25	0.056	0.057	0.047	0.046	0.149	0.143	0.132	0.130	0.055	0.057	0.048	0.055
a=0.5	0.057	0.052	0.045	0.046	0.126	0.128	0.125	0.118	0.056	0.057	0.040	0.050
a=0.75	0.038	0.039	0.049	0.061	0.151	0.144	0.129	0.123	0.058	0.049	0.050	0.045
a=1	0.047	0.038	0.052	0.056	0.134	0.130	0.129	0.134	0.051	0.048	0.058	0.044
	DGP 2: $\rho_g = \{a, 0\}$, $\beta = 0.1$											
a=0	0.050	0.059	0.042	0.061	0.623	0.628	0.610	0.623	0.048	0.040	0.036	0.056
a=0.25	0.096	0.125	0.191	0.369	0.492	0.590	0.813	0.958	0.119	0.190	0.333	0.588
a=0.5	0.213	0.344	0.632	0.915	0.793	0.956	0.997	1.000	0.307	0.586	0.854	0.988
a=0.75	0.415	0.691	0.938	0.996	0.972	1.000	1.000	1.000	0.592	0.897	0.995	1.000
a=1	0.624	0.927	0.997	1.000	0.999	1.000	1.000	1.000	0.872	0.991	1.000	1.000
	DGP 3: $\rho_g = \{a/\sqrt{n}, 0\}$, $\beta = 0$											
a=10	0.056	0.057	0.047	0.046	0.202	0.187	0.191	0.180	0.048	0.052	0.049	0.054
a=20	0.057	0.052	0.045	0.046	0.126	0.128	0.125	0.118	0.058	0.057	0.040	0.050
a=40	0.038	0.039	0.049	0.061	0.151	0.144	0.129	0.123	0.058	0.049	0.050	0.045
a=80	0.047	0.038	0.052	0.056	0.134	0.130	0.129	0.134	0.051	0.048	0.058	0.044
	DGP 4: $\rho_g = \{a/\sqrt{n}, 0\}$, $\beta = 0.1$											
a=10	0.077	0.085	0.071	0.078	0.490	0.455	0.478	0.472	0.072	0.073	0.092	0.091
a=20	0.150	0.148	0.144	0.144	0.705	0.656	0.687	0.689	0.239	0.231	0.235	0.226
a=40	0.444	0.448	0.465	0.436	0.984	0.981	0.985	0.988	0.647	0.681	0.673	0.677
a=80	0.956	0.977	0.957	0.952	1.000	1.000	1.000	1.000	0.998	0.997	0.997	0.999

and equal to 0, 0.25, 0.5, 0.75, 1, respectively. For DGP 3 and DGP 4, a degenerates and equal to $1.5/\sqrt{n}$, $3/\sqrt{n}$, $6/\sqrt{n}$, $12/\sqrt{n}$. The sample size n takes values 250, 500, 1000, and 2000.

Table 1 shows the small sample performance of the split-sample selective AR test. The split-sample proportion for the simulation results used in Table 1 is 0.5. Since 1/4 of the sample experience effective policy changes, Lemma 5 predicts that the split-sample test always has better power than the full sample test facing DGP 2. On the other hand, one could calculate that for DGP 4 S_{FS} in Proposition 2 equals to 0.48, 0.94, 1 and 1, respectively. So the split-sample IV regression improves efficiency if p <0.15, 0.44, 0.47, and 0.47, respectively. So it is expected that the split-sample IV regression method is more powerful than the full sample IV regression for all values of a except for the smallest one.

Table 1 shows that the naive selective instrumental variable approach over-rejects while the split-sample selective instrumental variable approach controls size very well, and that

Table 2: Power of AR Tests, Changing the Proportion in the Split-Sample Method

	Full Sample AR					Naive Selective AR					split-sample Selective AR				
n	250	500	1000	2000	4000	250	500	1000	2000	4000	250	500	1000	2000	4000
	$DGP5: p = 0.25$														
a=0.1	0.060	0.071	0.071	0.095	0.175	0.729	0.605	0.478	0.508	0.693	0.056	0.062	0.093	0.145	0.260
a=0.2	0.065	0.101	0.159	0.249	0.430	0.477	0.510	0.704	0.861	0.988	0.078	0.135	0.285	0.543	0.893
a=0.3	0.096	0.162	0.311	0.493	0.785	0.508	0.694	0.915	0.989	1.000	0.117	0.270	0.561	0.886	0.997
a=0.4	0.132	0.267	0.455	0.742	0.957	0.677	0.880	0.985	1.000	1.000	0.205	0.401	0.792	0.983	1.000
	$DGP6: p = 0.5$														
a=0.1	0.076	0.095	0.163	0.252	0.491	0.861	0.765	0.543	0.603	0.847	0.046	0.064	0.120	0.254	0.550
a=0.2	0.137	0.235	0.443	0.741	0.967	0.548	0.575	0.850	0.977	1.000	0.100	0.243	0.538	0.870	0.997
a=0.3	0.291	0.497	0.816	0.976	0.999	0.645	0.857	0.983	1.000	1.000	0.205	0.466	0.856	0.997	1.000
a=0.4	0.450	0.744	0.956	1.000	1.000	0.830	0.978	0.999	1.000	1.000	0.345	0.727	0.982	1.000	1.000

the split-sample selective instrumental variable approach improves the power of the AR test when the theory predicts so. Also note that when implementing the test, we first need to normalize the instrument $\tilde{Z}_g$ to an instrument $Z_g = M(X_g^0)\tilde{Z}_g$ such that $Z_g' X_g^0 = 0$.

Next we use simulation to show that if one let the proportion of data used for IV estimation increases with sample size, the split-sample selective IV method always improves the power as the sample size goes to infinity. We consider DGP 5 and DGP 6 where everything is the same as in DGP 2 except for q, p and a. For both DGP 5 and DGP 6, we set a =0.1, 0.2, 0.3 and 0.4, respectively. We also set $q = 1 - 0.5^{(n/500)^{0.5}}$. In another word, we use $0.5^{(n/500)^{0.5}}$ of the data in each group for first stage selection, while the rest $1 - 0.5^{(n/500)^{0.5}}$ of the data for selective instrumental variable regression. For the five different sample sizes 250, 500, 1000, 2000, 4000 that we considered, $q = 0.387$, 0.5, 0.625, 0.75 and 0.859 respectively. The proportion of effective policies, or p, equals to 0.25 in DGP 5 and 0.5 in DGP 6. For DGP 5, the split-sample selective method improves the power of the AR test for all sample sizes. For DGP 6, however, the split-sample selective IV approach improves the power only for sample size larger than 580 (according to Lemma 5). The simulation results reported in Table 2 perform just as the theory predicts.

Table 3 examine the size of the repeated split-sample AR test and compare it with the full sample and split-sample selective AR tests. In table 3 we consider DGP 7 and 8 where everything is the same as DGP 1 and 3 respectively except that a =0.05, 0.1 or 0.2 as is stated in the table, $p = 0.5$ and that $q = 0.5$. $R = [n^{1/2}]$ or $R = [n^{1/2}/2]$ with $[x]$ denote

Table 3: Size of the Repeated split-sample Selective AR Test

	Full Sample			split-sample			Repeated split-sample 1			Repeated split-sample 2		
n	100	300	900	100	300	900	100	300	900	100	300	900
	DGP 7: $\rho_g = 0$											
$a = 0.2$	0.047	0.050	0.050	0.050	0.045	0.056	0.057	0.052	0.063	0.048	0.057	0.057
	DGP 8: $\rho_g = 0.2$											
$a = 0.05$	0.064	0.087	0.178	0.057	0.069	0.186	0.065	0.085	0.240	0.070	0.089	0.245
$a = 0.1$	0.092	0.231	0.503	0.064	0.160	0.497	0.078	0.191	0.663	0.095	0.210	0.681
$a = 0.2$	0.226	0.555	0.955	0.104	0.435	0.943	0.130	0.513	0.995	0.152	0.554	0.997

the smallest integral larger than x. From results of DGP 7, we see that the repeated split-sample IV regression method method controls size reasonably well. For DGP 8, the criteria in Proposition 1 for the split-sample IV regression method to be more efficient than the full sample IV regression is not met. Therefore we see that the power of the split-sample method grows to 1 slower than the full sample method. However, the repeated split-sample IV regression method outperforms the full sample method as n increases, as is predicted by Lemma 6.

