Power and Size Analysis in Two Stage Least Squares (TSLS) Models with Weak Instruments.

Chienhsiang Yeh Supervisor: Juergen Meinecke

October 25, 2018

Abstract

The estimators of instrument variables (IV) regression may be non-normal when IV only has a low correlation with endogenous explanatory variables. Therefore, the traditional inference may be unreliable. This project studies the distribution of IV estimator and then, in particular, focuses on the power and size of test with weak IV. To correct the inference under weak IV, we simulate Anderson-Rubin test. We find that the size of AR-test improves significantly so that we can correctly reject the null when it is true, while the power may be much smaller in small sample size such that we may more likely to make Type II error. We further investigate more extreme cases of many IV and large sample size and then conclude that AR test is generally better than t-test.

1 Introduction

Empirically, economists often have the endogenous data which generate bias in ordinary least squares (OLS) model. One popular methodology for solving inconsistency is instrumental variables (IV) regression. To implement IV estimators, we usually do two-stage least squares(TSLS) process. Moreover, the conditions for IV are that they must be independent of error term and correlated with the endogenous variables. However, in practice, we may include weak IV such that they are insignificantly correlated with endogenous variables. For instance, Angrist & Krueger (1991) regress the earnings on years of education and used the season of birth as IV. They discover that the season of birth is weakly correlated with the year of education, so the estimated returns on education is similar in OLS and TSLS regression, suggesting that TSLS estimator has bias. They further point out that applied researchers use weak IV all the time.

Moreover, Staiger & Stock (1997) points out that the asymptotic distribution of estimators is non-normal under weak IV. Thus, the inference breaks down even in large sample size. There are some methodologies for this issue. For example, Anderson & Rubin (1949) provides Anderson-Rubin statistic for any number of weak instruments, Moreira (2003) provides conditional likelihood ratio test for one IV and Kleibergen (2002) presents an alternative Lagrange multiplier test.

This project implements Monte Carlo method to simulate simple regression models under weak IV to confirm the shutdown of inference, inconsistency and non-normal distribution. We examine one weak IV case in small sample size first and then imitate the specification in the paper of Angrist & Krueger (1991) which has many IV and large sample size to confirm that the problem exists in empirical works.

The result of TSLS estimators' distribution proves that the bias exists and the distribution is non-normal especially under weak IV. For statistical inference, we simulate t-test and the alternative Anderson-Rubin (AR) test. We find that the size of t-tests in small sample are too large under any number of weak IV. Additionally, the simulation of AR-statistic confirms that it can correct the size of tests while there is some trade-off for which the power in small sample may be smaller than that of t-test. Fortunately, we can also improve the power by larger sample size. Furthermore, in the extreme but realistic specification, we find that t-test has much larger size under many weak IV which can hardly be improved by increasing observations; while, we still can improve the power of AR test by increasing observations. Then, we conclude that AR-test is generally more reliable if sample size is much greater than the number of IV.

2 IV Regression and Weak Instrument Variables

2.1 IV Regression and Two Stage Least Squares

Consider a linear model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k} + u_i, \quad i = 1, ..., N.$$

Let x_k be a endogenous variable such that $Cov(x_k, u) \neq 0$. To construct consistent estimators, we apply instrument variables(IV) regression to solve endogenous problems. A random variable z_m is said to be an instrument variable if

1. relevance: $\mathbb{E}[z_m x_k] \neq 0$.
2. exogeneity: $\mathbb{E}[z_m u_i] = 0$

The relevance condition states that the correlation between instruments and endogenous variables should be greater than zero. Also, instruments must be exogenous such that they are uncorrelated with error term. Let $z_1, z_2, ..., z_m$ be instruments. Then, the column vector of exogenous variables is $z = (1, x_1, x_2, ..., x_{k-1}, z_1, z_2, ..., z_m)'$. Since $E(z_m u_i) = 0$, $\mathrm{E}(z'u) = 0$ where $u = (u_1, u_2, ..., u_N)'$. Thus, $\mathrm{E}(z'y) = \mathrm{E}(z'x)$ and then by the Large Number Theory the IV estimator is

$$\hat{\beta}_{IV} = \Big(\frac{1}{N}\sum_{i=1}^{N} z_i' x_i\Big)^{-1}\Big(\frac{1}{N}\sum_{i=1}^{N} z_i' y_i\Big).$$

It is well known that this estimator can be implemented by the process of Two Stage Least Squares(TSLS). TSLS is a two stages of linear regression:

$$\begin{aligned} y_i &= x_{i,-k}'\beta_{-k} + x_{i,k}\beta_k + u_i \\ x_{i,k} &= z_i'\pi + v_i, \end{aligned}$$

where $x_{i,-k} = (1, x_1, x_2, ..., x_{k-1})'$ and $\beta_{-k} = (\beta_0, \beta_1, ...\beta_{k-1})'$. The TSLS estimator is

$$\hat{\beta}_{TSLS} = \Big(\frac{1}{N}\sum_{i=1}^{N} \hat{x}_i'\hat{x}_i\Big)^{-1}\Big(\frac{1}{N}\sum_{i=1}^{N} \hat{x}_i' y_i\Big),$$

where $\hat{x}_{i,k} = z_i'\hat{\pi}$ is the predicted regressor from the first stage regression. Wooldridge (2010) shows that $\hat{\beta}_{TSLS}$ is unbiased and asymptotically normal under the following assumptions,

$$\begin{aligned} &\text{T1: } \mathrm{E}(z_i' u_i) = 0 \ \text{ for all i} \\ &\text{T2: rank } \mathrm{E}(z_i' z_i) = k + m \quad \& \quad \text{rank } \mathrm{E}(z_i' x_i) = k \\ &\text{T3: } \mathrm{E}(u_i^2 z_i' z_i) = \sigma^2 \mathrm{E}(z_i' z_i), \ \text{where } \sigma^2 = \mathrm{E}(u_i^2) \end{aligned}$$

Assumption T1 is exactly the exogenous condition of IV. T2 guarantees that the estimators of first and second stage exist(prevent multicollinearity). Then, under these assumptions,

$$\sqrt{N}(\hat{\beta}_{TSLS} - \beta) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2[\mathbb{E}(z'x)]^{-1}\mathbb{E}(z'z)[\mathbb{E}(z'x)]^{-1}\big)$$

Since TSLS is consistent and asymptotically normal, researchers use this estimator under endogeneity for empirical works and inference.

