THE UNIVERSITY OF CHICAGO

A Survey of the Weak Instruments Problem in Just-Identified Linear IV Regression

By

Boyang Zhang

July 2024

A paper submitted in partial fulfillment of the requirements for
the Master of Arts Program in
Social Science

Faculty Advisor: Joseph Hardwick
Preceptor: Joseph Hardwick

Abstract

The Weak Instruments problem, which characterizes the situation where the endogenous regressor is only weakly correlated with the excluded instrument, is a problem frequently encountered in empirical practice. In this paper, we focus on the most popular setting where there is a single endogenous regressor and single instrument, and try to provide empirical researchers with an accessible guide to the weak instruments literature. The focus is on the nature of weak instruments problem, methods to detect weak instrument, and methods to deal with weak instrument. A distinct feature of this paper is that it also explores the situation where the true first stage regression is nonlinear. This paper also surveys papers published on *American Economic Review* between 2020 and 2024. The result shows that empirical researchers still have relatively insufficient understanding of the weak instruments problem and the methods available to deal with it.

Keywords: Weak Instruments, Nonlinear First Stage, Anderson-Rubin

1 Introduction

In linear Instrumental Variable (IV) regression, the Weak Instruments problem characterizes the situation where the correlation between the endogenous regressor and the instrumental variable(s) is so weak that the limiting distribution of the conventional "Two Stage Least Squared" (TSLS) estimator and the corresponding t statistic are highly non-normal. Thus, the TSLS estimator will have large finite sample bias and the traditional t-ratio-based inference will have severe size distortion. I. Andrews et al. (2019) surveys 17 papers published on *American Economic Review* (*AER*) between 2014 and 2018 and finds that the weak instruments problem is prevalent in the empirical literature, so it has practical value in providing a survey of the literature and helping empirical researchers do correct estimation and inference in practice. This paper will exclusively focus on specifications with a single endogenous regressor and a single instrument (just-identification), as it is the most popular case in practice (See I. Andrews et al., 2019) and the case that we know most about.

The papers survey by I. Andrews et al. (2019) are published between 2014 and 2018 and the papers survey by D. S. Lee et al. (2022) are published between 2013 and 2019. Thus, to see if there are significant changes of empirical practice, we do a brief survey of papers published on *American Economic Review* that use IV model between 2020 and 2024[1] and

[1] Only papers that have IV regression results in the main text are included, so papers that use IV regression in the appendix are excluded. We also exclude papers that have multiple endogenous regressors or use other estimation techniques, such as non-linear GMM.

focus on the first stage F statistic they report[2].

We have a total of 78 papers that have a single endogenous regressor and use 2SLS to estimate the result. In this sample, 60 papers exclusively use just-identified model, 15 papers exclusively use over-identified model, and 3 papers use both just-identified and over-identified models. Therefore, just identified-model constitutes the vast majority of IV application we encounter in practice, which is in contrast to the survey by I. Andrews et al. (2019) for *AER* between 2014 and 2018 and D. S. Lee et al. (2022) for *AER* between 2013 and 2019, where only about half of the papers use just-identified models.

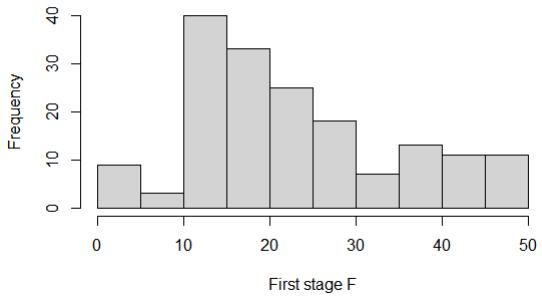

Figure 1: Distribution of First Stage F statistic under 50

Among just-identified models, Figure 1 shows the histogram of reported first stage F statistic under 50. Similar to the finding in I. Andrews et al. (2019) and D. S. Lee et al. (2022), there is a spike at 10, which means that researchers may have screened the First stage F statistic to make it pass the "rule of thumb".

It is also worth noting that there is a non-negligible number of specifications with first stage F statistic below 10, suggesting possible bias of TSLS and size distortion of corresponding t test. However, only 2 papers report Anderson-Rubin confidence interval, which we will see is robust to weak instruments problem and is the recommended procedure.

There are many first stage F statistics, including heteroskedasticity robust and non-robust first stage F statistic in the just-identified case, and effective first stage F statistic due to Olea and Pflueger (2013) in the over-identified case. We find that only 13 out of 63 papers that use just-identified models explicitly state the type of first stage F statistic that they use. Thus, providing a review of methods available in the weak instruments literature may be helpful.

[2]We only include first stage F statistic that is directly available in the table or can be directly computed based on the information in the table. For example, some papers with just-identified models only report the first stage coefficient and corresponding standard error rather than first stage F statistic explicitly, but obviously we can compute first stage F statistic based on these information.

The remaining parts of the paper are structured as follows: Section 2 sets up the Instrumental Variable context and relevant assumptions. Section 3 discusses the problems caused by having a weak instrument. Section 4 shows the bootstrap method does not work under weak instrument. Section 5 presents methods to detect the presence of weak instrument. Section 6 presents methods to deal with instruments, including alternative estimators and alternative inference methods. Section 7 allows nonlinear first stage and examines its consequences on estimation and inference. Section 8 extends the model to allow error term to be heteroskedastic. Section 9 applies methods discussed in Section 5 and 6 to the famous Acemoglu et al. (2001) paper. Section 10 discusses possible directions for future research.

2 IV set up

Consider the following just-identified linear IV regression model with a sample of independent and identically distributed (iid) data $\{Y_i, X_i, Z_i\}_{i=1}^n$, where Y_i is the outcome variable, X_i is the endogenous regressor, and Z_i is the exogenous instrumental variable. We assume no control variables (included instruments) for simplicity. $Y = (Y_1, Y_2, ..., Y_n)' \in R^n, X = (X_1, X_2, ..., X_n)' \in R^n$, and $Z = (Z_1, Z_2, ..., Z_n)' \in R^n$, we will have:

$$Y = X\beta + u \tag{1}$$

$$X = Z\pi + v \tag{2}$$

Before proceeding to the estimators and their limiting distributions, we make the following assumptions:

1. **Instrument Relevance and Exogeneity**

$$\pi \neq 0, E(u_i|Z_i) = 0, E(Z_i v_i) = 0$$

2. **Conditional Homoskedasticity and Normalization**

$$\frac{1}{n}Z'Z \xrightarrow{p} E[z_i^2] = 1$$

$$(\frac{1}{\sqrt{n}}Z'u, \frac{1}{\sqrt{n}}Z'v)' \xrightarrow{d} N(\begin{pmatrix}0\\0\end{pmatrix}, E[z_i^2]\begin{pmatrix}\sigma_u^2 & \sigma_{uv}\\ \sigma_{uv} & \sigma_v^2\end{pmatrix})$$

3. **Endogeneity**

$$\sigma_{uv} \neq 0$$

Several comments of Assumption **1** are in order. u is the structural error which has specific economic interpretation, for example the "ability" in the context of return to education. As Z is exogenous, it makes sense to assume $E[u_i|Z_i] = 0$. In contrast, v is a pure mathematical

object, namely the remainder after projecting X to Z, so it has no economic interpretation. It is simple to show that $E[Z_i v_i] = 0$ is true by construction, whereas $E[v_i|Z_i] = 0$ is not true in general unless $E[X_i|Z_i]$ is linear in Z.

3 The Weak Instruments Problem

The weak instruments problem happens when the correlation between the endogenous regressor and the instrument is so low that the bias of IV estimator is large and t-ratio-based inference procedure is potentially very misleading. To more clearly see this problem, assume the extreme case where $\pi = 0$, which means the instrument is completely irrelevant and β is unidentified. The IV estimator has limiting distribution:

$$\hat{\beta}_{IV} - \beta = \frac{Z'u}{Z'X} \xrightarrow{d} \frac{\sigma_{uv}}{\sigma_v^2} + \frac{\eta}{\psi_{zv}} \tag{3}$$

$$\begin{pmatrix} \eta \\ \psi_{zv} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2(1-\rho^2) & 0 \\ 0 & \sigma_v^2 \end{pmatrix}), \quad \rho = \frac{\sigma_{uv}}{\sigma_v \sigma_u} \tag{4}$$

The proof will be given in the Appendix.

Similarly the OLS estimator is

$$\hat{\beta}_{OLS} - \beta = \frac{X'u}{X'X} \xrightarrow{p} \frac{\sigma_{uv}}{\sigma_v^2}$$

We can see from Equation 4 that $\hat{\beta}_{IV}$ is not even consistent. More interestingly, the IV estimator is converging to the probability limit $\beta + \frac{\sigma_{uv}}{\sigma_v^2}$ of $\hat{\beta}_{OLS}$ plus a scaled Cauchy distribution. Thus, in the extreme case where the instrument is completely irrelevant, $\hat{\beta}_{IV}$ is neither consistent nor asymptotically normal, so the traditional t-ratio-based inference will generally provide misleading result. In fact, $\hat{\beta}_{IV}$ performs even worse than $\hat{\beta}_{OLS}$, as Cauchy distribution has heavy tails.