6 Return to Compulsory Schooling

The return to compulsory schooling literature studies how an extra year of (compulsory) schooling affects individual well-being such as earning and health outcomes later in life. Researchers often exploit variations in compulsory schooling laws across states and over time in the U.S. (Lleras-Muney, 2005, Oreopoulos, Page, and Stevens, 2006) and other countries (Oreopoulos, 2006) to instrument individual's endogenous schooling choice that worrily correlates with omitted variables such as individual ability and family background. The argument for this IV regression strategy is that any law change in minimum school leaving age affects individual education attainment, but not individual well-being later in life other than through the education channel.

Over time, different reduced form specifications have been adopted to estimate the return to compulsory schooling and the overall trend is to add in more and more controls as well as

state-of-birth dummies, cohort (birth year) dummies, state-specific time trends or regional cohort dummies. Earlier papers such as Lleras-Muney (2005) use 2SLS with cohort and state dummies to study the effect of (compulsory) education on mortality rates and find significant and positive effects. Oreopoulos (2006) and Oreopoulos, Page, and Stevens (2006) also find significant and positive effects of education on wages and health outcomes using similar parametric specifications. They also notice that this statistical significance disappears when state-specific time trends and state-year controls[12] on demographic and economic conditions are included into the regression in addition to or in replacement of the cohort dummies.

Stephens and Yang (2014) is the most recent paper in the literature that carefully examines the benefits of compulsory education in the U.S. Instead of using state-specific time trends, they divide the U.S. into 4 geographic regions and add into their specification regional specific cohort dummies. They also find insignificant returns to additional year of compulsory schooling with their flexible specification.

Neither specification used in Oreopoulos (2006) and Stephens and Yang (2014) are nested within each other. In this section we extend the specification used in Oreopoulos (2006) with state-specific time trend and state controls to allow for heterogenous effect of the compulsory law changes in the first stage. Formally, we consider the following specification.

$$Logwage_{i,sdt} = \beta Educ_{i,sdt} + X_{sdt}r + u_{i,slt}$$

$$Educ_{i,sdt} = \rho_{sd} CL_{st} + X_{sdt}\alpha + v_{i,sdt}$$

[12]The state-year controls include average age, the faction of the state population living in urban areas, living on a farm, being black, being in the labor force and working in the manufacturing industry in Oreopoulos (2006). The controls include the state population, the number of doctors per capita, the value per acre of farm land, the percentage of the state population that is foreign born, the percentage of the state population living in urban areas, being black, working in the manufacturing industry and average manufacturing wages per worker in Oreopoulos (2006). In both papers, the state-year controls are aligned with individuals' state of birth at age 14.

where $Educ_{i,sdt}$ is the years of schooling if individual i of demographic group (while male, white female, black male, black female) d born in state s in year t while CL_{st} is the compulsory schooling year an individual face at age 14 if he/she was born in year t and live in state s. The regressor X_{st} includes a quartic in age, census year indicators, state-specific time trends, state-year controls and vectors of demographic group, state-of-birth fixed effects. The first stage correlation between the instrument and the endogenous regressor are allowed to be heterogenous across states and among demographic groups, which is consistent with the documentation in Lleras-Muney (2005) and (Gu and Shen, 2014) where researchers find that some law changes (in certain states) may not be effective and that people from different part of the population may have heterogenous respond to the same law change. We extend in this section the specification of Oreopoulos (2006) rather than that of Stephens and Yang (2014) because state-specific first stage effects of the law changes cannot be identified together with cohort fixed effects.

The data we use in this section merges the individual-level 1960-1980 public-use census data compiled by Stephens and Yang (2014) with the state-year controls provided by Oreopoulos (2006). Before we discuss the implementation of our selective IV regression method, we compare in Table 4 the full sample IV regression results from different specifications that have been used in the literature. The results are reported in Column 1-3. We see that the statistical significance of positive return to compulsory schooling disappears once state-specific time trends or regional cohort dummies are controlled, as is described in Oreopoulos (2006) and Stephens and Yang (2014). We also report in the table two different inference results with different clustering strategies.[13] In the literature, the dominating strategy is to cluster at the state-year level where state refers to the state-of birth and year refers to the

[13]In the theoretical sections, homoskedasticity is assumed on the outcome equation. This is because if we allow σ_u^2 to depend on X, then the AR CI won't have the nice close form solution that we rely on in all our comparisons.

Table 4: Classic IV Regression Results

	Full Sample			States with Law Changes*		
2SLS	0.103	0.046	-0.054	0.142	0.045	-0.007
cluster at state-year						
2SLS CI	(0.073, 0.134)	(-0.017, 0.108)	(-0.169, 0.061)	(0.104, 0.180)	(-0.023, 0.113)	(-0.159, 0.136)
AR CI	(0.074, .136)	(-0.027, 0.113)	(-0.265, 0.034)	(0.107, 0.184)	(-0.035, 0.121)	(-0.653, 0.122)
First Stage F	96.13	18.92	12.93	77.07	15.44	5.08
cluster at state						
2SLS CI	(0.035, 0.171)	(-0.029, 0.120)	(-0.469, 0.361)	(0.070, 0.214)	(-0.021, 0.111)	(-0.390, 0.376)
AR CI	(0.023, 0.189)	(-0.604, 0.158)	$(-\infty, 0.171) \cup (0.422, \infty)$	(0.065, 0.244)	$(-\infty, \infty)$	$(-\infty, \infty)$
First Stage F	11.04	4.06	1.44	21.31	3.20	0.77
N (in thousands)	3741	3741	3741	2882	2882	2882
State-of-birth Dummies	Y	Y	Y	Y	Y	Y
State-year Controls	Y	Y	N	Y	Y	N
State-specific Trends	N	Y	N	N	Y	N
Region-cohort Dummies	N	N	Y	N	N	Y

Note: The subsample also excludes individual where less than 1000 individuals of his/her demographic groups (typicall black men and black women) are observed in a given state, as is described in the main text.

birth year. However, if individuals born in different states but around the same time period are exposed to similar shocks, clustering at state-year level may result in over-rejection in hypotheses testing. An alternative strategy is to cluster at the state level only. We see from Column 1-3 that second more robust inference method produce substantially wider confidence intervals. Under the specification of Stephens and Yang (2014), cluster at state level even yields unbounded AR confidence interval.

With our simultaneous equation with heterogenous first stage policy effect, we focus our regression to states with changes in compulsory schooling observed in the dataset. Since we also allow the effect of compulsory schooling laws to be different for different race gender groups, we also exclude from our data demographic state groups with less than 1000 individuals. For example, black men and women in Idaho, are also dropped out. This leaves us with 30 states and 2,882,377 individuals.