2.2 Weak IV

Suppose that the relevance condition of IV does not hold. Then, IV are slightly correlated with endogenous variables. More specifically, we define weak IV as

$$\mathbb{E}[z_i x_i] \approx 0.$$

Now, could we get consistent estimators and correct inference under such weak IV? To obtain the intuition, consider the simple regressions with weak IV z,

$$\begin{aligned}
&y_i = \beta x_i + u_i \\
&x_i = \pi z_i + v_i, \qquad \text{for all i} \\
&\mathbb{E}[z_i x_i] \approx 0 \\
&\text{where } x_i, y_i \text{ and } z_i \text{ are all scalar.}
\end{aligned}$$

Fort this scalar regressor and scalar instrument, the weak IV can be identified if the squared correlation coefficient $r^2_{x,z}$ is small. More generally, for a vector of IV, if the coefficient of determination R^2 or the F-statistic for test $H_0 : \pi = 0$ in first stage regression is small, then we identify that IV are weak. Since R^2 is the measure of goodness for the first stage regression, we further define R^2 as the strength of instrument. Under this specification, the estimator is

$$\hat{\beta}_{TSLS} = (\frac{1}{N}\sum_{i=1}^{N} z_i x_i)^{-1}(\frac{1}{N}\sum_{i=1}^{N} z_i y_i).$$

The estimator converges in probability to

$$\hat{\beta}_{TSLS} \xrightarrow{p} \beta + \frac{Cov(z,u)}{Cov(z,x)}.$$

Suppose that $Cov(z, u)$ is close to zero but not exactly zero (this is reasonable for empirical work). Given a weak IV z, z is weakly correlated with x and then $Cov(z, x)$ is also close to zero. Then, the magnitude of the second term,

$$\frac{Cov(z,u)}{Cov(z,x)},$$

is not zero and cannot be ignored such that there exists a bias. Moreover, note that the bias, $Cov(z, u)/Cov(z, x)$, is a function of the degree of endogeneity because the bias may be neglectable as $Cov(z, x)$ increases.

Further, the bias may be larger than the OLS estimator. The OLS estimator is

$$\hat{\beta}_{OLS} = (\frac{1}{N}\sum_{i=1}^{N} x_i^2)^{-1}(\frac{1}{N}\sum_{i=1}^{N} x_i y_i).$$

Thus, it converges in probability to

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta + \frac{Cov(x,u)}{Cov(x,x)}.$$

Then the ratio of bias between TSLS and OLS estimators is

$$\frac{\hat{\beta}_{TSLS} - \beta}{\hat{\beta}_{OLS} - \beta} \xrightarrow{p} \frac{Cov(z,u)}{Cov(z,x)}\frac{Cov(x,x)}{Cov(x,u)} = \frac{Corr(z,u)}{Corr(z,x)}\frac{1}{Corr(x,u)}.$$

So, the bias of $\hat{\beta}_{TSLS}$ may be greater than the bias of $\hat{\beta}_{OLS}$ if

$$Corr(z,u) > Corr(z,x)Corr(x,u).$$

In additoin, Staiger & Stock (1997) prove that the asymptotic distribution of $(\hat{\beta}_{TSLS} - \beta)$ is a ratio between two random variables which are both standard normally distributed and correlated to each other. Since the asymptotic distribution is non-normal, the hypothesis testing would fail.

3 Anderson-Rubin test

Consider again the model in section 2.1

$$\begin{aligned} y = x\beta &= x_{-k}\beta_{-k} + x_k\beta_k + u \\ x_k &= z\pi + v, \end{aligned}$$

and test

$$\begin{aligned} H_0 &: \beta = \gamma \\ H_1 &: \beta \neq \gamma. \end{aligned}$$

The procedure of Anderson-Rubin (AR) test is

(i) calculate $y^* = y - \gamma x_k$

(ii) regress y^* on instrument z

$$y^* = z\xi + w,$$

where w is the error term and ξ is the coefficient.

(iii) test if the coefficients on IV are all zero,

$$H_0 : \xi = \mathbf{0}$$

$$H_1 : \xi \neq \mathbf{0}$$

This test is equivalent to the test of $H_0 : \beta = \gamma$. It is a F-test in small samples.

Recall that assumption T1, $Cov(z,u) = 0$. Then, under H_0,

$$\mathrm{E}(y^* z) = \mathrm{E}((x_{-k} + u)z) = 0, \quad \text{where } x_k \perp\!\!\!\perp z.$$

Thus, y^* is uncorrelated with IV and the corresponding coefficients are zero under H_0. The intuition is that γx_k is the endogenous part under the null, and the remaining part y^* is then exogenous so that it does not correlate with the error. Additionally, we can reorganize the equations,

$$y^* = x_{-k}\beta_{-k} + (\beta_k - \gamma)\pi z + (\beta_k - \gamma)v + u.$$

Then, we use $\hat{\xi} = (\hat{\beta}_k - \gamma)\hat{\pi}$ for AR-statistic instead of biased TSLS estimator $\hat{\beta}_{k,TSLS}$ for t-statistic. That is, $\hat{\xi}$ or AR-statistic has more sufficient information for inference under the null. Moreover, the asymptotic distribution of AR-statistic at the null is a χ^2_s distribution of some non-central parameters, where s is the number of IV. Then, AR-statistic is a F-statistic in small sample size.