Now let us consider the more realistic case where the first stage coefficient is close to 0, but not exactly 0. Before diving into the math, we first show some simple simulations to more intuitively understand the problem. The sample size $N = 1000$, iterations $S = 5000$, degree of endogeneity $\rho = 0.99$, first stage strength $\pi \in \{0, 0.05, 0.10, 0.15, 0.20, 0.35\}$, $Z_i = 1$, $\sigma_u = \sigma_v = 1$

Figure 2 shows the distribution of the IV estimator under different identification strength. We can see that when the instrument is weak, the median of IV estimator is far from 0, showing substantial median bias. When the instrument gets stronger, the median gradually converges to the true value 0.

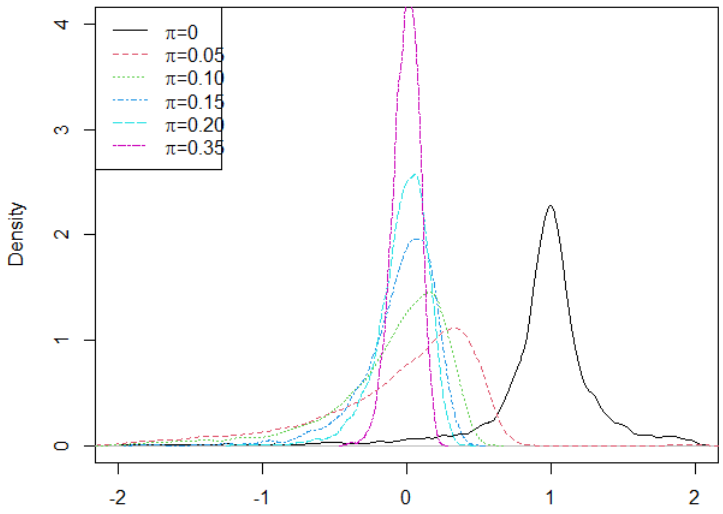

Figure 2: Distribution for IV estimator with $\rho = 0.99$

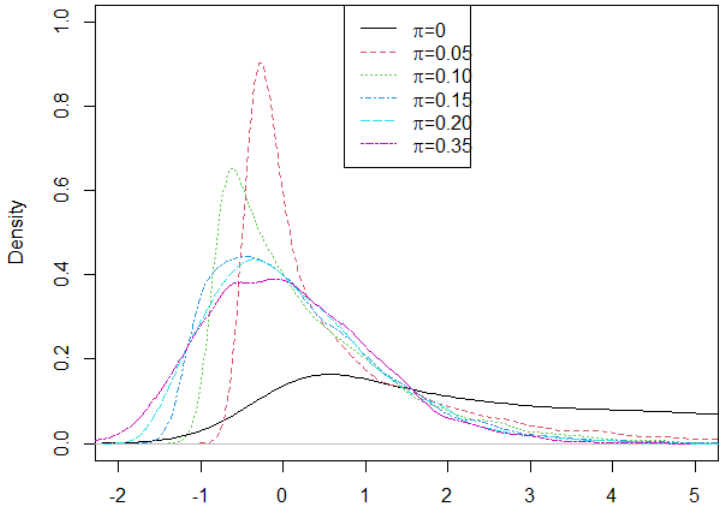

Figure 3: Distribution of t statistic with $\rho = 0.99$

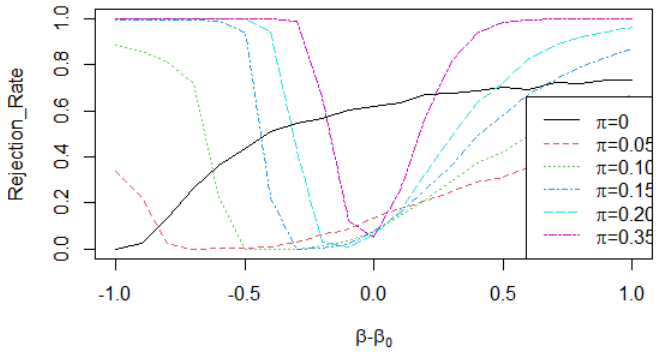

Figure 4: Power of t-test with $\rho = 0.99$

Figure 3 shows that the distribution of t-statistic under different instrument strength, holding the degree of endogeneity (correlation between first stage and second stage error term) fixed at 0.99. The t statistic under $\beta = \beta_0$ is:

$$t = \frac{\sqrt{n}(\hat{\beta}_{IV} - \beta_0)}{\sqrt{\widehat{\text{AsyVar}}(\hat{\beta}_{IV})}}$$

$$\widehat{\text{AsyVar}}(\hat{\beta}_{IV}) = \hat{\sigma}_u^2(\frac{1}{n}Z'X)^{-2}\frac{1}{n}Z'Z, \quad \hat{\sigma}_u^2 = \frac{1}{n}(Y - X\hat{\beta}_{IV})'(Y - X\hat{\beta}_{IV})$$

We can see that when the instrument is strong, the distribution resembles a normal distribution, but as the instrument gets weaker, the t-statistic performs less and less similar to a normal distribution, suggesting t-ratio-based inference will produce misleading result.

Figure 4 shows the power of the t-test with different degree of instrument strength. We use a grid of parameter values, in this case from -1 to 1, with interval 0.1, and collect cases where the $|t| > 1.96$. By doing so we can get the power of t test for each parameter value. When $\beta - \beta_0 = 0$, namely the null hypothesis is true, the rejection probability is exactly the size of the test. It is obvious that when $\pi = 0$ or $\pi = 0.05$, the t test has virtually no ability to detect a large true negative effect, but has some ability to detect large true positive effect. As the instrument becomes stronger, the power curve becomes more symmetric and has roughly correct size and power.

Before preceding, it is interesting to see from Figure 3 the distribution of t statistic is highly asymmetric, namely it has a much fatter tail on the right, which is even true when $\pi = 0.35$, which corresponds to a first stage F statistic (See more in Section 5) larger than 100. Another way of interpreting this asymmetry is that positive parameter estimate corresponds to smaller standard error while negative parameter estimate corresponds to larger standard

error. The reason for this pattern is simple. As pointed out by Keane and Neal (2023), when the bias of OLS estimate is positive (or the correlation between the endogenous regressor and the structural error is positive), a positive sample correlation between the instrument and structural error (the population correlation is 0 by the instrument validity assumption) will inflate both the TSLS estimate and the correlation between instrument and endogenous regressor, namely dragging down the standard error. Thus, there is a negative correlation between TSLS estimate and standard error. Keane and Neal (2023) shows that this phenomenon may cause the traditional t test to have low power in detecting true negative effect while high power to find false positive effect, a phenomenon which they argue to be largely neglected in the weak instruments literature.

The simulation results in Figure 2, 3, and 4 illustrate the poor performance of $\hat{\beta}_{IV}$ and t statistic even in large samples under weak instruments. Thus, the sample size does not affect whether instruments is weak. The literature suggests that the "Concentration Parameter", which I will define below, is the parameter that truly dictates the performance of IV estimator and t statistic.

Further assume $v_i \sim N(0, \sigma_v^2), u_i \sim N(0, \sigma_u^2), \frac{1}{n} Z'Z = 1$. The concentration parameter $\mu^2 = \frac{\pi' Z' Z \pi}{\sigma_v^2}$, we have

$$\begin{aligned} \hat{\beta}_{IV} - \beta &= \frac{Z'u}{Z'X} \\ &\sim \frac{\sigma_u}{\sigma_v} \frac{z_u}{\mu + z_v} \\ \begin{pmatrix} z_u \\ z_v \end{pmatrix} &\sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}), \frac{\rho}{\sigma_{uv}} \sigma_u \sigma_v \end{aligned}$$

It is clear from the expression that when μ^2 is large, $\hat{\beta}_{IV}$ well approximates β and the conventional estimation and inference methods work. If μ^2 is small, $\hat{\beta}_{IV}$ will be far from β in finite samples, so the conventional t-ratio-based inference methods do not work.

As $\mu^2 = \frac{\pi' Z' Z \pi}{\sigma_v^2}$, when the sample size is large μ will be large at well, $\hat{\beta}_{IV}$ will be approximately normal and weak instruments is no longer a problem. However, as shown by Bound et al. (1995), the weak instruments problem can happen even when the sample size is extremely large. Therefore, Staiger and Stock (1997) tries to model the weak instruments problem by setting $\pi = \frac{C}{\sqrt{n}}$, where C is some constant. It is also known as the "weak instruments asymptotics".

The intuition for "weak instruments asymptotics" is straightforward: After setting $\pi = \frac{C}{\sqrt{n}}$, we have $\mu^2 = \frac{C' \frac{1}{n} Z' Z C}{\sigma_v^2} \xrightarrow{p} \frac{C' Q_z C}{\sigma_v^2}$, a constant. In other words, we do not allow the concentration parameter to depend on the sample size, as it should only depend on the strength of instruments. By doing so, the concentration parameter can indicate the

instrument strength. Without assuming the finite sample distribution of v_i and u_i, we have $\hat{\beta}_{IV} - \beta \xrightarrow{d} \frac{\sigma_u}{\sigma_v} \frac{z_u}{C/\sigma_v + z_v}$

Even though the intuition is straightforward, the approximation can still be bad in practice. Figure 5 is a simulation that examines the quality of modelling $\pi = \frac{C}{\sqrt{n}}$. The idea is to compare the distribution of $\hat{\beta}_{IV} - \beta$ and distribution of $\frac{z_u}{C+z_v}$, the limiting distribution of $\hat{\beta}_{IV} - \beta$ under weak instrument asymptotics. We can see from Figure 5 that the actual distribution and the limiting distribution is virtually indistinguishable, confirming that the "weak instrument asymptotics" is a good model for weak instruments.