The first column of Table 5 reports the classic 2SLS and AR regression results for the full sample and for the white men. We notice that the AR CIs are not informative at all when state level clusters are used instead of the state-year level clusters. The AI confidence interval clustering at the state-year level is also very wide for the subsample of white men. The second column reports the naive selective IV regression results. As we discussed in the theoretical section, the naive strategy gives inconsistent results. Column 3-6 report

two sets of split-sample and repeated split-sample IV regression results. The two sets of estimators are different because different random seeds are used when splitting the dataset. These columns show how our proposed split-sample strategies could potentially improve accuracy of IV regression. First, we notice that with state-level clustering, the proposed split-sample and the repeated split-sample methods improves the accuracy of IV regression substantially for both regressions on the full sample and the white men. For white men, the AR CI with state-year level clustering is also greatly improved with the proposed split-sample strategies. Second, we notice that the repeated split-sample methods perform better than the split-sample methods, yielding shorter confidence intervals and more stable results. In addition to the results in Column 1-6, we also report in Column 7 and 8 repeated split-sample IV regression results with $R = 100$. They yield qualitatively similar results compared to repeated split-sample IV regression with $R = 25$ although to be able to perform repeated split-sample with $R = 100$ more black groups are dropped from the data because of lack of observations.

Generally, one is able to conclude with 95% confidence level that an additional year of education increases an individual's wage by no more than 12% on average. For white men, the additional year of education increases an individual's wage by at least 2% on average.

In Table 6, we take the split-sample and repeated split-sample selection results calculated above under the Oreopoulos (2006) specification and perform various IV regressions under the most robust specification in (Stephens and Yang, 2014). This is not how we mean to do our selective IV regression methods. But we still see that the selection tends to improve the inference when the state level clusters are used instead of the state-year level clusters.

Table 5: Selective IV Regression Results: State-year Controls and State-specific Trends

	Baseline	Naive	SS 1	RSS 1 (r=25)	SS 2	RSS 2 (r=25)	RSS 1 (r=100)	RSS 2 (r=100)
	Full sample							
2SLS	0.045	0.013	0.005	0.041	0.016	0.044	0.020	0.051
cluster at state-year								
2SLS CI	(-0.023, 0.113)	(-0.058, 0.085)	(-0.102, 0.113)	(-0.012, 0.093)	(-0.098, 0.129)	(-0.018, 0.105)	(-0.035, 0.075)	(0.000, 0.103)
AR CI	(-0.035, .121)	(-0.067, 0.081)	(-0.147, 0.118)	(-0.017, 0.094)	(-0.175, 0.119)	(-0.026, 0.105)	(-0.044, 0.074)	(-0.003, 0.104)
First Stage F	15.44	51.11	11.58	33.65	10.90	31.79	31.23	37.80
cluster at state								
2SLS CI	(-0.021, 0.111)	(-0.070, 0.096)	(-0.137, 0.147)	(-0.027, 0.108)	(-0.150, 0.181)	(-0.044, 0.132)	(-0.050, 0.091)	(-0.010, 0.113)
AR CI	(-∞, ∞)	(-0.085, 0.089)	(-0.313, 0.115)	(-0.055, 0.102)	(-0.304, 0.151)	(-0.098, 0.129)	(-0.132, 0.074)	(-0.035, 0.111)
First Stage F	3.20	38.87	8.22	16.11	9.55	10.08	9.18	13.74
N (in thousands)	2882	481	206	2535	196	2409	2842	2841
	White men:							
2SLS	0.217	0.061	0.264	0.114	0.074	0.178	0.167	0.139
cluster at state-year								
2SLS CI	(0.0042, 0.392)	(-0.019, 0.141)	(0.118, 0.411)	(0.042, 0.186)	(-0.015, 0.162)	(0.082, 0.273)	(0.077, 0.256)	(0.060, 0.219)
AR CI	(0.091, 0.799*)	(-0.038, 0.142)	(0.158,0.602)	(0.044, 0.204)	(-0.04, 0.171)	(0.093, 0.317)	(0.091, 0.299)	(0.067 0.246)
First Stage F	6.56	18.97	9.49	21.06	12.79	16.54	17.83	19.76
cluster at state								
2SLS CI	(-0.127, 0.561)	(-0.034, 0.156)	(0.148, 0.381)	(0.017, 0.211)	(-0.055, 0.202)	(0.025, 0.331)	(0.026, 0.307)	(0.028, 0.250)
AR CI	(-∞, ∞)	(-0.177 0.12)	(0.187, 0.795)	(-0.017, 0.213)	(-0.229, 0.164)	(0.044, 0.799*)	(0.051, 0.556)	(0.018, 0.294)
First Stage F	1.50	8.23	4.58	13.35	8.01	4.84	6.60	10.59
N (in thousands)	1505	144	106	1293	61	1267	1505	1504

Note: All confidence intervals are calculated using 95% confidence level. * *the 0.799 is from the bound set in Stata for searching AR CI. Need to be reexamined before submission.*

Table 6: Selective IV Regression Results: State-year Controls and State-specific Trends

	Baseline	Naive	SS 1	RSS 1	SS 2	RSS 2
	Specification SY, full sample					
2SLS	-0.007	0.011	-1.33	0.047	0.092	0.09
	Inference clustering at the state-year level					
2SLS CI	(-0.159,0.136)	(0.014, 0.207)	(-8.045, 5.382)	(-0.292, 0.208)	(-0.030, 0.213)	(-0.019, 0.199)
AR CI	(-0.653, .122)	(0.008, 0.216)	(-∞, -0.042) ∪ (0.381, ∞)	(-0.292,0.208)	(-0.056, 0.224)	(-0.045, 0.205)
First Stage F	5.08	28.32	0.18	7.85	15.17	16.15
	Inference clustering at the state level					
2SLS CI	(-0.390,0.376)	(0.004, 0.217)	(-21.062, 18.4)	(-0.365, 0.459)	(-0.055, 0.238)	(-0.129, 0.299)
AR CI	(-∞, ∞)	(-∞, ∞)	(-∞, -0.015) ∪ (0.069, ∞)	(-∞, ∞)	(-∞, -0.785) ∪ (-0.551, ∞)	(-∞, ∞)
First Stage F	0.77	2.15	0.02	0.73	3.41	2.26
N						
	Specification SY, white men:					
2SLS	0.026	0.059	0.480	0.061	0.080	0.136
	Inference clustering at the state-year level					
2SLS CI	(-0.217,0.270)	(-0.004, 0.123)	(-0.06, 1.02)	(-0.074, 0.195)	(-0.056, 0.215)	(-0.030, 0.301)
AR CI	(-∞, ∞)	(-0.011, 0.122)	(0.217, 0.799*) ∪ (0.381, ∞)	(-0.307,0.202)	(-0.164, 0.244)	(-∞, ∞)
First Stage F	2.02	37.25	2.21	5.88	7.01	3.47
	Inference clustering at the state level					
2SLS CI	(-0.474,0.526)	(0.003, 0.116)	(-0.502, 1.461)	(-0.201, 0.322)	(-0.112, 0.271)	(-0.190, 0.462)
AR CI	(-∞, ∞)	(-0.048, 0.099)	(-∞, -0.213) ∪ (0.183, ∞)	(-∞, ∞)	(-∞, ∞)	(-∞, ∞)
First Stage F	0.44	12.01	0.54	1.34	1.98	0.81
N						

Note: All confidence intervals are calculated using 95% confidence level. * *the 0.799 is from the bound set in Stata for searching AR CI. Need to be reexamined before submission.*

7 Appendix

Proof for Lemma 1

Proof. First we show some results that will be repeatedly used in this proof. Let $D_{\rho_g-\rho}$ be the diagonal matrix with the $\sum_{l=1}^{g-1} n_l + i$ element equal to $\rho_g - \rho$ for $i = 1, \ldots, n_g$, then $\epsilon = v + D_{\rho_g-\rho} Z$.