4 Hypothesis testing and Power Function

Hypothesis testing is a procedure for selecting between statistical models. Suppose that we have a simple linear regression model on a set of random variables (x, y),

$$y = a + \beta x + \varepsilon.$$

A statistical hypothesis is a proposed statement that narrows or restricts the model. This statement is also defined as the null hypothesis, H_0, which claims the distribution of random variables. Against the restricted model, there is a broader claim called alternative hypothesis, H_1, which rivals the null hypothesis. For instance, the null hypothesis could be $\beta = 0$ and then the compared alternative hypothesis is $\beta \neq 0$. The notation is

$$\begin{aligned} H_0 &: \beta = 0, \\ H_1 &: \beta \neq 0. \end{aligned}$$

Since the null hypothesis and the alternative hypothesis are rivals and exclusive, only one of them is true. To determine which hypothesis is true, a test is defined as a rule to accept or reject the null. More specific, a test assigns a rejection region and an acceptance region for a test statistic which derived from the assumed distribution under the null. Note that the rejection region and the acceptance region are also exclusive and complementary to each other in the sample space, since they are the spaces for the null and the alternative respectively. Here, the test statistic is

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}},$$

where $s_{\hat{\beta}}$ is the standard error of the estimator. Besides, H_0 is rejected if $t \geq t_{(0.975,N-1)}$ and $t \leq t_{(0.025,N-1)}$ with significance level 0.05.

If the null hypothesis is true, the relevant test statistic is a t-distribution with $(n-1)$ degree of freedom given that the error ε is normally distributed. If the null is false, the test statistic would be invalid or extreme such that it falls in the rejection region under the proposed null hypothesis or t-distribution.

Since a test statistic is a probability distribution and the estimators are random variables, we can find its confidence interval with some significance level. Thus, it is possible that we do an incorrect inference and generate a false result. The following errors are defined:

Type I error: the null is true, but it is rejected.

Type II error: the null is false, but it is accepted.

The probability of correctly rejecting the null hypothesis is called the power of a test. Then, when H_1 is true, one subtracts the probability of Type II error is the power of a test. Furthermore, the size or significance level of a test is defined by the probability of making Type I error. Therefore, the value of power function is equal to the size when H_0 is true. The textbook example of power function is presented in figure 1, which is the t-test for H_0: $\beta = 0$ under strong IV and large sample size of 5000 with exogenous independent variable. As the above discussion, the power function is one under H_1, $\beta \neq 0$, and 0.05 under H_0: $\beta = 0$.

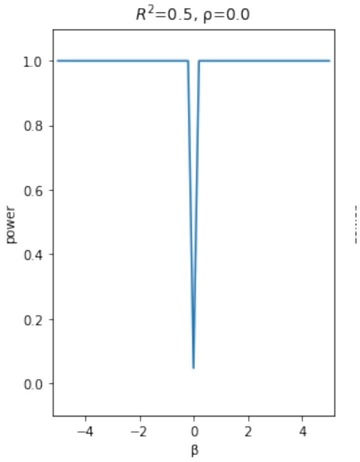


Figure 1: The textbook power function for t-test, under sample size of 5000 and strong IV

5 Numerical Study

5.1 Data Generating Process

To begin with, the data are randomly generated for Monte Carlo simulation setup using the simple regression model without intercepts,

$$\begin{aligned} y_i =& \beta x_i + u_i \\ x_i =& \pi z_i + v_i, \\ z_i \sim \mathcal{N}(0,1),\ (u,v) \sim& \mathcal{N}(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}) \quad \text{for all i,} \end{aligned}$$

where $\rho = Corr(x,u) = Corr(u,v)$ and it is also defined as the degree of endogeneity.

Given that R^2 is the coefficient of determination for the first stage regression, R^2 presents the strength of IV. For this simple regression,

$$\begin{aligned} R^2 &= (Corr(z,x))^2 = (\frac{Cov(x,z)}{\sigma_z \sigma_x})^2 = (\frac{Cov(\pi z + v, z)}{\sigma_z \sigma_x})^2 \\ &= \frac{\sigma_z^2}{\sigma_x^2}\pi^2 = \frac{1 \times \pi^2}{Var(x)} = \frac{\pi^2}{Var(\pi z + v)} = \frac{\pi^2}{\pi^2 + 1}. \end{aligned}$$

Then, we let

$$\pi = \sqrt{\frac{R^2}{1 - R^2}}$$

to set up the degree of weakness for IV. Next, we implement this model with all combinations of following specifications.