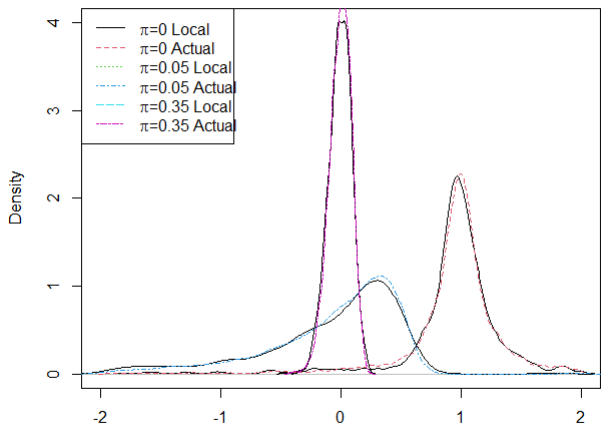

Figure 5: Distribution of original IV estimate and sample from $\frac{z_u}{C+z_v}$

After modelling the weak instruments problem, let us go back to the problem shown in Figure 2, 3, and 4, namely why IV estimator and t statistic perform so poorly under weak instruments. We first consider the size (maximum rejection probability of t test under the null and across nuisance parameters) of the t test.

Suppose we are testing

$$H_0 : \beta = \beta_0 \quad vs \quad H_1 : \beta \neq \beta_0$$

A note of caution is that under weak instrument asymptotics, the t statistic will no longer have a limiting distribution that is symmetric around 0, so the sign of certain quantities (as we will see in the Appendix, $Z'X$) will affect the limiting distribution of t statistic. Thus, For Equation 6, 26, 27, and 37, we will present the limiting distribution of t^2 (or Wald)

statistic. The asymptotic distribution of Wald statistic as:

$$t^2 = \left(\frac{\sqrt{n}(\hat{\beta}_{IV} - \beta_0)}{\sqrt{\widehat{\text{AsyVar}}(\hat{\beta}_{IV})}}\right)^2 \tag{5}$$

$$\xrightarrow{d} \frac{z_u^2}{1 - 2\rho\frac{z_u}{C/\sigma_v + z_v} + \left(\frac{z_u}{C/\sigma_v + z_v}\right)^2} \tag{6}$$

The asymptotic distribution of t statistic is highly non-normal, causing the traditional confidence interval $[\hat{\beta}_{IV} - 1.96\sqrt{\frac{\widehat{\text{AsyVar}}(\hat{\beta}_{\text{IV}})}{n}}, \hat{\beta}_{IV} + 1.96\sqrt{\frac{\widehat{\text{AsyVar}}(\beta)_{\text{IV}}}{n}}]$ to be unreliable. Figure 6 shows the effect of weak instruments on the size of t-test. It is shown that when C is small,

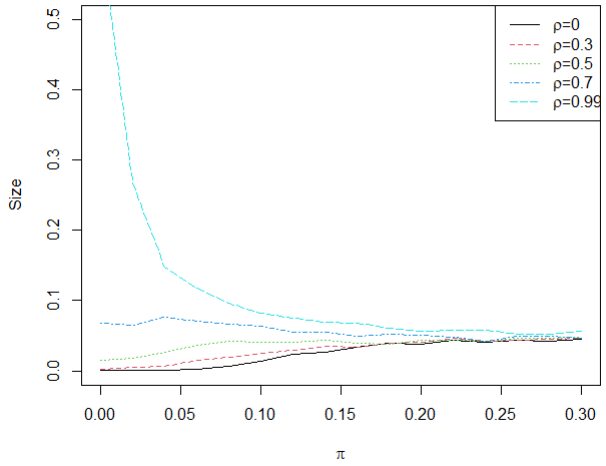

Figure 6: size of t-test with different instrument strength and degree of endogeneity

the size of the t-test can be much larger than 5%, in other words, the confidence interval significantly undercovers the true β. While the size becomes much closer to 5% when the instrument becomes stronger. It is worth noting that when the degree of endogeneity is small, the size of t test is less than 5% however weak the first stage is. In fact, J. Angrist and Kolesár (2023) argues that it is unlikely to have degree of endogeneity larger than 0.5 in practice, so t test for just-identified IV regression is reliable irrespective of instrument strength.

Another problem caused by weak instruments is that IV estimator is "biased", or not centered at the true value. Under weak instrument asymptotics,

$$\hat{\beta}_{IV} - \beta = \frac{Z'u}{Z'X} \xrightarrow{d} \frac{\psi_{zu}}{C + \psi_{zv}} \tag{7}$$

$$\begin{pmatrix}\psi_{zu}\\ \psi_{zv}\end{pmatrix} \sim N(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_u^2 & \sigma_{uv}\\ \sigma_{uv} & \sigma_v^2\end{pmatrix}) \tag{8}$$

I use quotation marks because IV estimator does not have moments when the model is just-identified, so the bias in the usual sense does not exist. Stock and Yogo (2005) considers the Nagar bias to approximate the deviation from the true value.

Under the weak instrument asymptotics[3],

$$\hat{\beta}_{IV} - \beta \xrightarrow{d} \frac{z_v}{\mu_\infty + z_v}\frac{\sigma_{uv}}{\sigma_v^2} + \frac{\eta}{C + \psi_{zv}}, \mu_\infty^2 = \frac{C^2}{\sigma_v^2} \tag{9}$$

The second term is a scaled Cauchy distribution, so we only focus on the first term in terms of the location of the estimator. Using Taylor expansion, we have

$$E[\hat{\beta}_{IV} - \beta] \approx E[\frac{z_u}{\mu_\infty + z_v}\frac{\sigma_{uv}}{\sigma_v^2}] \approx -\frac{\sigma_{uv}}{\sigma_v^2}\frac{1}{\mu_\infty^2} \tag{10}$$

Thus, the "bias" of the IV estimator increases as the concentration parameter becomes smaller and the degree of endogeneity becomes higher. Note $\frac{\sigma_{uv}}{\sigma_v^2}$ is exactly the bias of OLS estimator, so Stock and Yogo (2005) considers the relative "bias" of IV with respect to OLS,

$$\frac{E[\hat{\beta}_{IV} - \beta]}{E[\hat{\beta}_{OLS} - \beta]} \approx -\frac{1}{\mu_\infty^2}$$

which only depends on the concentration parameter. In fact, as we will see in Section 5, the first stage F statistic has mean $\mu_\infty^2 + 1$, if the researcher believes 10% relative bias is a reasonable threshold for weak instruments, he will find first stage F statistic greater than 10 an indication of strong instruments, and this is where the "rule of thumb", 10, for the first stage F statistic comes from.

4 Bootstrap Failure Under Weak Instruments

Bootstrap is a popular alternative when the conventional normal approximation fails. However, Bootstrap does not work for the weak instruments problem because the distribution of t statistic is not continuous with respect to the data generating process.

We are interested in forming confidence interval for θ that has coverage at least $1 - \alpha$. Let $R_n = \sqrt{n}(\hat{\theta} - \theta)$, traditionally we use normal distribution to approximate R_n and form confidence interval of θ based on it. However, as shown in Section 3, this normal approximation can be bad. The idea of Bootstrap is to take the sample $\{W_i\}_{i=1}^n = \{X_i, Y_i, Z_i\}_{i=1}^n \sim P$ as the population, get new sample by sampling with replacement from the new population, and take advantage of the fact that the relationship between the new sample and the new

[3] I use μ_∞ here to separate it from the finite sample μ

population is similar to that of original sample and original population. As we can approximate the distribution of the new population and the new sample using their empirical counterpart, we can form confidence interval for θ. In fact, the reason for Bootstrap failure under weak instruments is that the relationship between new sample and new population is not close enough to that of original sample and population.

More formally[4], let $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$, $J_n(t, P) = P(R_n < t)$. Sample B times with replacement from the original sample $\{W_i\}_{i=1}^n$ to get $W^1, ..., W^B$, compute $R_n^b = \sqrt{n}(\hat{\theta}_n^b - \hat{\theta}_n)$ for each W^b. Let $J_n(t, \hat{P}_n) = P(R_n^b \leq t)$. $\hat{P}_n$ denotes the empirical distribution of the original samples. As $J_n(t, \hat{P}_n)$ is still observed, we use its sample analogue $\hat{J}_n(t, \hat{P}_n) = \frac{1}{B}\sum_{b=1}^B \mathbf{I}_{\{R_n^b \leq t\}}$. We hope $\hat{J}_n(t, \hat{P}_n)$ and $J_n(t, P)$ are closed as $\hat{P}_n$ gets close to P, but it is not true here because $\sqrt{n}(\hat{\beta}_{IV} - \beta)$ diverges when $\pi = 0$ while $\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N(0, \frac{\sigma_\varepsilon^2}{\pi^2})$ when $\pi \neq 0$, namely the distribution of J is not continuous.

A simple simulation demonstrates the point. Let sample size $n = 500$, Bootstrap iteration $B = 1000$, $\rho = 0.99$, $\pi \in \{0, 0.05, 0.1, 0.15, 0.2, 0.35\}$, $Z_i = 1$. Repeat the above procedure 1000 times.