Lemma 12. *1.* $P_Z Z = Z,\ P_{X^0} X^0 = X^0,\ M_X Z = 0,\ M_X X^0 = 0,\ H_X Z = Z,$

2. $v' P_Z v \Rightarrow \sigma_v^2 \chi^2(1),\ v' M_X v \Rightarrow \sigma_v^2 \chi^2(N-d-1),\ \frac{v' M_X v}{N-d-1} \xrightarrow{p} \sigma_v^2,\ \frac{u' M_X u}{N-d-1} \xrightarrow{p} \sigma_u^2,$

3. $P_Z X^0 = 0,\ P_{X^0} Z = 0,\ P_X = P_Z + P_{X^0},\ H_X X^0 = 0,$

4. $Z' D_{\rho_g-\rho} Z/N \xrightarrow{p} 0,\ Z' D_{\rho_g-\rho} X^0/N \xrightarrow{p} 0,\ Z' P_Z \epsilon/N \xrightarrow{p} 0,\ \epsilon' P_Z \epsilon/N \xrightarrow{p} 0,\ \epsilon' M_X \epsilon/N \xrightarrow{p} \sigma_v^2 + a^2 kp(1-p),\ \epsilon' H_X \epsilon/N \xrightarrow{p} 0.$

5. $u' P_Z \epsilon/\sqrt{N} \xrightarrow{p} 0,\ u' M_X \epsilon/(N-d-1) \xrightarrow{p} 0$

Proof. The first and second parts are trivial. The third part follows from $Z' X_0 = 0$. Let

$N_1 = \sum_{g=1}^{G} 1(\rho_g = a) n_g$ and $N_0 = \sum_{g=1}^{G} 1(\rho_g = 0) n_g$; $N_1 + N_0 = N$ and by definition, $N_1/N = p$. In the fourth part of the lemma, because Z_{ig} and X^0_{ig} are i.i.d. across groups,

$$\begin{aligned}
Z' D_{\rho_g - \rho} Z/N &= (a-\rho)\frac{N_1}{N}\frac{1}{N_1}\sum_{g\in G_+}\sum_{i=1}^{n_g} Z_{ig}^2 + (0-\rho)\frac{N_0}{N}\frac{1}{N_0}\sum_{g\in G_+^c}\sum_{i=1}^{n_g} Z_{ig}^2 \\
&\xrightarrow{p} (a-\rho)\frac{N_1}{N}k + (0-\rho)\frac{N_0}{N}k = 0 \\
Z' D_{\rho_g - \rho} X^0/N &= (a-\rho)\frac{N_1}{N}\frac{1}{N_1}\sum_{g\in G_+}\sum_{i=1}^{n_g} Z_{ig}X_{ig}^0 + (0-\rho)\frac{N_0}{N}\frac{1}{N_0}\sum_{g\in G_+^c}\sum_{i=1}^{n_g} Z_{ig}X_{ig}^0 \\
&\xrightarrow{p} (a-\rho)\frac{N_1}{N}E[Z_{ig}X_{ig}^0] + (0-\rho)\frac{N_0}{N}E[Z_{ig}X_{ig}^0] = 0
\end{aligned}$$

Then it is easy to show that

$$\begin{aligned}
Z' P_Z \epsilon/N &= Z'v/N + Z' D_{\rho_g-\rho} Z/N \xrightarrow{p} 0 \\
\epsilon' P_Z \epsilon/N &= v' P_Z v/N + (Z' D_{\rho_g-\rho} Z/N)^2/(Z'Z/N) \xrightarrow{p} 0 \\
\epsilon' P_{X^0} \epsilon/N &= v' P_Z v/N + (Z' D_{\rho_g-\rho} X^0/N)(X^{0\prime} X^0/N)^{-1}(Z' D_{\rho_g-\rho} X^0/N)' \xrightarrow{p} 0, \\
\epsilon' M_X \epsilon/N &= v'v/N + Z' D_{\rho_g-\rho} D_{\rho_g-\rho} Z/N - \epsilon' P_Z \epsilon/N - \epsilon' P_{X^0}\epsilon/N \to \sigma_v^2 + a^2 kp(1-p), \\
\epsilon' H_X \epsilon/N &= \epsilon' P_Z \epsilon/N - \frac{c}{N-d-1}\epsilon' M_X \epsilon/N \xrightarrow{p} 0.
\end{aligned}$$

For the last part of the lemma,

$$\begin{aligned}
u' P_Z \epsilon/\sqrt{N} &= u' P_Z v/\sqrt{N} + u' P_Z D_{\rho_g-\rho} Z/\sqrt{N} \\
&= (u'Z/\sqrt{N})(Z'Z/N)(Z'v/N) + (u'Z/\sqrt{N})(Z'Z/N)(Z' D_{\rho_g-\rho} Z/N) \\
&\xrightarrow{p} 0, \\
u' M_X \epsilon/(N-d-1) &= u'\epsilon/(N-d-1) - (u'X/N)(X'X/N)^{-1} X'\epsilon/(N-d-1) \\
&\xrightarrow{p} 0.
\end{aligned}$$

□

Now we prove statements in Lemma 1

Since $|CI^{2SLS}| = 2\mathsf{t}_{N-d-1,\alpha/2}\sqrt{\frac{\hat{\sigma}_u^2}{W'P_ZW}}$, $\mathsf{t}_{N-d-1,\alpha/2} \to \mathsf{Z}_{\alpha/2}$, $\hat{\sigma}_u \xrightarrow{p} \sigma_u$, the asympotic result of the 2SLS confidence interval is straightforward given that

$$W'P_ZW/N = (\rho Z + X^0\gamma + \epsilon)'P_Z(\rho Z + X^0\gamma + \epsilon)/N = \rho^2(Z'Z/N) + 2\rho Z'P_Z\epsilon/N + \epsilon'P_Z\epsilon/N \xrightarrow{p} ka^2p^2.$$

To prove the part of the lemma concerning the asymptotics of the AR confidence interval, we first have that

$$W'H_XW/N = W'P_ZW/N - \frac{c}{N-d-1}W'M_XW/N = W'P_ZW/N - \frac{c}{N-d-1}\epsilon'M_X\epsilon/N \xrightarrow{p} ka^2p^2.$$

Therefore, the probability that the AR confidence interval is bounded also converges to 1. In addition,

$$\begin{aligned}
(Y'H_XW)^2 &= \beta_0^2(W'H_XW)^2 + 2\beta_0(W'H_XW)(u'H_XW) + (u'H_XW)^2, \\
(Y'H_XY)(W'H_XW) &= \beta_0^2(W'H_XW)^2 + 2\beta_0(W'H_XW)(u'H_XW) + (u'H_Xu)(W'H_XW), \\
u'H_XW/\sqrt{N} &= u'H_X(Z\rho + X^0\alpha + \epsilon)/\sqrt{N}, \\
&= u'Z\rho/\sqrt{N} + u'P_Z\epsilon/\sqrt{N} - (c/\sqrt{N})u'M_X\epsilon/(N-d-1), \\
&= u'Z\rho/\sqrt{N} + o_p(1) + o_p(1/\sqrt{N}) \\
u'H_Xu &= u'P_Zu - \frac{c}{N-1-d}u'M_Xu = (Z'Z)^{-1}(u'Z)^2 - c\sigma_u^2 + o_p(1).
\end{aligned}$$