- degree of endogeneity: $\rho \in \{0.00, 0.50, 0.90, 0.99\}$
- strength of instrument: $R^2 \in \{0.01, 0.05, 0.10, 0.20, 0.50\}$
- sample size: $n \in \{100, 200, 500, 1000, 2000\}$
- the coefficient of endogenous variable β: $\beta \in [-5, 5]$

For each combination, we generate 2000 samples(or 2000 runs) of size n data and then estimate $\hat{\beta}_{TSLS}$ to find the bias and distribution. For hypothesis testing, we compute t-statistic to test the null

$$H_0 : \beta = 0.$$

Here, H_0 is generally rejected if t-statistic is larger than 1.96 or less than -1.96 at significance level 5%. Note that the data may be generated under H_1 for the combinations which $\beta \neq 0$,

$$H_1 : \beta \neq 0 \text{ or } \beta = b,\ b \in [-5, 0) \cup (0, 5].$$

5.2 Empirical Power and Size

For each replication of sample, we may or may not reject the null. In Monte Carlo method, the power under H_1 is calculated as

$$\text{power} \equiv \frac{\text{the number of replications for which } H_0 \text{ is rejected}}{\text{total number of replications}},$$

while the size of test under H_0 (given H_0 is true) is

$$\text{size} \equiv \frac{\text{the number of replications for which statistic is in rejection region}}{\text{total number of replications}}.$$

As discussed in section 4, the size is equal to the power function when H_0 is true, so we simply calculate the power at $\beta = 0$ as the size. Thus, if the data is generated under the null, we expect that the power is around 0.05 which is equal to size of test at a 5% significance level. Additionally, if the data is generated under H_1, the power should close to one.

5.3 Result

5.3.1 The bias and distribution of $\hat{\beta}_{TSLS}$

Initially, we simulate the distribution of $\hat{\beta}_{TSLS}$ under small sample size and one IV. Figure 2 shows the result of one weak IV with sample size $n = 100$ and $\beta = 0$ under all degree of endogeneity. It indicates that there exists bias ($\mu \neq 0$) and the distributions are non-normal for the cases of weak IV. Note that the distribution is independent of the degree of endogeneity. Furthermore, the pattern is similar to the simulated distribution of the ratio between two normal distributed random variables (figure 4). This is the evidence for which the asymptotic normality of $(\hat{\beta}_{TSLS} - \beta)$ breaks down. Moreover, as the sample size n increases, the bias converges to zero and the distribution is more close to normal (see figure 3). It seems that this result conflicts the theory in section 2.2. However, there exits some skewness and this simulation is the simplest model with only one endogenous variable and one IV, so the theory still could be true generally.

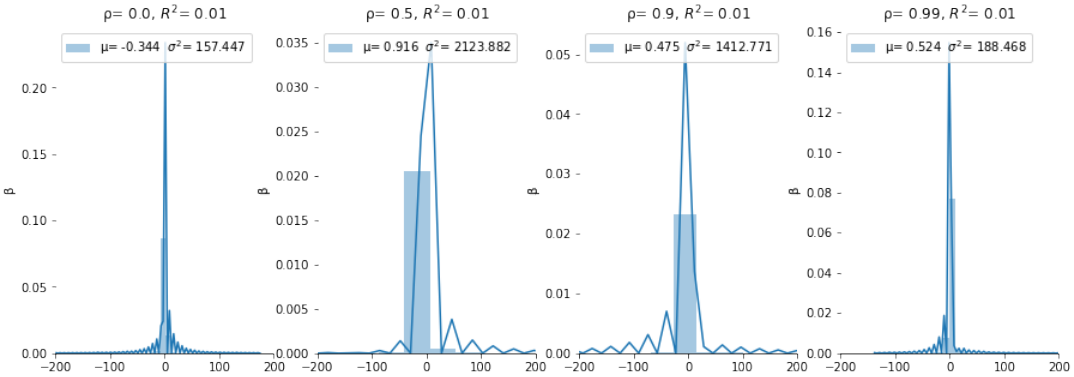


Figure 2: The distribution of $\hat{\beta}_{TSLS}$ for $R^2 = 0.01$, $n = 100, \beta = 0$.

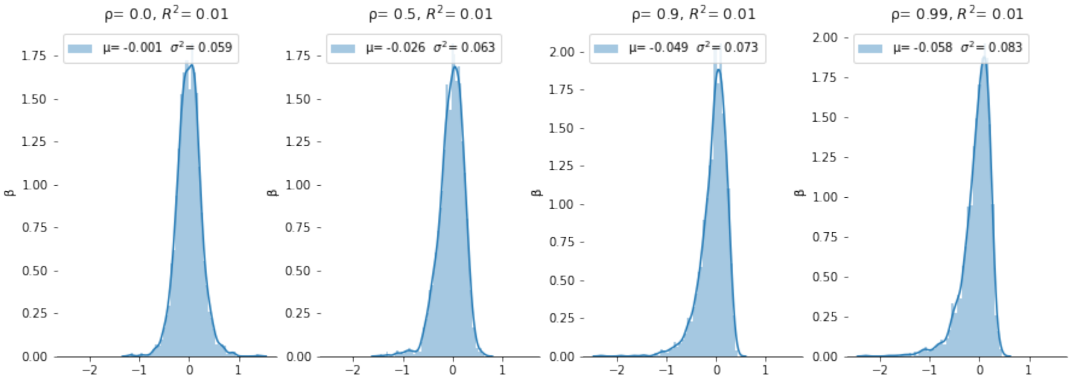


Figure 3: The distribution of $\hat{\beta}_{TSLS}$ for $R^2 = 0.01$, $n = 2000, \beta = 0$.

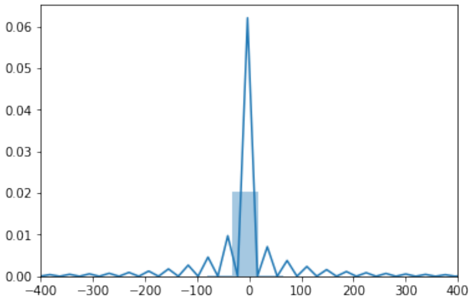


Figure 4: The simulated distribution of two standard normal distributed random variables.