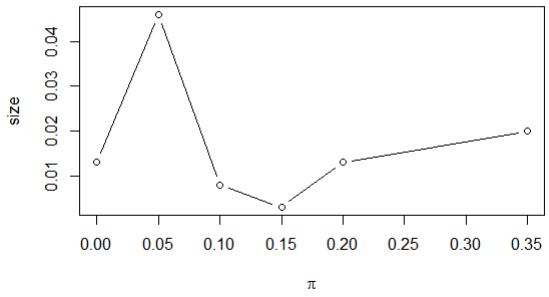

Figure 7: Size of Bootstrap t test

Figure 7 shows the size of the Bootstrap t test. We can see that the size is extremely close to 0 in most of the cases, suggesting that the confidence interval may too wide. Interestingly in this specific simulation design the bootstrap test does not over-reject, even when instruments is extremely weak. Probably in some other cases it will over-reject. Modifications of Bootstrap, including m out of n bootstrap and subsampling, also do not work when instrument is weak, see D. W. Andrews and Guggenberger (2009) for more discussion.

[4]This part borrows heavily from Dr Ponomarev's machine learning class note and Professor Torgovitsky's Applied Microeconometrics class note

5 Detection of Weak Instruments

We have shown that weak instruments can have serious consequences, both in terms of bias of IV estimator and size distortion of the corresponding t statistic. so it is important to detect it. It is shown that the instrument strength depends exclusively on the first stage coefficient π (or C). However, C is neither known nor consistently estimable. Thus, the idea is to find a statistic whose distribution depends on C, so that we can form hypothesis based on this statistic and test the strength of the instrument. The most popular way in practice is to compute the first stage F statistic testing the hypothesis that first stage coefficient $\pi = 0$, suggested by Staiger and Stock (1997) and formalized by Stock and Yogo (2005).

$$\hat{F} = \frac{(\sqrt{n}\hat{\pi})^2}{\widehat{\text{Asy}}\text{Var}(\hat{\pi})} \xrightarrow{d} (z_v + \frac{C}{\sigma_v})^2 \overset{def}{=} F \tag{11}$$

$$\widehat{\text{Asy}}\text{Var}(\hat{\pi}) = \hat{\sigma}_v^2 (Z'Z)^{-1}, \quad \hat{\sigma}_v^2 = \frac{1}{n}(X - Z\hat{\pi})'(X - Z\hat{\pi}) \tag{12}$$

$$\text{Thus} \quad F \sim \chi_1^2((\frac{C}{\sigma_v})^2), \quad E(F) = (\frac{C}{\sigma_v})^2 + 1 \tag{13}$$

Obviously the limiting distribution of first stage F statistic is a non-central chi square distribution, with non-centrality parameter $(\frac{C}{\sigma_v})^2$. If we set $\sigma_v = 1$ and test the hypothesis $H_0 : C = \sqrt{10}$ vs $H_1 : C > \sqrt{10}$. Under the null hypothesis the distribution of F is known, so we can compute the sample first stage F statistic and compare it with the critical value of this known non-central chi square distribution. If the hypothesis is rejected, or the first stage F statistic is large enough, we do not need to worry about the weak instruments problem, otherwise we may worry about the consequences of weak instruments.

Hahn and Hausman (2002) proposes another method to detect weak instruments. The idea is to compare the conventional 2SLS estimator with the "backward" 2SLS estimator, namely estimating the same parameter in two different ways. In the conventional case where the first order asymptotics provides a good approximation to the distribution, the difference between these two estimators, after proper rescaling, should converge in probability to 0. If it does not, we may worry that the first order asymptotics does a bad job and the weak instruments problem may be a concern.

More formally, the traditional TSLS estimator is

$$\hat{\beta}_{TSLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'P_zX)^{-1}X'P_zY$$

We now consider the backward TSLS estimator $\frac{1}{\hat{c}}$ derived from the "backward" TSLS regression.

$$X = Yc + u$$

$$Y = Z\pi + v$$

$$\hat{c} = (\hat{Y}'\hat{Y})^{-1}\hat{Y}'X = (Y'P_zY)^{-1}Y'P_zX$$

We can show that

$$\sqrt{n}(\hat{\beta} - \frac{1}{\hat{c}}) \xrightarrow{p} 0 \tag{14}$$

However, as Hahn and Hausman (2002) points out, in practice $\hat{\beta}$ and $\frac{1}{\hat{c}}$ can differ by a large amount.

Hahn and Hausman (2002) uses the second order asymptotics to derive the asymptotic distribution of $\hat{\beta} - \frac{1}{\hat{c}}$, under the assumption that $\frac{K}{n} \to \alpha$, as in the many instruments literature initiated by Bekker (1994). It is shown in Hahn and Hausman (2002) (omitted in this paper) that

$$\hat{\beta} - \frac{1}{\hat{c}} \xrightarrow{p} B$$

$$\sqrt{n}(\hat{\beta} - \frac{1}{\hat{c}}) \xrightarrow{d} N(0, V)$$

where both B and V are proportional to α. Then we can test the hypothesis

$$H_0 : \text{plim}\sqrt{n}(\hat{\beta} - \frac{1}{\hat{c}} - \hat{B}) = 0$$

where $\hat{B}$ is the estimator for B. This idea is exactly the opposite as the test using first stage F statistic, as it tests the null hypothesis that the instrument is strong, while the test based on first stage F statistic tests the null that instrument is weak. Unfortunately, this test is shown to have poor power by Hausman et al. (2005), namely it may not do a good job detecting the presence of weak instruments.

In the just-identified setting that we focus on, this test does not work because $\hat{\beta}$ and $\frac{1}{\hat{c}}$ are exactly the same, namely

$$\hat{\beta} = \frac{1}{\hat{c}} = (Z'X)^{-1}Z'Y$$

6 Dealing with Weak Instruments

6.1 Alternative Estimators

After detecting the presence of weak instruments, another important question is how to deal with it. A natural suggestion is to use another estimator so that bias can be eliminated. However, Hirano and Porter (2014) shows that no unbiased or asymptotically unbiased estimator[5] is available if instrument can be arbitrarily weak ("singular" in their terminology)

[5] "Asymptotically unbiased" means that the estimator converges in distribution (After proper centering and rescaling) to a random variable, which has expected value equal to the true parameter value.

and no additional assumption is imposed. Thus, researchers aim for alternative estimators that can reduce bias. Limited Information Maximum Likelihood estimator (LIML) is a popular alternative among econometricians, as Staiger and Stock (1997) and Stock and Yogo (2005) show in simulations that it is less biased than TSLS and can have better size control. However, none of the papers with over-identified model in my *AER* sample reports LIML estimate. J. D. Angrist et al. (2023) is a nice paper that reports LIML estimate.

LIML is a maximum likelihood estimator, which can be shown to be:

$$\hat{\beta}_{LIML} = (X'(I - \kappa M_z)X)^{-1}X'(I - \kappa M_z)Y \tag{15}$$

where $M_z = I - Z(Z'Z)^{-1}Z'$, and κ is the smallest root of equation

$$\det(\bar{Y}'\bar{Y} - \kappa\bar{Y}'M_z\bar{Y}) = 0$$

where $\bar{Y} = (Y, X)$. It can be shown that $\kappa = 1$ in the just-identified model, so $\hat{\beta}_{LIML} = \hat{\beta}_{2SLS}$ under just identification.

However, I. Andrews and Armstrong (2017) shows that if the sign of one or more first stage coefficient is known, unbiased estimator (and asymptotically unbiased, if the error term is unknown and distribution is non-normal) is available.

More formally, Consider the just identified reduced form regression with fixed instruments.

$$y = Z\pi\beta + v_1 \tag{16}$$

$$X = Z\pi + v_2 \tag{17}$$

and known variance covariance matrix

$$\begin{pmatrix} v_{1i} \\ v_{2i} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix})$$

Then

$$\begin{pmatrix} \widehat{\pi\beta} \\ \hat{\pi} \end{pmatrix} \sim N(\begin{pmatrix} \pi\beta \\ \pi \end{pmatrix}, (Z'Z)^{-1}\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}) \tag{18}$$

$\hat{\beta}_{IV} = \frac{\widehat{\pi\beta}}{\hat{\pi}}$, but the expectation of ratio is not the ratio of expectations, so $\hat{\beta}_{IV}$ is not unbiased. I. Andrews and Armstrong (2017), however, considers $\hat{\delta} = \widehat{\pi\beta} - \frac{\sigma_{12}}{\sigma_2^2}\hat{\pi}$, and $\hat{\tau}$, an unbiased estimator for $\frac{1}{\pi}$. Also $\hat{\tau}$ is only a function of $\hat{\pi}$.

As $\widehat{\pi\beta}$ and $\hat{\pi}$ are jointly normal, and

$$Cov(\hat{\delta}, \hat{\pi}) = (Z'Z)^{-1}\sigma_{12} - (Z'Z)^{-1}\sigma_{12} = 0$$

$\hat{\delta}$ is independent of $\hat{\pi}$, thus independent of $\hat{\tau}$. Therefore,

$$E[\hat{\delta}\hat{\tau}] = E[\hat{\delta}]E[\hat{\tau}] = \beta - \frac{\sigma_{12}}{\sigma_2^2}$$

and

$$\hat{\beta} = \hat{\delta}\hat{\tau} + \frac{\sigma_{12}}{\sigma_2^2}$$

is an unbiased estimator for β

The problem left is finding $\hat{\tau}$. I. Andrews and Armstrong (2017) shows that,

$$\hat{\tau} = \frac{1}{\sigma_2}\frac{1 - \Phi(\hat{\pi}_2/\sigma_2)}{\phi(\hat{\pi}_2/\sigma_2)}$$

$E[\hat{\tau}] = \frac{1}{\pi}$ if $\pi > 0$[6]

Thus,

$$\hat{\beta} = (\widehat{\pi\beta} - \frac{\sigma_{12}}{\sigma_2^2}\hat{\pi})(\frac{1}{\sigma_2}\frac{1 - \Phi(\hat{\pi}_2/\sigma_2)}{\phi(\hat{\pi}_2/\sigma_2)}) + \frac{\sigma_{12}}{\sigma_2^2}$$

is an unbiased estimator of β. I. Andrews and Armstrong (2017) also shows that, if the distribution of error term is unknown, under strong instrument, $\hat{\beta}$ has the same asymptotic behaviour as $\hat{\beta}_{TSLS}$, namely asymptotically normal, but is asymptotically unbiased if is under the weak instrument asymptotics.