Recall that $\sqrt{c} \to \mathsf{Z}_{\alpha/2}$, we have that

$$\begin{aligned}
\sqrt{N}|CI^{AR}| &= 2 \times \sqrt{N\frac{(Y'H_XW)^2 - (Y'H_XY)(W'H_XW)}{(W'H_XW)^2}} \\
&= 2 \times \sqrt{\frac{(u'H_XW/\sqrt{N})^2 - (u'H_Xu)(W'H_XW/N)}{(W'H_XW/N)^2}} \\
&= 2 \times \sqrt{\frac{(u'Z\rho)^2/N - ((Z'Z)^{-1}(u'Z)^2 - c\sigma_u^2)(\rho^2Z'Z)/N}{((\rho^2Z'Z)/N)^2}} + o_p(1) \\
&\xrightarrow{p} 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{ap\sqrt{k}}
\end{aligned}$$

Note that in finite sample the AR confidence interval could be empty, but such probability goes to zero in the limit. □

Proof of Lemma 2

Proof. Assume that $\hat{\sigma}_v \neq 0$ and $Z_g'Z_g \neq 0.$; both statements are true with probability approaching one, under Assumption 2. Recall that $Z_g'X_g^0 = 0$ for all $g = 1, ..., G$, then all selected groups satisfy that

$$t_g = \frac{\rho_g + Z_g'v_g(Z_g'Z_g)^{-1}}{(Z_g'Z_g)^{-1/2}\hat{\sigma}_v} > c_{g,FS} \Rightarrow \frac{Z_g'v_g}{\sqrt{n_g}} > c_{g,FS}\hat{\sigma}_v(Z_g'Z_g/n_g)^{1/2} - \sqrt{n_g}(Z_g'Z_g/n_g)\rho_g.$$

For those groups with $\rho_g = a$,

$$\begin{aligned}
E[Z_{ig}v_{ig}|t_g > c_{g,FS}, \rho_g = a] =& \frac{1}{\sqrt{n_g}}E\left[\frac{Z_g'v_g}{\sqrt{n_g}}\middle|\frac{Z_g'v_g}{\sqrt{n_g}} > c_{g,FS}\hat{\sigma}_v(Z_g'Z_g/n_g)^{1/2} - \sqrt{n_g}(Z_g'Z_g/n_g)a\right] \\
=& o_p\left(\frac{1}{\sqrt{n_g}}\right).
\end{aligned}$$

For those groups with $\rho_g = 0$,

$$\begin{aligned} E[Z_{ig}v_{ig}|t_g > c_{g,FS}, \rho_g = 0] =& \frac{1}{\sqrt{n_g}} E\left[\frac{Z_g'v_g}{\sqrt{n_g}}\middle|\frac{Z_g'v_g}{\sqrt{n_g}} > c_{g,FS}\hat{\sigma}_v(Z_g'Z_g/n_g)^{1/2}\right] \\ =& O_p\left(\frac{1}{\sqrt{n_g}}\right) \end{aligned}$$

When $u_{ig} = \eta v_{ig} + e_{ig}$ with $E[e_{ig}|v_{ig}, Z_{ig}] = 0$, we also have that

$$\begin{aligned} E[Z_{ig}u_{ig}|t_g > c_{g,FS}, \rho_g = 0] &= \eta E[Z_{ig}v_{ig}|t_g > c_{g,FS}, \rho_g = a] + E[Z_{ig}e_{ig}|t_g > c_{g,FS}, \rho_g = a] \\ &= O_p\left(\frac{1}{\sqrt{n_g}}\right) + E[Z_{ig}E[e_{ig}|Z_{ig}, v_{ig}]|t_g > c_{g,FS}, \rho_g = a] \\ &= O_p\left(\frac{1}{\sqrt{n_g}}\right). \end{aligned}$$

That is, the exclusion restriction is violated to the rate of $\frac{1}{\sqrt{n_g}}$. With significance level α_{FS}, the probability of having at least one group falsely selected out of all G_+^c groups is $1 - (1 - \alpha_{FS})^{G_0}$. The lemma is hence proved. □

Proof for Lemma 5

Proof. By results in lemma 3, for all three multiple testing methods, any group g passing the first stage selection satisfy that $t_g > c_{g,FS}$ with $c_{g,FS} = t_{\lfloor n_g(1-q)\rfloor - d - 1, \alpha_{FS}}$. Given dataset, the proportion of data selected for 2SLS or AR regression equals to

$$\begin{aligned} N_{IV}/N =& \frac{1}{N}\sum_{g=1}^{G} 1(t_g > c_{g,FS})\lceil n_g q\rceil = \sum_{g=1}^{G} 1(t_g > c_{g,FS})\frac{\lceil n_g q\rceil}{\sum_{g=1}^{G} n_g} \\ =& \sum_{g=1}^{G} w_{G,g}X_{G,g} \end{aligned}$$

where $X_{G,g} = 1(t_g > c_{g,FS})$ is a triangular array of random variables and $w_{G,g} = \frac{\lceil n_g q\rceil}{\sum_{g=1}^{G} n_g}$ a a triangular array of weights. Then the limiting result N_{IV}/N can be derived by applying

weak law of large numbers for weighted triangular arrays studied in Rosalsky and Sreehari (1998). Notice that

$$G \max_{g=1,..,G} \left| w_{G,g} - \frac{1}{G} \right| < \infty,$$
$$\frac{1}{G} \left| X_{G,g} - \Phi\left(\frac{\sqrt{k}}{\sigma_v} n_{g_0} \rho_g - c_{g,FS} \right) \right| \xrightarrow{p} 0.$$

Then

$$\begin{aligned} N_{IV}/N &\xrightarrow{p} \sum_{g=1}^{G} w_{G,g} \Phi\left(\frac{\sqrt{k}}{\sigma_v} n_{g_0} \rho_g - c_{g_0,FS} \right) \\ &= \sum_{g \in G_+} w_{G,g} \Phi\left(\frac{\sqrt{k}}{\sigma_v} n_{g_0} a - c_{g_0,FS} \right) + \sum_{g \in G_+^c} w_{G,g} \alpha_{FS} \\ &\xrightarrow{p} qp + q(1-p)\alpha_{FS}. \end{aligned}$$

The first convergence holds when $G \to \infty$ and the second when $n_g(1-q) \to \infty, G \to \infty$ for all $g = 1, ..., G$. For fixed q, the selective IV regression methods, including both the 2SLS and the AR test, have, in the limit, sample size equal to $(p + \alpha_{FS}(1-p))q$ proportion of the full sample and the correlation between the instrument and the endogeneous covariate is $\frac{pa}{\alpha_{FS}(1-p)+p}$. Following the derivation for Lemma 1 and replacing the sample size (originally N) by $(p+\alpha_{FS}(1-p))qN$ and the average first stage correlation (originally ap) by $\frac{pa}{\alpha_{FS}(1-p)+p}$, the limiting result of Lemma 5 is clear, which includes the asymptotic lengths of both 2SLS and AR confidence intervals and the fact that the probability that AR confidence interval is bounded goes to one in the limit. □