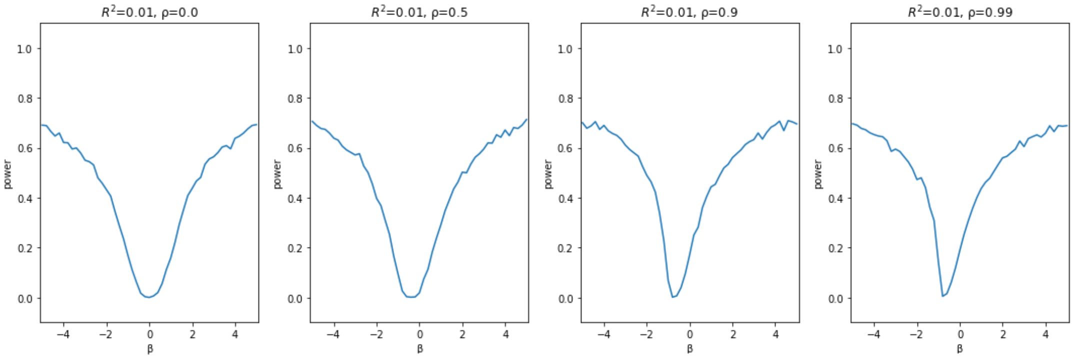


Figure 5: The power of t-test for sample size of 100

Table 1: Size table for t-test.

(a) sample size $n = 100$

		Exogenous			Endogenous
		$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.90$	$\rho = 0.99$
Weak IV	$R^2 = 0.01$	0.0005	0.0190	0.1705	0.1880
	$R^2 = 0.05$	0.0060	0.0460	0.0950	0.1085
	$R^2 = 0.1$	0.0135	0.0405	0.07457	0.0995
	$R^2 = 0.2$	0.0305	0.0455	0.0685	0.0705
Strong IV	$R^2 = 0.5$	0.0550	0.0580	0.0565	0.0610

(b) The worst case, $R^2 = 0.01$, for each sample size.

		Exogenous			Endogenous
	sample size	$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.90$	$\rho = 0.99$
	$n = 100$	0.0005	0.0190	0.1705	0.1880
Weak IV	$n = 200$	0.0035	0.0350	0.1165	0.1520
$R^2 = 0.01$	$n = 500$	0.0040	0.0345	0.0875	0.1025
	$n = 1000$	0.0105	0.0460	0.0775	0.0970
	$n = 2000$	0.0585	0.0515	0.0385	0.0480

5.3.2 The power and size for the test $H_0 : \beta = 0$

Next, we simulate the power function to check that the inference is indeed unreliable when IV are dramatically weak. Figure 5 shows the worst result with small sample size $n = 100$ and weak IV $R^2 = 0.01$. In the figure, the degree of endogeneity increases as from left to right. Comparing to figure 1, the power function shifts down so that it has smaller power and higher probability of failing to reject the null when, say, $\beta = 4$. Besides, it shifts to left as the degree of endogeneity increases. This is reasonable since the bias is positive rather than negative, $\hat{\beta}_{TSLS} - \beta > 0$, in figure 2.

The size of tests under one IV is summarized in table 1. Table 1a shows that when IV is weak ($R^2 = 0.01$) and the degree of endogeneity is high($\rho = 0.99$), the size is greater than 0.15 under small sample size $n = 100$. Table 1b indicates that the size approximates to 0.10 even with the sample size of 1000. These large sizes, which are greater than 5%, suggest that we may fail to accept the null when IV are weak. Moreover, table 1b summarizes that the size is decreasing in sample size, so this problem can be improved by large sample size.

5.3.3 AR test Numerical Result

The worst case of power function for AR-test simulation is presented in figure 6 of which IV is weak and sample size is small. It is clear that the power is asymmetric under high degree of endogeneity. When β is much greater or less than the null value, say $\beta = 4 > 0$, the power is around or less than 0.2 so that there is a high probability of 80% for falsely accepting the null. Also, the power at $\beta = -1.5$ is greater than that of $\beta = 4$ (the rightest plot of figure 6). These results agree with Marmer (2017) which proves that the finite probability of detecting big deviation from the null, small $\beta - 0$, would cause the power at small deviation to surpass the power at large deviation. Nevertheless, the power at $\beta = 0$ or the size is close to 0.05 for all strength of IV, all degree of endogeneity and all sample size (see table 2). Thus, the probability of type I error is corrected although the probability of type II error may increases. Moreover, figure 8 and figure 7 under large sample size indicate that we can improve the power by increasing the sample size even though the power may slightly worse than that of t-test when the deviation from null (for instance, $\beta - 0 = 0.5$) is small.

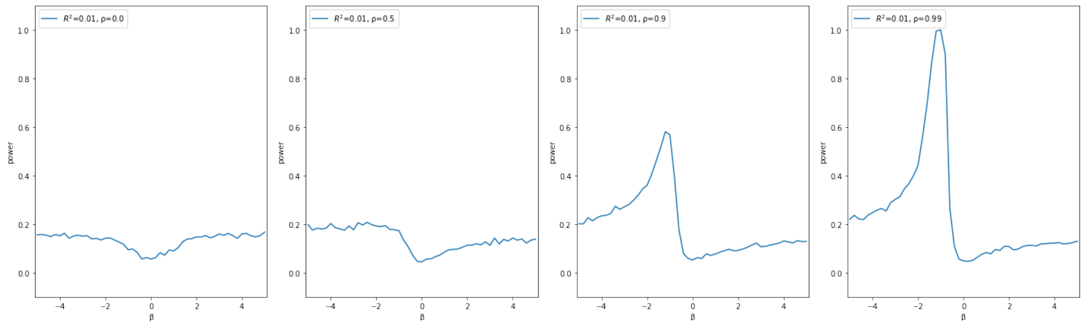


Figure 6: The power of AR-test under one weak IV $R^2 = 0.01$ and small sample size $n = 100$.