6.2 Valid Inference

No unbiased or consistent estimator exists does not mean that no valid statistical inference is available. By "valid statistical inference" I mean using a statistical test that has the correct size (normally 5%). By test inversion a valid 95% confidence interval can be constructed once a test with 5% size is available.

Currently there are mainly two approaches to conduct valid inference when facing weak instruments. The first method is to completely discard estimators but focus on using pivotal test statistic, which has a limiting distribution that does not depend on the strength of instruments or other nuisance parameters. Anderson Rubin statistic (AR) and Lagrange Multiplier statistic (LM) (or Kleibergen's K statistic) are two main examples in this class of tests.

The second method uses non-pivotal statistic, for example t statistic, but adjusts the critical value to ensure correct size of the test. Sometimes the critical value is adjusted by combining the original test with pivotal test statistic mentioned above. Researchers first conduct a pre-test based on first stage F statistic. If F is small, for example less than 10,

[6]Though the theorem only applies when $\pi > 0$, we can run the first stage $X = -Z\pi + v$ if $\pi < 0$, so that the theorem still applies.

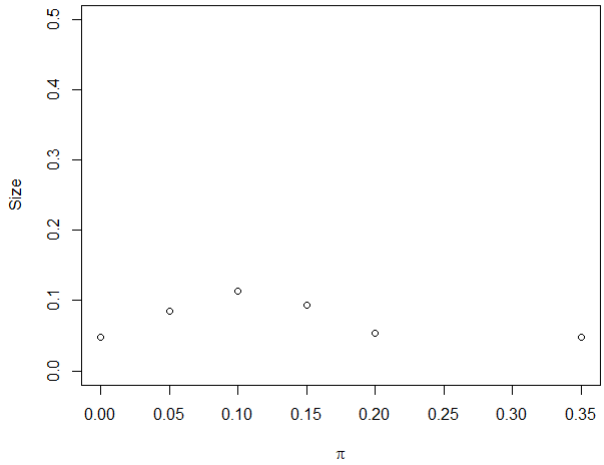

Figure 8: Size of the two step method with $\rho = 0.99$

researchers use pivotal test statistic. If F is large, researchers use the conventional point estimate and standard error. As this is a two-step method, the overall size of the procedure is not correct if the traditional 1.96 critical value is used. Thus, if the researcher wants to have a 5% size for the overall procedure, he needs to change the threshold for first stage F statistic or the critical value of t test, or both. Figure 8 simulates the size of the two step method, which uses a AR test if the first stage F statistic is less than 10, and a t test with 1.96 as critical value when the first stage F statistic is larger than 10. Obviously relying on the "rule of thumb" 10 and critical value 1.96 will not give correct size for some instrument strength, though it has much better size control compared to using the t test alone.

There are other methods to adjust critical value. The most famous one is the "Conditional test" approach popularized by Moreira (2003) and Moreira (2009). The idea is straightforward: Although the original test statistic is not pivotal, if there is a statistic that is sufficient for the nuisance parameter, then conditioning on this sufficient statistic will give a pivotal statistic for every value of this sufficient statistic, thus valid inference can be conducted. Conditional Likelihood Ratio statistic (CLR) and Conditional Wald statistic (CW) are two main examples in this class of tests.

Some very recent papers (D. S. Lee et al., 2022 and D. Lee et al., 2023), however, use some complicated methods to adjust critical value, which turns out to work well. I will introduce these valid inference methods one-by-one.

I will start by introducing the AR test proposed by Anderson and Rubin (1949). The idea is to make use of the exclusion restriction and deal with the null hypothesis directly, without using information about the first stage, thus is robust to weak instruments.

More formally, suppose researchers want to test the hypothesis

$$H_0 : \beta = \beta_0 \quad vs \quad H_1 : \beta \neq \beta_0$$

If the exclusion restriction of instruments holds under the null hypothesis, namely Z does not enter the second stage regression, researchers can test whether Γ in the regression $y - X\beta_0 = Z\Gamma + u$ is 0. The only reason for rejecting the hypothesis $\Gamma = 0$ is that the null hypothesis H_0 is not true, because the exclusion restriction is satisfied. Otherwise we do not find enough evidence to reject $\beta = \beta_0$. Therefore, the Anderson-Rubin statistic is the F statistic testing $\Gamma = 0$, with k being the number of instruments

$$AR(\beta_0) = \frac{(y - X\beta_0)' P_z (y - X\beta_0)}{\frac{1}{n-k}(y - X\beta_0)' M_z (y - X\beta_0)}$$
$$\text{where} \quad P_z = Z(Z'Z)^{-1}Z', M_z = I - P_z$$

It is obvious that AR does not depend on first stage information, hence robust to weak instruments. To verify the performance, I conduct a similar simulation exercise to compute the rejection probability of AR statistic under different degree of instrument strength in Figure 9.

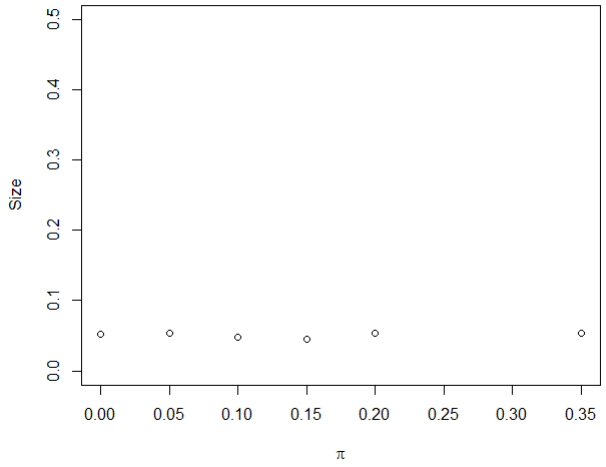

Figure 9: AR Size with $\rho = 0.99$

It is obvious that the size of AR statistic remains at 5% no matter how weak the instrument is. Therefore, researchers suggest using AR statistic for inference when the instrument is weak.

Kleibergen's K statistic, due to Kleibergen (2002), is also pivotal. The K statistic is in fact equivalent to the Lagrange Multiplier (LM) statistic, but K statistic is proposed

without distributional assumption while obviously LM statistic requires the normality assumption. The LM statistic will be introduced in the next section, where we have all the notations ready, and the equivalence between LM and K statistic will be given in the Appendix.

$$K = \frac{(y - X\beta_0)' P_{\tilde{Z}} (y - X\beta_0)}{\frac{1}{n-k}(y - X\beta_0)' M_z (y - X\beta_0)}$$

where

$$\tilde{Z} = P_z (X - (y - X\beta_0) \frac{\frac{1}{n-k}(y - X\beta_0) M_z X}{\frac{1}{n-k}(y - X\beta_0)' M_z (y - X\beta_0)})$$

Again k is the number of instruments. We can show that the limiting distribution of K is $\chi^2(1)$, irrespective of instrument strength.

The K statistic is quite similar to AR statistic, except that K statistic projects $y - X\beta_0$ to $\tilde{Z}$, which only has one column, whereas AR statistic projects $y - X\beta_0$ to Z, which has k columns. It is known that AR statistic has poor power when there are a large number of instruments, as it has a limiting distribution of $\chi^2(k)$, but the K statistic can avoid the problem as it has a limiting distribution of $\chi^2(1)$. In fact, it can be shown that the K statistic is exactly the AR statistic under just-identification.

Let us consider two conditional test, Conditional Likelihood Ratio (CLR) statistic and Conditional Wald (CW) statistic. We first introduce the idea of conditional tests. We have seen that under weak instrument asymptotics the t statistic has a limiting distribution that depends on nuisance parameters C and ρ. Therefore, the t test will not have correct size as some value of nuisance parameter can cause size to well exceed 5%, as we have shown in Figure 6. The idea of conditional test is to express the original test statistic, say T_n, as a function of two statistics, say S and T, where the distribution of S does not depend on the nuisance parameter, but the distribution of T depends on the nuisance parameter. Thus, the conditional distribution $T_n|T$ does not depend on the nuisance parameter, and we can get critical value functions that depend on T to ensure the test has correct size.