Proof of Lemma 6

Proof. First we show that the weighting matrix $\hat{Q} \xrightarrow{p} Q$, where Q is a $N \times N$ diagonal matrix with the $\sum_{l=1}^{g-1} n_l + i$-th diagonal element equaling to 1 if $g \in G_+$ and α_{FS} otherwise, then $trace(Q)/N \xrightarrow{p} (1-(1-p)(1-\alpha_{FS}))$. Let $\hat{\rho}_g^r$ be a sequence of first stage OLS estimators of ρ_g in group g and subsample r and $\hat{\sigma}_g^r$ the corresponding estimator for σ_v/k so that $\hat{\sigma}_g^r/\sqrt{n_g/R}$ is the standard error of $\hat{\rho}_g^r$. Likewise, let $\hat{\rho}_g$ be the first stage OLS estimator of ρ_g using data of all subsamples in group g and $\hat{\sigma}_g/\sqrt{n_g}$ be its standard error estimator. Define $\hat{L}_{g,R}(x) = \frac{1}{R}\sum_{r=1}^{R} 1\left(r_{g,R}(\hat{\rho}_g^r - \rho_g) \leq x\right)$, where $r_{g,R} = \sqrt{n_g/R}/\hat{\sigma}_g^r$. By Theorem 2.2.1 and Corollary 2.4.1 of Politis, Romano, and Wolf (1999) and the fact that standard normal is the limiting distribution of $r_{g,R}(\hat{\rho}_g^r - \rho_g)$, if $R \to \infty$, $n_g/R \to \infty$, $\hat{L}_{g,R}(\mathsf{Z}_{\alpha_{FS}}) \xrightarrow{p} \alpha_{FS}$ where α_{FS} is the fixed significance level of the first stage testing. The proof parallels that for the first part of Theorem 2.2.1. in Politis, Romano, and Wolf (1999). It is done by first showing that $\hat{L}_{g,R}(\mathsf{Z}_{\alpha_{FS}}) \xrightarrow{p} Prob[r_{g,R}(\hat{\rho}_g^r - \rho_g) \leq \mathsf{Z}_{\alpha_{FS}}]$ as $R \to \infty$. Then since $Prob[r_{g,R}(\hat{\rho}_g^r - \rho_g) \leq \mathsf{Z}_{\alpha_{FS}}] \to \alpha_{FS}$ as $n_g/R \to \infty$.

Let $t_g^r = r_{g,R}\hat{\rho}_g^r$ be the sequence of t-statistic of first stage correlation for group g and subsample r. Then it is clear that $\frac{1}{R}\sum_{r=1}^{R} 1(t_g^r > \mathsf{Z}_{\alpha_{FS}}) \xrightarrow{p} 1$ if $\rho_g > 0$ and $\frac{1}{R}\sum_{r=1}^{R} 1(t_g^r > \mathsf{Z}_{\alpha_{FS}}) \xrightarrow{p} \alpha_{FS}$ if $\rho_g = 0$ for all $g = 1, ..., G$.

Next we show that

$$\sqrt{N}(\hat{\beta}^{2SLS,RSS} - \beta) \Rightarrow N\left(0, \frac{\sigma_u^2}{a^2 p^2 k}\left(1-(1-p)(1-\alpha_{FS}^2)\right)\right).$$

Recall that $Z'\hat{Q}X^0 = 0$ by normalization,

$$\hat{\beta}^{2SLS,RSS} = (Z'\hat{Q}W)^{-1}Z'\hat{Q}Y = \beta + (Z'\hat{Q}W/N)^{-1}Z'\hat{Q}u/N.$$

Write $h_g = 1(g \in G_+) + \alpha_{FS}(g \in G_+^c)$ and $\hat{h}_g^r = \frac{1}{R-1}\sum_{l\neq r, l=1,...,R} 1(t_g^l > c_{g,FS}) = \frac{1}{R-1}\sum_{l=1,...,R} 1(t_g^l > c_{g,FS}) - \frac{1}{R-1}1(t_g^r > c_{g,FS}) \equiv \hat{h}_g\left(t_g^1, ..., t_g^R\right) - \frac{1}{R-1}1(t_g^r > c_{g,FS})$. Note that $\hat{h}_g^r = w_{ig}$ and h_g are elements in matrix Q.

$$
\begin{aligned}
Z'\hat{Q}u/\sqrt{N} =& \sum_{g=1}^{G}\sqrt{\frac{n_g}{N}}\frac{1}{\sqrt{R}}\sum_{r=1}^{R}\frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}\hat{h}_g^r = \sum_{g=1}^{G}\sqrt{\frac{n_g}{N}}\frac{1}{\sqrt{R}}\sum_{r=1}^{R}\left\{\frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}\hat{h}_g^r - E\left[\frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}\hat{h}_g^r\right]\right\} \\
=& \sum_{g=1}^{G}\sqrt{\frac{n_g}{N}}\frac{1}{\sqrt{R}}\sum_{r=1}^{R}\left\{\left[\frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}\hat{h}_g\left(t_g^1, ..., t_g^R\right) - \frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}(R-1)}1(t_g^r > c_{g,FS})\right]\right. \\
& \left. -E\left[\frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}\hat{h}_g\left(t_g^1, ..., t_g^R\right) - \frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}(R-1)}1(t_g^r > c_{g,FS})\right]\right\} \\
\equiv& \sum_{g=1}^{G}\sqrt{\frac{n_g}{N}}\frac{1}{\sqrt{R}}\sum_{r=1}^{R}\left\{\hat{f}_{n,g}(\mathbf{X}_g^r; \mathbf{X}_g^1, ..., \mathbf{X}_g^R) - E\left[\hat{f}_{n,g}(\mathbf{X}_g^r; \mathbf{X}_g^1, ..., \mathbf{X}_g^R)\right]\right\}.
\end{aligned}
$$

where $\mathbf{X}_g^r = (Z_g^r \ u_g^r \ t_g^r)$ with joint distribution F. The second equality holds because $\hat{h}_g^r$ is independent of Z_g^r and u_g^r. The L^2 distance between $\hat{f}_{n,g}$ and $f_{0,g}$ where $f_{0,g}(X_g^r) \equiv \frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}h_g$.

$$
\begin{aligned}
&\int\left(\hat{f}_{n,g}(\mathbf{X}_g^r; \mathbf{X}_g^1, ..., \mathbf{X}_g^R) - f_{0,g}(\mathbf{X}_g^r)\right)^2 dF(\mathbf{X}_g^r) = \int\frac{\left((Z_g^r)'u_g^r\right)^2}{n_g/R}\left(\hat{h}_g^r - h_g\right)^2 dF(\mathbf{X}_g^r) \\
&= \frac{\left(\hat{h}_g^r - h_g\right)^2}{n_g/R}E\left[\left((Z_g^r)'u_g^r\right)^2\right] = \left(\hat{h}_g^r - h_g\right)^2 E[Z_{ig}^2]E[u_{ig}^2] \\
&\xrightarrow{p} 0
\end{aligned}
$$

The second equality comes from the fact that $\hat{h}_g^r$ is a function of $\mathbf{X}_g^1, ..., \mathbf{X}_g^{r-1}, \mathbf{X}_g^{r+1}, ..., \mathbf{X}_g^R$ but not $\mathbf{X}_g^r$. The third equality comes form the fact that Z_{ig} and u_{ig} are independent. The last convergence result comes from the boundedness assumption of $E[Z_{ig}^2]$ and σ_u^2 and the fact that $\hat{h}_g^r \xrightarrow{p} h_g$, as is shown in the first part of the proof. By Lemma 19.24 of van der