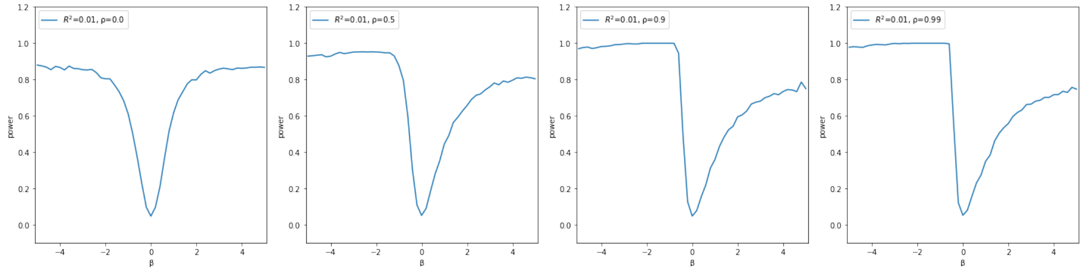


Figure 7: The power of AR-test under one weak IV $R^2 = 0.01$ and large sample size $n = 1000$.

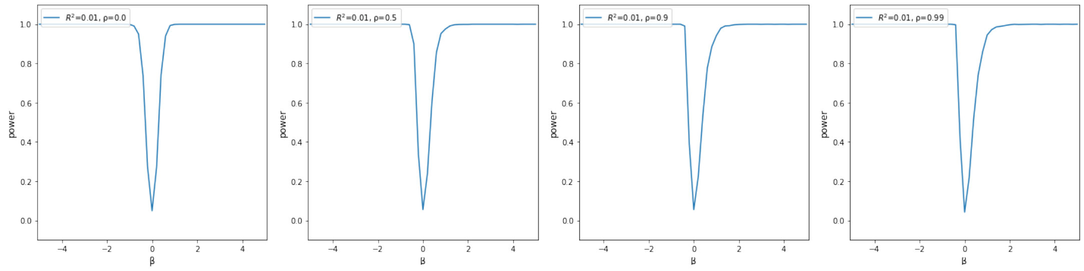


Figure 8: The power of AR-test under one weak IV $R^2 = 0.01$ and large sample size $n = 5000$.

Table 2: Size table for AR-test.

(a) Sample Size $n = 100$

		Exogenous			Endogenous
		$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.90$	$\rho = 0.99$
Weak IV	$R^2 = 0.01$	0.0580	0.0545	0.0565	0.0415
	$R^2 = 0.05$	0.0535	0.0560	0.0560	0.0555
	$R^2 = 0.1$	0.0570	0.0505	0.0500	0.0575
	$R^2 = 0.2$	0.0425	0.0525	0.0525	0.0610
Strong IV	$R^2 = 0.5$	0.0515	0.0500	0.0585	0.0470

(b) The worst case, $R^2 = 0.01$, for each sample size.

		Exogenous			Endogenous
	sample size	$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.90$	$\rho = 0.99$
	$n = 100$	0.0580	0.0545	0.0565	0.0415
Weak IV	$n = 200$	0.0610	0.0510	0.0430	0.0480
$R^2 = 0.01$	$n = 500$	0.0485	0.0420	0.0520	0.0520
	$n = 1000$	0.0460	0.0515	0.0600	0.0475
	$n = 2000$	0.0425	0.0375	0.0420	0.0495

5.3.4 More Extreme Cases of Many Weak IV under the Specifications in Angrist & Krueger (1991)

For cases of more weak IV, we set the model as section 5.1 with k IV,

$$\begin{aligned} y_i =& \beta x_i + u_i \\ x_i =& \pi z_{1,i} + \pi z_{2,i} + ... + \pi z_{k,i} + v_i, \\ \pi =& \sqrt{\frac{R^2}{k(1-R^2)}} \\ z_{1,i}, z_{2,i}, ..., &\text{ and } z_{k,i} \sim \mathcal{N}(0,1) \text{ and are i.i.d..} \end{aligned}$$

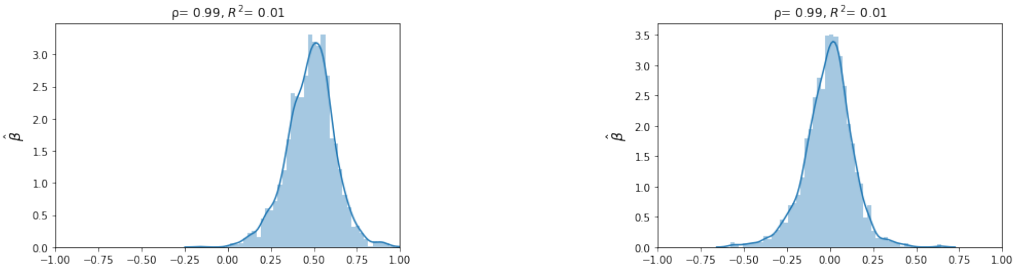
where all the other variables are assigned as section 5.1.