More formally, consider the reduced form IV regression:[7]

$$y_1 = Z\Pi\beta + v_1 \tag{19}$$

$$y_2 = Z\Pi + v_2 \tag{20}$$

$$\begin{pmatrix} v_{1i} \\ v_{2i} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}) = N(\mathbf{0}, \Omega) \tag{21}$$

where $v_1 = \beta v_2 + u$, u is the error term in the structural equation. Z is non-stochastic and

[7] This part borrows heavily from Xiaoxia Shi's lecture note, available here https://users.ssc.wisc.edu/~xshi/econ715/Lecture_11_WeakIV.pdf, and Moreira (2003)

Ω is known. We are interested in the following hypothesis testing problem:

$$H_0 : \beta = \beta_0 \quad \text{vs} \quad H_1 : \beta \neq \beta_0$$

We can then write down the likelihood function for $\{y_{1i}, y_{2i}\}_{i=1}^n$

$$f(y_1, y_2; \beta, \Pi) = (2\pi)^{-n}|\Omega|^{-\frac{n}{2}}\exp(-\frac{1}{2}[\sum_{i=1}^{n} Y_i'\Omega^{-1}Y_i - 2\Pi'Z'Y\Omega^{-1}A + A'\Omega^{-1}A\Pi'Z'Z\Pi])$$

where $Y_i = (y_{1i}, y_{2i})'$ and $A = (\beta, 1)'$. As Z and Ω are non stochastic, by factorization theorem, $Z'Y$ is the sufficient statistic for $(\beta, \Pi')'$. Thus, $Z'YD$, where $D = [b_0, \Omega^{-1}A_0], b_0 = (1, -\beta_0)', A_0 = (\beta_0, 1)'$, is also a sufficient statistic for $(\beta, \Pi')'$, as D is a constant matrix.

Now we have $Z'YD = [Z'(y_1 - \beta_0 y_2), Z'Y\Omega^{-1}A_0] = [S, T]$. Under H_0,

$$S = Z'(y_1 - \beta_0 y_2) \sim N(\mathbf{0}, Z'Zb_0'\Omega b_0) \tag{22}$$

$$T = Z'Y\Omega^{-1}A_0 \sim N(A_0'\Omega^{-1}A_0 Z'Z\Pi, Z'ZA_0'\Omega^{-1}A_0) \tag{23}$$

It is obvious that the distribution of S does not depend on Π, but the distribution of T depends on Π. Therefore, if we are using a test statistic $\phi(S, T)$ that has a distribution that depends on Π, as we do not know the value of Π nor can we consistently estimate Π when the instrument is weak, using such test statistic and the traditional fixed critical value will not have correct size.

There are test statistics that has a distribution that does not depend on Π, either because the test statistic is not a function of T, for example the AR statistic that we have introduced before,

$$AR(\beta_0) = \frac{(y_1 - \beta_0 y_2)'Z(Z'Z)^{-1}Z'(y_1 - \beta_0 y_2)}{\sigma_u^2} = \frac{S'(Z'Z)^{-1}S}{\sigma_u^2}$$

or the test statistic is a function of T, but the distribution does not depend on Π, for example the LM (K) statistic that we have alluded to. Define the standard statistic $\bar{T} = (Z'Z)^{-\frac{1}{2}}T(A_0'\Omega A_0)^{-\frac{1}{2}}, \bar{S} = (Z'Z)^{-\frac{1}{2}}S(b_0'\Omega^{-1}b_0)^{-\frac{1}{2}}$, we can show that the LM statistic is

$$LM = \bar{S}'\bar{T}(\bar{T}'\bar{T})^{-1}\bar{T}'\bar{S} \tag{24}$$

and since $\bar{S}$ and $\bar{T}$ are independent, LM statistic has a χ_1^2 distribution under the null, which does not depend on Π. When the model is just identified, $\bar{T}$ is a scalar, so LM statistic is equivalent to AR statistic

$$LM = \bar{S}'\bar{S} = \frac{S'(Z'Z)^{-1}S}{\sigma_u^2} = AR(\beta_0)$$

Many other test statistic, however, have nuisance parameters in the distribution. To

deal with this problem, a natural idea is to use a critical value function that depends on T, and this is exactly what Moreira (2003) does. As S and T are independent (proven in Appendix), we can consider $\phi(S,T)$ at each value $T=t$. $\phi(S,t)$ has a distribution that does not depend on Π, so we get a critical value for each value of $T=t$. Although the functional form of the critical value function could be hard to derive, especially for complicated $\phi(S,T)$, in practice we can simulate the function.

Consider the Wald statistic (square of the t statistic)

$$W=\frac{(\sqrt{n}(\hat{\beta}_{IV}-\beta_0))^2}{\widehat{\text{Asy}\text{Var}}(\hat{\beta}_{IV})}=\frac{((y_2'P_zy_2)^{-1}y_2'P_zu)^2}{\hat{\sigma}_u^2(y_2'P_zy_2)^{-1}}=\frac{(y_2'P_zu)^2}{\hat{\sigma}_u^2(y_2'P_zy_2)}$$

Under the null, $S=Z'u$ and $\hat{\sigma}_u^2=\sigma_1^2-2\hat{\beta}_{IV}\sigma_{12}+\hat{\beta}_{IV}^2\sigma_2^2$. Note both $\hat{\sigma}_u^2$ and $y_2'Z$ depend on π through $y_2'Z$, so we need to express $y_2'Z$ as a function of S and T. Note that

$$\begin{cases}S=Z'y_1-\beta_0Z'y_2\\T=c_1Z'y_1+c_2Z'y_2\end{cases}$$

where c_1 and c_2 are functions of Ω^{-1} and a_0, then we have

$$Z'y_2=\frac{T-c_1S}{c_2+c_1\beta_0}$$

Thus, we can express W as a function of S, T, Ω, and β_0. Once we conditional on $T=t$, the distribution of W does not depend on Π, and we can form test that has correct size, which is the conditional Wald (CW) statistic. Figure 10 shows the size of the CW test. It is obvious that CW test is robust to weak instruments.

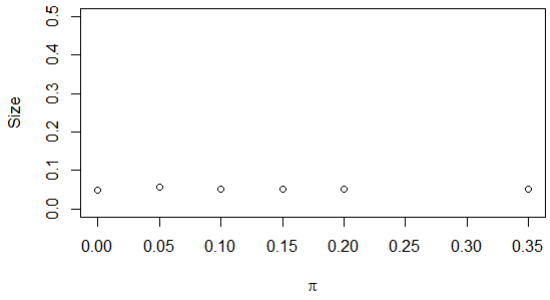

Figure 10: CW Size with $\rho=0.99$

Here we motivate the usefulness of conditional tests with extremely restrictive assumptions: finite sample normal distribution and known variance covariance matrix. This result,

however, can be easily extended to unknown distribution, as $\sigma_2, \sigma_1, \sigma_{12}$ can all be consistently estimated.

Moreira (2003) utilizes the conditional test approach to propose the Conditional Likelihood Ratio (CLR) test. The idea is to use the traditional likelihood ratio statistic

$$LR = 2[\max_{\beta,\Pi} L_n(y_1, y_2; \beta, \Pi) - \max_{\Pi} L_n(y_1, y_2; \beta_0, \Pi)]$$

to do conditional test.

We can show that

$$LR = \frac{b_0' Y' P_z Y b_0}{b_0' \Omega b_0} - \lambda_{\min}$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\Omega^{-\frac{1}{2}} Y' P_z Y \Omega^{-\frac{1}{2}}$. we can write

$$LR = \frac{1}{2}[\bar{S}'\bar{S} - \bar{T}'\bar{T} - \sqrt{(\bar{S}'\bar{S} + \bar{T}'\bar{T})^2 + 4[\bar{S}'\bar{S}\bar{T}'\bar{T} - (\bar{S}'\bar{T})^2]}] \tag{25}$$

Obviously LR is a function of S and T, so we can use the conditional test idea that we introduced before.

In the just-identified case, $\bar{S}$ and $\bar{T}$ are scalars, so we have $LR = \bar{S}'\bar{S}$, which is exactly the AR statistic, because

$$LR = \bar{S}'\bar{S} = \frac{S'(Z'Z)^{-1}S}{b_0'\Omega b_0} = \frac{S'(Z'Z)^{-1}S}{\sigma_u^2} = AR$$

Although both CW and AR have correct size, they have different power properties. In fact, Moreira (2009) shows AR test is the Uniformly Most Powerful Unbiased Test (UMPU) under just identification, where "unbiased test" means power can never be lower than size while "uniformly most powerful" means that the test always has the highest power under the constraint of size, regardless of the alternative hypothesis[8]. However, D. W. Andrews et al. (2007) shows that CW can have poor power properties. Mills et al. (2014) proposes the one-sided conditional t test and modifies the original CW to propose a new two-sided conditional t test, and show that they can perform as well as CLR. The literature except Keane and Neal (2023) does not seem to pay much attention to the one sided test and continues to recommend using AR under just-identification (I. Andrews et al., 2019). However, as I will mention below, some very recent papers provide evidence that AR may not be the best choice even under just-identification.

I will briefly introduce two recent papers (D. S. Lee et al., 2022 and D. Lee et al., 2023) that adjust critical value of the t test directly to conduct valid statistical inference. At a high level, both papers are motivated by the need by applied researchers, who prefer sticking to the familiar t test based inference and are reluctant to use robust inference methods, as

[8] See Chapter 3 of Lehmann and Romano (2022) for further discussion.

shown in Section 1. D. S. Lee et al. (2022) proposes the tF procedure, which uses a critical value function that is a smooth function of the first stage F statistic while D. Lee et al. (2023) proposes the VtF procedure, which uses a critical value function that is a function of both the first stage F statistic and the correlation between first stage and second stage error.