Vaart (1998) we have that

$$Z'\hat{Q}u/\sqrt{N} \stackrel{d}{=} \sum_{g=1}^{G}\left\{\frac{1}{\sqrt{R}}\sum_{r=1}^{R}\left(f_{0,g}(\mathbf{X}_g^r) - E\left[f_{0,g}(\mathbf{X}_g^r)\right]\right)\right\}$$
$$= \sum_{g=1}^{G}\frac{1}{\sqrt{R}}\sum_{r=1}^{R}\frac{(Z_g^r)'u_g^r}{\sqrt{n_g/R}}h_g = Z'Qu/\sqrt{N},$$

where $\stackrel{d}{=}$ denote that two random variable have the same limiting distribution. Likewise, we can show that

$$Z'\hat{Q}W/N = Z'QW/N + o_p(1)$$

given boundedness of $E[Z_{ig}^4]$ and σ_v^2. It is then clear that the length of the 2SLS confidence interval goes in the limit to $\frac{2\mathsf{Z}_{\alpha/2}\sigma_u}{\sqrt{(Z'QW)^{-1}(Z'QQZ)(Z'QW)^{-1}}} = 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{\sqrt{W'P_Z^QW}}$. Calculations similar to that in proof for Lemma 1 show that

$$W'QZ/N = Z'D_{\rho_g}QZ/N + o_p(1) = a\frac{N_1}{N}\frac{1}{N_1}\sum_{g\in G_+}\sum_{i=1}^{n_g}Z_{ig}^2 + o_p(1) \xrightarrow{p} apk,$$
$$Z'QQZ/N = \frac{N_1}{N}\frac{1}{N_1}\sum_{g\in G_+}\sum_{i=1}^{n_g}Z_{ig}^2 + \frac{N_0}{N}\frac{1}{N_0}\sum_{g\in G_+^c}\sum_{i=1}^{n_g}Z_{ig}^2\alpha_{FS}^2 \xrightarrow{p} (1-(1-p)(1-\alpha_{FS}^2))k,$$
$$W'P_Z^QW/N = (W'QZ/N)^2(Z'QQZ/N)^{-1} \xrightarrow{p} \frac{a^2p^2k}{1-(1-p)(1-\alpha_{FS}^2)}.$$

The asymptotic length of the 2SLS confidence interval is then clear.

By similar arguments as before, given boundedness of $E[Z_{ig}^4]$, one can show that $Z'\hat{Q}\hat{Q}Z/N = Z'QQZ/N + o_p(1)$. Together with the fact that $Z'\hat{Q}u/\sqrt{N} \stackrel{d}{=} Z'Qu/\sqrt{N}$ and that $Z'\hat{Q}W/N =$

$Z'QW/N + o_p(1)$, it is easy to show that the repeated split-sample AR test

$$AR(\beta_0) = \frac{(Y - \beta_0 W)' P_Z^{\hat{Q}} (Y - \beta_0 W)}{(Y - \beta_0 W)' M_X^{\hat{Q}} (Y - \beta_0 W)/(trace(\hat{Q}) - d - 1)} \Rightarrow \chi^2(1)$$

under the null hypothesis $H_0 : \beta = \beta_0$.

Moreover, since

$$W' H_X^{\hat{Q}} W/N = W' P_Z^{\hat{Q}} W/N + o_p(1) = W' P_Z^{Q} W/N + o_p(1) \xrightarrow{p} \frac{a^2 p^2 k}{1 - (1-p)(1-\alpha_{FS}^2)},$$

the probability of having a bounded AR confidence interval goes to one.

For the asymptotic lengths of the AR confidence interval, we have that

$$u' H_X^{\hat{Q}} W/\sqrt{N} = u' P_Z^{\hat{Q}} W/\sqrt{N} + o_p(1) = (u'QZ'/\sqrt{N})(Z'QQZ/N)^{-1}(Z'QW/N) + o_p(1)$$
$$u' H_X^{\hat{Q}} u = u' P_Z^{\hat{Q}} u - \frac{c}{trace(Q) - d - 1} u' M_X^{Q} u = (u'QZ'/\sqrt{N})^2 (Z'QQZ/N)^{-1} - c\sigma_u^2 + o_p(1)$$

Then the asymptotic length of the repeated split-sample AR confidence interval follows.

$$\begin{aligned}
\sqrt{N}|CI^{AR}| &= 2 \times \sqrt{\frac{(u' H_X^{\hat{Q}} W/\sqrt{N})^2 - (u' H_X^{\hat{Q}} u)(W' H_X^{\hat{Q}} W/N)}{(W' H_X^{\hat{Q}} W/N)^2}} \\
&= 2\sqrt{\frac{c\sigma_u^2}{W' H_X^{\hat{Q}} W/N}} + o_p(1) \\
&\xrightarrow{p} 2\mathsf{Z}_{\alpha/2} \frac{\sigma_u}{ap\sqrt{k}} \sqrt{(1 - (1-p)(1-\alpha_{FS}^2))}.
\end{aligned}$$

□

Proof for Lemma 7

Proof. Using Similar calculations as those in the proofs of Lemma 1, we can show that with the full sample,

$$\begin{aligned} W'H_XW = W'P_ZW - \frac{c}{N-d-1}W'M_XW &= \left[\frac{\epsilon'Z}{\sqrt{Z'Z}} + \rho\sqrt{Z'Z}\right]^2 - c\frac{W'M_XW}{N-d-1} \\ &= \left[\frac{\epsilon'Z/\sqrt{N}}{\sqrt{Z'Z/N}} + ap\sqrt{Z'Z/N}\right]^2 - c\frac{W'M_XW}{N-d-1} \\ &\overset{d}{=} \mathcal{Z}^2_{FULL} - c\left(\sigma_v^2 + a^2kp(1-p)\right) \end{aligned}$$

where $\mathcal{Z}_{FULL} \sim N\left(ap\sqrt{k}, \sigma_v^2 + a^2p(1-p)k'/k\right)$. The scalar $k' = E\left[Z_{ig}^4\right]$ is defined in Assumption 2.

Then the probability that the full sample AR confidence interval is bounded is

$$\begin{aligned} \mathsf{P}^{AR} = P\left(\mathcal{Z}^2_{FULL} > c(\sigma_v^2 + a^2kp(1-p))\right) &= 2 \times P\left(\mathcal{Z} > \frac{\sqrt{c(\sigma_v^2 + a^2kp(1-p))} - ap\sqrt{k}}{\sqrt{\sigma_v^2 + a^2p(1-p)k'/k}}\right) \\ &= 2\Phi\left(\frac{ap\sqrt{k} - \sqrt{c(\sigma_v^2 + a^2kp(1-p))}}{\sqrt{\sigma_v^2 + a^2p(1-p)k'/k}}\right) \end{aligned}$$

where $\mathcal{Z}$ is a random variable that follows standard normal distribution.

Similarily, we can derive the probability of having bounded infeasible selective AR confidence interval.

$$\begin{aligned} \mathsf{P}^{AR,INFSEL} = P\left(\mathcal{Z}^2_{INFSEL} > c\sigma_v^2\right) &= 2 \times P\left(\mathcal{Z} > \sqrt{c\frac{\sigma_v^2}{\sigma_v^2}} - a\sqrt{pk}/\sigma v\right) \\ &= 2\Phi\left(a\sqrt{pk}/\sigma_v - \sqrt{c}\right) \end{aligned}$$

where $\mathcal{Z}_{INFSEL} \sim N(a\sqrt{pk}, \sigma_v^2)$.