We simulate the power and size under this model and the specifications in Angrist & Krueger (1991) to imitate the more extreme empirical works. This extreme example has large sample size and many number of weak instruments. Angrist & Krueger (1991) have almost 200,000 observations and 200 instruments in their study. To set up doable specifications in computer, we have a compromise and focus on the following parameters,

- number of instruments: $k \in \{5, 10\}$
- sample size: $n \in \{500, 1000\}$

The distribution of $\hat{\beta}_{TSLS}$ under 10 IV and sample size of 1000 is presented in figure 9. When β is zero (figure 9a), the mean of $\hat{\beta}_{TSLS}$ is greater than zero; while, $\hat{\beta}_{TSLS}$ has a mean around zero if $beta = -0.48$ (figure 9b). This suggest that $\hat{\beta}_{TSLS}$ has a positive bias.

Figure 10 shows the result of 10 weak IV $R^2 = 0.01$ and sample size of 1000. The power is asymmetric around $\beta = 0$ when explanatory variable is endogenous, $\rho \geq 0.9$. That is, comparing to 1, the power function shifts to left as the strength of endogeneity increases. The lowest power



(a) $\beta = 0$.

(b) $\beta = -0.48$.

Figure 9: Distribution of $\hat{\beta}_{TSLS}$ under 10 weak IV, $R^2 = 0.01$, sample size of 1000

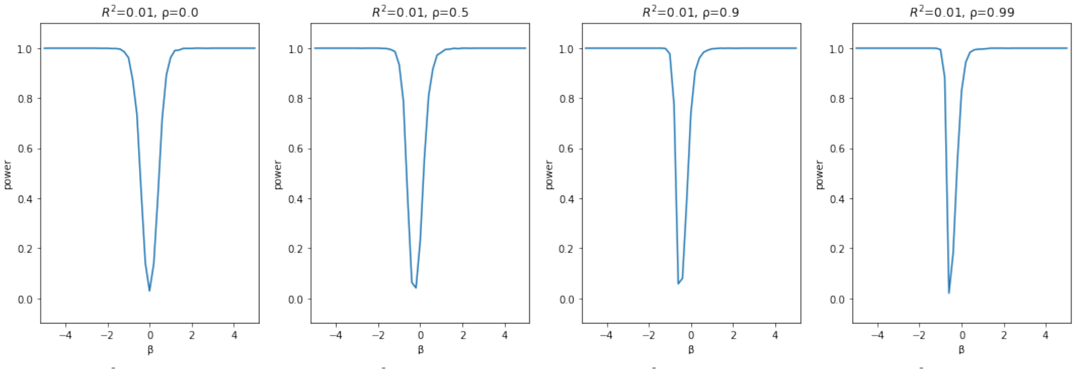


Figure 10: The power under 10 weak IV with $R^2 = 0.01$, sample size of 1000.

under $\rho = 0.99$, the rightest plot, is at around $\beta = -0.48$ for which $\hat{\beta}_{TSLS}$ has mean of zero in figure 9b. Then, the positive bias explains the left shift of power function. Moreover, when the true β is zero, the figure indicates that the power is around 0.9 under highly endogeneity, $\rho = 0.99$. Yet, the power function does shift down versus the case of small sample size and one IV.

The size of test with sample size 1000 and five or ten IV are summarized in table 3. We can observe few results from the table. Firstly, the size are generally greater than that of one IV given $R^2 \leq 0.2$ and $\rho \geq 0.5$. In the most extreme instance, there is 85% of probability such that we falsely reject the null under $\rho = 0.99$ and $R^2 = 0.01$. Second, the size grows significantly as the number of IV increases from five to ten. Third, even with strong IV $R^2 = 0.2$, the size is greater than 5% if we include ten weak IV. Fourth, the size increases exponentially in ρ. This is reasonable since the bias is increasing in the degree of endogeneity. Finally, even in large sample size of 1000, the size is relatively high under weak IV. Recall that under one IV, the size of t-test can be improved by large sample size. However, under 10 IV, the size improvement of increasing sample size is insignificant.

Table 3: Size table of t-test under many IV.

(a) 5 IV, Sample Size $n = 1000$

	ρ									
	Exogenous								Endogenous	
R^2	0.00	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99
Weak 0.01	0.0230	0.0325	0.0555	0.0935	0.1330	0.1600	0.2400	0.2865	0.361	0.4270
0.05	0.0435	0.0520	0.0515	0.0535	0.0715	0.0830	0.1020	0.0975	0.129	0.1360
0.1	0.0355	0.0495	0.0440	0.0540	0.0475	0.0645	0.0665	0.0805	0.075	0.1040
0.2	0.0475	0.0510	0.0540	0.0490	0.0635	0.0545	0.0555	0.0535	0.057	0.0675
Strong 0.5	0.0530	0.0475	0.0470	0.0490	0.0415	0.0535	0.0545	0.0545	0.052	0.0580

(b) 10 IV, Sample Size $n = 1000$

	ρ									
	Exogenous								Endogenous	
R^2	0.00	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99
Weak 0.01	0.0300	0.0515	0.1050	0.1645	0.2625	0.3390	0.4745	0.6075	0.7375	0.8500
0.05	0.0455	0.0520	0.0680	0.0885	0.1135	0.1485	0.1855	0.2415	0.2685	0.3100
0.1	0.0505	0.0570	0.0560	0.0755	0.0900	0.0935	0.1185	0.1255	0.1730	0.2065
0.2	0.0500	0.0505	0.0550	0.0660	0.0655	0.0785	0.0755	0.0970	0.0990	0.1060
Strong 0.5	0.0515	0.0490	0.0395	0.0500	0.0475	0.0520	0.0600	0.0520	0.0670	0.0680

5.3.5 AR test under Many IV

We further simulate the AR test under the above more extreme specifications and the following the model

$$\text{regress } (y - 0 \times x) \text{ on } z_1, z_2, z_3, ..., z_k.$$

The statistic is F-statistic of F-test

$$H_0 : \text{all coefficients of } z_1, z_2, ..., \text{ and } z_k \text{ are zero.}$$

We first focus on the power function under the weak IV. Figure 11 presents the power of AR test with sample size of 1000. The power is asymmetric and less than 0.4 when $\beta = 4$. The probability of type II error is relative high comparing that of one IV. Generally, it has worse power under 10 IV than that of one IV in figure 7 given the same large sample size. However, the size is close to zero under the null. We then increase the sample size further. Figure 12 shows the power function under 10 IV and large sample size of 5000. The power is close to one if $|\beta|$ is large enough, say, two. Then, we can conclude that the trade-off is insignificant in much larger sample size. Note that in figure 12, the power is still worse than the textbook example as $0 < \beta < 2$. We can improve this further by increasing the sample size such that it is more than 1000 times as many as the number of IV.