Let us dive deeper into D. S. Lee et al. (2022). Recall Equation (6) which shows the limiting distribution of Wald statistic under weak instrument asymptotics:

$$\hat{t}^2 \xrightarrow{d} t^2 \sim \frac{z_u^2}{1 - 2\rho\frac{z_u}{C/\sigma_v + z_v} + (\frac{z_u}{C/\sigma_v + z_v})^2} \tag{26}$$

where

$$\begin{pmatrix} z_u \\ z_v \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}), \rho = \frac{\sigma_{uv}}{\sigma_u \sigma_v}$$

Note $AR(\beta_0) \xrightarrow{d} z_u^2$ and $\hat{F} \xrightarrow{d} (\frac{C}{\sigma_v} + z_v)^2$. Following the notation in D. S. Lee et al. (2022), let $t_{AR} = z_u, f = \frac{C}{\sigma_v} + z_v, f_0 = \frac{C}{\sigma_v}$, Thus,

$$\hat{t}^2 \xrightarrow{d} t^2 \sim \frac{t_{AR}^2}{1 - 2\rho\frac{t_{AR}}{f} + (\frac{t_{AR}}{f})^2} \tag{27}$$

The goal is to ensure t test has correct size, namely under H_0,

$$P(|t| > c_\alpha) \leq \alpha$$

Figure 6 provides a heuristic[9] argument that for a fixed, weak instrument strength, larger value of ρ corresponds to higher rejection probability. Thus it is natural to impose

$$P_{\beta=\beta_0, \rho=1}(|t| > c_\alpha) = \alpha$$

Note this approach is inherently conservative as for smaller value of ρ, which as shown by J. Angrist and Kolesár (2023) may be the prevalent case in practice, the null rejection probability is less than α, which may produce unnecessarily wide confidence intervals. In fact, D. Lee et al. (2023) exactly solves this conservativeness problem, a point I will come back later.

As $\rho = 1$, $f = t_{AR} + f_0$, we have

$$t^2 \sim \frac{f^2 t_{AR}^2}{(f - t_{AR})^2} = \frac{f^2 (f - f_0)^2}{f_0^2}$$

This is a function of f, an example is given in Figure 11, with $f_0 = 3$, which corresponds

[9] Stock and Yogo (2005) initially makes the conjecture, and D. S. Lee et al. (2022) proves the result.

to $E[F] = 10$, the rule of thumb.

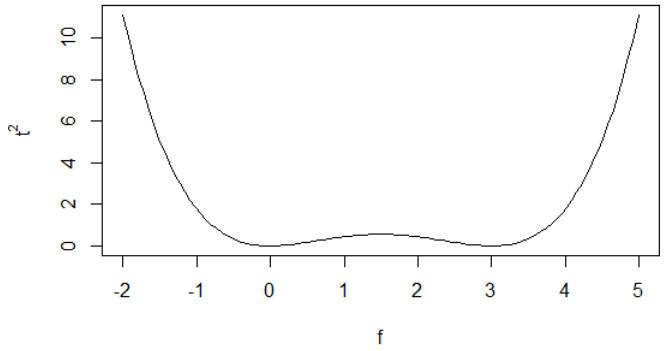

Figure 11: An Example of distribution of t^2

We know $P(|t| > c_\alpha) = P(t^2 > c_\alpha)$, so the idea to construct a critical value function that is a function of $c_\alpha(f)$, and the region where $c_\alpha(f)$ is above t^2 is exactly the acceptance region. D. S. Lee et al. (2022) restricts attention to $c_\alpha(f)$ that only intersects t^2 at two points, denoted by $\bar{f}$ and $\underline{f}$, for some technical reasons. In this way we can solve for c_α, $\bar{f}$, and $\underline{f}$ using the following system of equations:

$$\begin{cases} \frac{\bar{f}^2(\bar{f}-f_0)^2}{f_0^2} - c_\alpha(\bar{f}) = 0 \\ P(\underline{f} < f < \bar{f}) = 1-\alpha \\ \frac{\underline{f}^2(\underline{f}-f_0)^2}{f_0^2} - c_\alpha(\underline{f}) = 0 \end{cases} \tag{28}$$

For the specific solution of $c_\alpha(f)$ and its properties, see D. S. Lee et al. (2022).

An interesting feature of tF procedure is that the test is not unbiased, namely the power can dip below the nominal level α. However, AR test is only uniformly most powerful among the class of unbiased tests, so it is possible that tF is more powerful than AR in some cases. Indeed, the survey of D. S. Lee et al. (2022) shows that tF confidence interval has shorter expected length than AR.

D. Lee et al. (2023) tries to eliminate the conservativeness by imposing

$$P_{\beta=\beta_0}(|t| > c_\alpha(\rho, f)) = \alpha$$

which resembles the AR test that gives exactly level α. A notable feature is that the critical value function is both a function of both first stage strength, denoted by f, and degree of

endogeneity, denoted by ρ. The idea behind constructing the critical value function is complicated, so readers are referred to Appendix D of D. Lee et al. (2023) for further information.

This critical value function seems counter-intuitive that the function can depend on ρ. In fact, we can consistently estimate ρ under the null. Consider the reduced form regression as in Equation 17 and normalization so that $\frac{1}{n}Z'Z \xrightarrow{p} Q_{zz} = 1$, we have

$$\sqrt{n}\begin{pmatrix} \widehat{\pi\beta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix} \xrightarrow{d} N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{RF}\sigma_1\sigma_2 \\ \rho_{RF}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix})$$

There are no weak instruments problems in reduced form regressions, so $\sigma_1, \sigma_2, \rho_{RF}$ are all consistently estimable[10]. It is not difficult to show that

$$\rho = \frac{\rho_{RF} - \beta\frac{\sigma_2}{\sigma_1}}{\sqrt{1 - 2\rho_{RF}\beta\frac{\sigma_2}{\sigma_1} + \beta^2\frac{\sigma_2^2}{\sigma_1^2}}} \tag{29}$$

Thus, under the null, ρ can be consistently estimated by $\hat{\rho}$, with $\beta = \beta_0$ and $\sigma_1, \sigma_2, \rho_{RF}$ replaced by their consistent estimates.

Similar to the tF test, VtF test is also not unbiased. Indeed, VtF produced shorter confidence intervals than AR in all 10 papers surveyed by D. Lee et al. (2023).

D. Lee et al. (2023) also provides extensive simulation evidence to compare the power among different tests under just-identification, including AR, tF, VtF, and CW. VtF seems to have higher power than AR in large area of the parameter space, though it is not uniformly higher than AR as AR is admissible, shown in Chernozhukov et al. (2009). It is surprising that CW also seems to have higher power than AR for many parameter value, as it seems to contradict what is observed in D. W. Andrews et al. (2007). In fact, D. W. Andrews et al. (2007) never shows simulation evidence for just identified model, so it is not a contradiction here, though Keane and Neal (2023) argues that in just-identified model, CW (or conditional t) provides similar results as AR but is much harder to implement. In terms of the reason for the better performance of VtF and CW, similar to that of tF, all three tests are not unbiased, so they may have power lower than 5% for some alternatives, but have higher power than AR in other alternatives. It turns out that even in just identified model AR may be overly conservative. Figure 12, 13, and 14 give power curves for AR and CW under different alternatives. In the simulation, we vary the true parameter value in the grid, in this case between -1 and 1, with interval 0.1, and collect cases where the test statistic is larger than 3.84, the 95% quantile for χ_1^2. By doing so, we can get the rejection probability (power) for each alternative value. It seems that when instrument is strong, AR

[10]Intuitively, the weak instruments problem really comes from dividing the first stage, which is close to 0. As long as you are not dividing by first stage, conventional estimator works just fine.

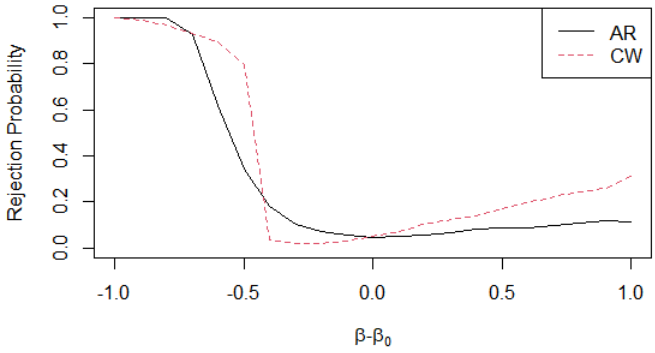

Figure 12: Power when $\pi = 0.05$

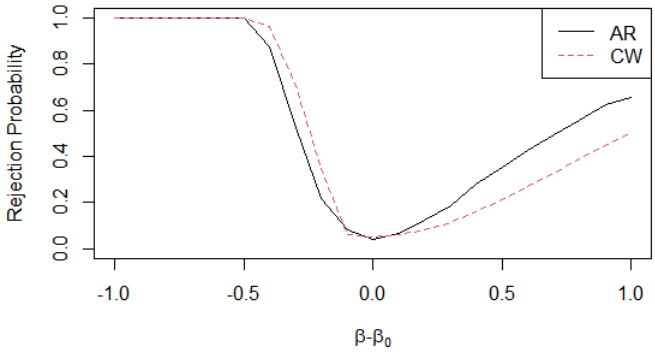

Figure 13: Power when $\pi = 0.15$

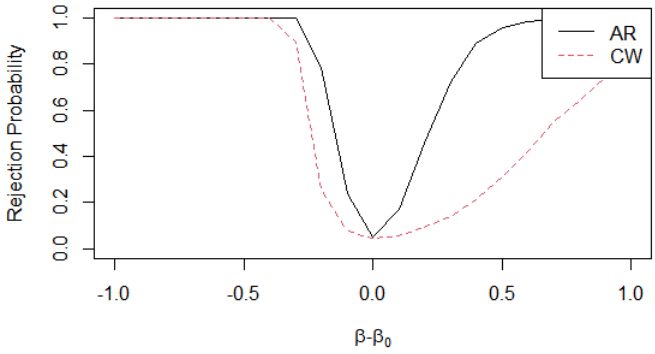

Figure 14: Power when $\pi = 0.35$

performs much better than CW, when the instrument is weak, CW is a bit better than AR.