Since $k'/k = E[Z_{ig}^4]/k > E[Z_{ig}^2]^2/k = k$, the probability for the full sample AR is always

smaller than the probability for the infeasible AR as long as $0 < p < 1$. □

Proof for Lemma 11

Proof. The proof parallels the one for Lemma 6. First, the weighting matrix $\hat{Q} \xrightarrow{p} Q_{wk}$, where Q_{wk} is a $N \times N$ diagonal matrix with the $\sum_{l=1}^{g-1} n_g + i$-th diagonal element equaling to $S_{g,FS}$ if $g \in G_+$ and α_{FS} otherwise. Then by Lemma 19.24 of van der Vaart (1998), we can show that the 2SLS estimator is asymptotic normal and its $(1-\alpha)\times 100$ percent confidence interval length equals to $\frac{2\mathsf{Z}_{\alpha/2}\sigma_u}{\sqrt{(Z'Q_{wk}W)^{-1}(Z'Q_{wk}Q_{wk}Z)(Z'Q_{wk}W)^{-1}}} + o_p(1) = 2\mathsf{Z}_{\alpha/2}\frac{\sigma_u}{\sqrt{W'P_Z^{Q_{wk}}W}} + o_p(1)$. Calculations similar to that in proof for Lemma 1 show that

$$
\begin{aligned}
W'Q_{wk}Z/R &= Z'D_{\rho_g}Q_{wk}Z/R + o_p(1) = \sum_{g\in G_+} p_g E\left[Z_{ig}^2\right]\left(a/\sqrt{\frac{N}{R}}\right) S_{g,FS} + o_p(1) \\
&= \left(ak/\sqrt{\frac{N}{R}}\right)\frac{N}{R}\sum_{g\in G_+} p_g S_{g,FS} + o_p(1) \\
&= \left(ak\sqrt{\frac{N}{R}}\right)\sum_{g\in G_+} p_g S_{g,FS} + o_p(1), \\
Z'Q_{wk}Q_{wk}Z/R &= \frac{N}{R}\left(\sum_{g\in G_+} p_g E\left[Z_{ig}^2\right] S_{g,FS}^2 + (1-p)E\left[Z_{ig}^2\right]\alpha_{FS}^2\right) + o_p(1) \\
&= k\frac{N}{R}\left(\sum_{g\in G_+}\frac{n_g}{N} S_{g,FS}^2 + (1-p)\alpha_{FS}^2\right) + o_p(1), \\
W'P_Z^{Q_{wk}}W/R &= (W'Q_{wk}Z/R)^2(Z'Q_{wk}Q_{wk}Z/R)^{-1} \xrightarrow{p} \frac{a^2k\left(\sum_{g\in G_+} p_g S_{g,FS}\right)^2}{\sum_{g\in G_+} p_g S_{g,FS}^2 + (1-p)\alpha_{FS}^2}.
\end{aligned}
$$

The asymptotic length of the 2SLS confidence interval is then clear.

Also by Lemma 19.24 of van der Vaart (1998), we can show that the repeated split-sample

AR test

$$AR(\beta_0) = \frac{(Y - \beta_0 W)' P_Z^{\hat{Q}} (Y - \beta_0 W)}{(Y - \beta_0 W)' M_X^{\hat{Q}} (Y - \beta_0 W)/(trace(\hat{Q}) - d - 1)} \Rightarrow \chi^2(1)$$

under the null hypothesis $H_0 : \beta = \beta_0$.

Then since

$$W' H_X^{\hat{Q}} W/R = W' P_Z^{\hat{Q}} W/R + o_p(1) = W' P_Z^{Q_{wk}} W/R + o_p(1) = \frac{a^2 k \left(\sum_{g \in G_+} p_g S_{g,FS}\right)^2}{\sum_{g \in G_+} p_g S_{g,FS}^2 + (1-p)\alpha_{FS}^2} + o_p(1),$$

Since $a > 0$ and $S_{g,FS} > \alpha_{FS}$, the RHS is bounded away from zero. Therefore we show that the probability of having a bounded repeated split-sample AR confidence interval goes to one.

Also, we can derive that the asymptotic length of the repeated split-sample selective AR confidence interval

$$\begin{aligned}\sqrt{R}|CI^{AR,RSS}| &= 2\sqrt{\frac{c\sigma_u^2}{W' H_X^{\hat{Q}} W/R}} + o_p(1) \\ &= 2\mathsf{Z}_{\alpha/2} \frac{\sigma_u}{a\sqrt{k}\sum_{g \in G_+} p_g S_{g,FS}} \times \sqrt{\sum_{g \in G_+} p_g S_{g,FS}^2 + (1-p)\alpha_{FS}^2} + o_p(1)\end{aligned}$$

is the same as the asymptotic length of the repeated split-sample selective 2SLS confidence interval. □

References

Abadie, A., M. M. Chingos, and M. R. West (2014): "Endogenous Stratification in Randomized Experiments," *working paper*.

Acemoglu, D., S. Johnson, J. Robinson, and P. Yared (2008): "Income and Democ-

racy," *American Economic Review*, 98(3), 808842.

Anderson, T., and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. K., and J. H. Stock (2005): "Inference with Weak Instruments," *working paper*.

Benjamini, Y., and Y. Hochberg (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of Royal Statistical Society, B*, 57, 289–300.

Benjamini, Y., and Y. Hochberg (2000): "On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83.

Benjamini, Y., A. Krieger, and D. Yekutieli (2006): "Adaptive Linear step-up procedures that control the false discovery rate," *Biometrika*, 93, 491–507.

Berkowitz, D., M. Caner, and Y. Fang (2012): "The validity of instruments revisited," *Journal of Econometrics*, 166, 255–266.

Cai, T., and J. Jin (2007): "Estimating the Null and the Proportion of Non-Null Effects in Large-Scale Multiple Comparisons," *Journal of the American Statistical Association*, 102, 495–506.

Cao, H., W. Sun, and M. Kosorok (2013): "The Optimal Power Puzzle: Scrutiny of the Monotone Likelihood Ratio assumption in Multiple Testing," *Biometrika*, 100, 495–502.

Cervellati, M., F. Jung, U. Sunde, and T. Vischer (2014): "Income and Democracy: Comment," *American Economic Review*, 104(2), 707–719.

Conley, T., C. Hansen, and P. Rossi (2012): "Plausibly Exogenous," *the Review of Economics and Statistics*, 94, 260–272.

Dufour, J.-M. (1997): "Some Impossiblity Theorems in Econometrics, with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1389.

Efron, B., R. Tibshirani, J. Storey, and V. Tusher (2001): "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Genovese, C., and L. Wasserman (2002): "Operating Characteristic and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society, B*, 64, 499–517.

Gu, J., and S. Shen (2014): "Multiple Testing for Positive Treatment Effects," *working paper*.

Guggenberger, P. (2012): "On The Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption," *Econometric Theory*, 28, 387–421.

Lleras-Muney, A. (2002): "Were State Laws on Compulsory Education Effective? An analysis from 1915 to 1939," *Journal of Law and Economics*, 45.

Lleras-Muney, A. (2005): "The Relationship Between Education and Adult Mortality in the United States," *The Review of Economic Studies*, 72, 189–221.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Sructural Models," *Econometrica*, 71, 1027–1048.

Oreopoulos (2006): "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter," *American Economic Review*, 96(1), 152–175.

Oreopoulos, P., M. E. Page, and A. H. Stevens (2006): "The Intergenerational Effects of Compulsory Schooling," *Journal of Labor Economics*, 24(4), 729–760.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer.

Rosalsky, A., and M. Sreehari (1998): "On the Limiting Behavior of Randomly Weighted Partial Sums," *Statistics and Probability Letters*, 40, 403–420.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 1997.

Stephens, M., and D.-Y. Yang (2014): "Compulsory Education and the Benefits of Schooling," *American Economic Review*, 104(6), 1777–1792.

Storey, J. D. (2002): "A direct approach to False Discovery Rates," *Journal of the Royal Statistical Society*, 64, 479–498.

——— (2003): "The positive False Discovery Rate: a Bayesian interpretation and the q-value," *Annals of Statistics*, 31, 2013–2035.

Sun, W., and T. T. Cai (2007): "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal of American Statistical Association*, 102, 901–912.

van der Laan, M., and S. Dudoit (2007): *Multiple Testing Procedures with Applications to Genomics*. Springer.

van der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press.