We further examine the size table. Table 4 shows the size of AR statistic for the most extreme case with 10 IV and sample size $n = 1000$. For other specifications, the results are similar. It is obvious that the size are all around 0.05. Thus, we can say that AR test reduces the probability of type I error to 0.05 at significance level 5%.

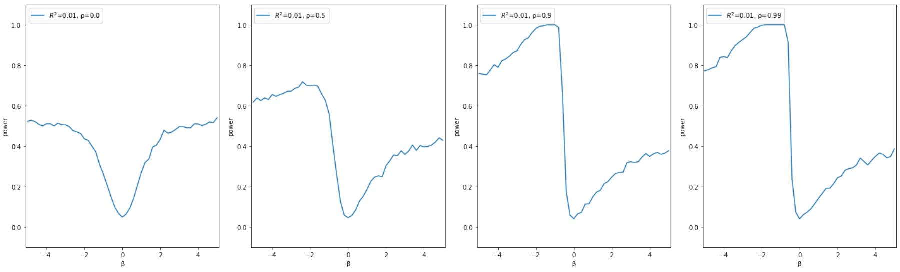


Figure 11: The power for AR-test with $R^2 = 0.01$, sample size of 1000 and 10 IV.

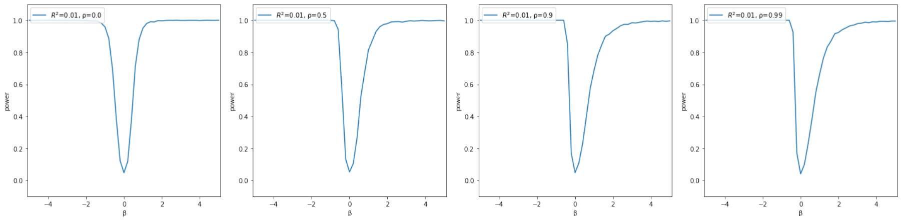


Figure 12: The power for AR-test with $R^2 = 0.01$, sample size of 5000 and 10 IV.

Table 4: Size table for AR-statistic with ten IV and sample size $n = 1000$.

		Exogenous			Endogenous
		$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.90$	$\rho = 0.99$
Weak IV	$R^2 = 0.01$	0.0540	0.0535	0.0420	0.0495
	$R^2 = 0.05$	0.0400	0.0460	0.0495	0.0450
	$R^2 = 0.1$	0.0445	0.0440	0.0535	0.0500
	$R^2 = 0.2$	0.0670	0.0495	0.0550	0.0535
Strong IV	$R^2 = 0.5$	0.0500	0.0495	0.046	0.0495

5.4 Practical Guide

In practice, if we have dramatically many observations and meany weak IV as Angrist & Krueger (1991), then I suggest use AR-test for inference. First, AR-test is easy to implement. If AR-statistic gives different result versus t-test, then we would discover the existence of bias. Then, we may consider to forgo some weak IV which can be identified by F-test in 2.2. If we do not remove weak IV, we can go back and trust AR-test. The reasons are that we may not know the degree of endogeneity, but the distribution of AR-statistic is independent of endogenity so that the only concern is the strength of IV. However, since we have many observations comparing to the number of IV, the power of AR-test is comparable to the textbook power function.

6 Conclusion

This project first confirms that TSLS estimator under weak is non-normal and biased. Moreover, given small sample size, t-statistic has incorrect size of test under one or many weak IV, while AR-test can improve the size but sacrifices the power. Furthermore, increasing sample size can only improve the size of t-test under one weak IV but hardly improves that under many weak IV. On the other hand, AR-test has correct size for all specifications but may have worse power in small sample. However, the power of AR-test is growing in sample size. Then, I recommend AR-statistic under large observations if the degree of endogeneity and strength of IV are unknown.

References

Anderson, T. & Rubin, H. (1949), 'Estimators of the parameters of a single equation in a complete set of stochastic equations', *The Annals of Mathematical Statistics* **21**.

Angrist, J. D. & Krueger, A. B. (1991), 'Does compulsory school attendance affect schooling and earnings?', *The Quarterly Journal of Economics* **106**(4), 979–1014.
URL: *http://www.jstor.org/stable/2937954*

Kleibergen, F. (2002), 'Pivotal statistics for testing structural parameters in instrumental variables regression', *Econometrica* **70**(5), 1781–1803.

Marmer, V. (2017), 'Econometrics with Weak Instruments: Consequences, Detection, and Solutions'. https://faculty.arts.ubc.ca/vmarmer/. Last accessed on 22.09.2018.

Moreira, M. J. (2003), 'A conditional likelihood ratio test for structural models', *Econometrica* **71**(4), 1027–1048.

Staiger, D. & Stock, J. H. (1997), 'Instrumental variables regression with weak instruments', *Econometrica* **65**(3), 557–586.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.