7 Non-Linear First Stage

Assumption **1** in Section 2 makes it clear that the first stage is a projection relationship between X and Z, so it is possible that the true relationship between X and Z is nonlinear, for example $E[X_i|Z_i] = Z_i^2$. This section explores if nonlinearity could cause any problems to the estimation and inference procedure discussed before.

As a working example, consider the following model:

$$Y_i = X_i\beta + u_i \tag{30}$$

$$X_i = Z_i^2\delta + w_i \tag{31}$$

$$E[u_i|Z_i] = E[w_i|Z_i] = 0 \tag{32}$$

$$Z_i \sim N(0,1) \tag{33}$$

$$\begin{pmatrix} u_i \\ w_i \end{pmatrix} |Z_i \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uw} \\ \sigma_{uw} & \sigma_w^2 \end{pmatrix}) \tag{34}$$

The model differs from the general setup in Section 2 because Equation 31 spells out the true relationship between X and Z, rather than a pure projection from X to Z. The reason for setting $Z \sim N(0,1)$ is to ensure $Cov(X,Z) = E[XZ] - E[X]E[Z] = 0$, because $E[Z^3] = 0$ for $Z \sim N(0,1)$. In other words, there is no linear relationship between X and Z. Thus, under weak instrument asymptotics, $\hat{F} \xrightarrow{d} \chi_1^2$, and $E[F] = 1$, namely the first stage F statistic will always indicate a weak instruments problem, no matter how big δ is.

The behaviour of $\hat{\beta} = (Z'X)^{-1}Z'Y$ is also quite interesting. In fact, it resembles the behaviour when the instrument is weak, for example Equation 7. It is simple to show that

$$\hat{\beta} - \beta = \frac{Z'u}{Z'X} \xrightarrow{d} \frac{\phi_{zu}}{\phi_{zx}} = \frac{\sigma_{uw}}{\delta^2 E[Z_i^6] + \sigma_w^2} + \frac{\xi}{\phi_{zx}} \tag{35}$$

$$\begin{pmatrix} \phi_{zu} \\ \phi_{zx} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uw} \\ \sigma_{uw} & \delta^2 E[Z_i^6] + \sigma_w^2 \end{pmatrix})$$

$$\begin{pmatrix} \xi \\ \phi_{zx} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 - \frac{\sigma_{uw}^2}{\delta^2 E[Z_i^6] + \sigma_w^2} & 0 \\ 0 & \delta^2 E[Z_i^6] + \sigma_w^2 \end{pmatrix})$$

When δ is small, the limiting distribution of $\hat{\beta} - \beta$ again resembles a Cauchy distribution plus a biased term, which means $\hat{\beta}$ will have large median bias. When δ is big, $\hat{\beta} - \beta$ still has a scaled Cauchy as the limiting distribution, but without the biased term , so $\hat{\beta}$ will be

approximately median unbiased. Figure 16 illustrates the point.

Next we examine the limiting distribution of Wald statistic (t^2). It is not difficult to show that

$$t^2 = \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{\sigma}_u^2 (Z'Z)(Z'X)^{-2}}} \tag{36}$$

$$\xrightarrow{d} \frac{\phi_{zu}^2}{(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6] + \sigma_w^2} + \frac{\xi}{\phi_{zx}})^2 (\delta^2 E[Z_i^4] + \sigma_w^2) - 2\sigma_{uw}(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6] + \sigma_w^2} + \frac{\xi}{\phi_{zx}}) + \sigma_u^2} \tag{37}$$

$$\hat{\sigma}_u^2 = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \tag{38}$$

When δ is small, namely the instrument provides little information, We would expect a similar behaviour of t test as discussed in Section 3, namely there is substantial size distortion. When δ is large, the denominator will be quite large, so t^2 will be close to 0. Figure 15 confirms the above findings. The simulation setup for Figure 15 and 16 is: Sample Size $N = 1000$, Iterations $S = 5000$, $\delta \in \{0.1, 0.5, 1, 5, 10\}$, and

$$\begin{pmatrix} u_i \\ w_i \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix})$$

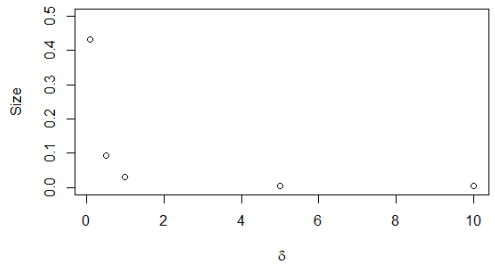

Figure 15: Size of t test with $\rho = 0.99$

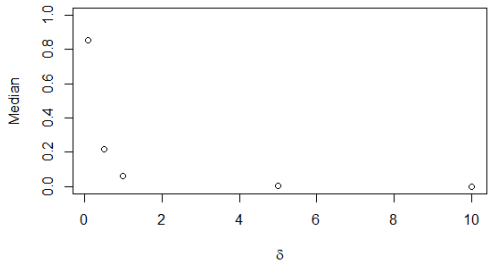

Figure 16: Median bias of $\hat{\beta}$

Obviously t test may not be a desirable choice in this context. AR test is a natural alternative. Figure 17 shows that AR test has correct size regardless of δ.

As a side note, in this working example we have shown that even when the first stage F statistic is quite small, $\hat{\beta}$ may have low bias and t test may have correct size. This result matches the pattern shown in Figure 3 of I. Andrews et al. (2019), which is a calibrated simulation to AER papers. Figure 3 of I. Andrews et al. (2019) shows that the first stage F statistic is small, the median bias of $\hat{\beta}$ and size of t test can be quite volatile, ranging from unbiased and correct size to heavily biased and large size distortion.

In summary, it seems that the issue of using traditional TSLS and t test when the true

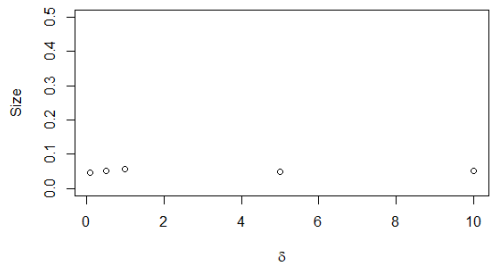

Figure 17: Size of AR test with $\rho = 0.99$

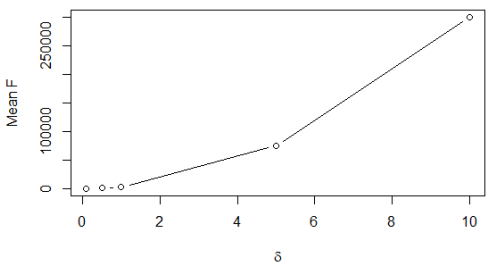

Figure 18: $E[F_{nonlinear}]$ and δ

first stage is nonlinear in Z is that the t-test will converge to 0 when δ is large and the first stage F statistic does not indicate instrument strength. We may want to correctly specify $E[X|Z]$ and estimate δ, then derive the first stage F statistic that is suitable for this case. It is simple to show that when applying OLS on $X_i = Z_i^2\delta + w_i$,

$$\hat{\delta} = (\sum_{i=1}^{n} Z_i^4)^{-1} \sum_{i=1}^{n} Z_i^2 X_i$$

and if we use the weak instrument asymptotics, namely $\delta = \frac{C}{\sqrt{n}}$, we have

$$\hat{F}_{nonlinear} = \hat{\delta}'[\widehat{Var}(\hat{\delta})]^{-1}\hat{\delta} \xrightarrow{d} (\frac{CE[Z_i^4]^{\frac{1}{2}}}{\sigma_w} + N(0,1))^2 \overset{d}{=} F_{nonlinear}$$

The concentration parameter $\mu^2 = \frac{C^2 E[Z_i^4]}{\sigma_w^2}$, $E[F_{nonlinear}] = \mu^2 + 1$. Figure 18 shows how $E[F_{nonlinear}]$ evolves with δ, obviously $F_{nonlinear}$ regains the ability to indicate identification strength after correctly specifying $E[X|Z]$

We may also want to see if correctly specifying $E[X|Z]$ can bring some other benefits. If we follow the idea of "Two Stage Least Square", namely obtaining the fitted value by running OLS in the first stage, then using the fitted value as an IV[11] for β

$$\hat{\beta}_{nonlinear} = (\sum_{i=1}^{n} \hat{X}_i' X_i)^{-1} \sum_{i=1}^{n} \hat{X}_i Y_i$$

[11]The idea of using first stage fitted value as an regressor in the second stage and running OLS in the second stage also works in this case, but is susceptible to misspecification (for example specifying $E[X|Z] = \Phi(\pi Z)$ but in fact $E[X|Z] = \frac{\exp(\pi Z)}{1+\exp(\pi Z)}$). The intuition is that the estimated residual in the first stage using probit or logit or other nonlinear models under misspecification is not guaranteed to uncorrelated with the fitted value $\hat{X}$, thus plugging in fitted value in the second stage and run OLS may not be valid. This is termed "Forbidden Regression". See J. D. Angrist and Pischke (2009), pp. 143-144 for discussion.

where $\hat{X}_i = Z_i^2\hat{\delta}$. It is simple to show

$$\hat{\beta}_{nonlinear} \xrightarrow{p} \beta \tag{39}$$

$$\sqrt{n}(\hat{\beta}_{Nonlinear} - \beta) = (\frac{1}{n}\sum_{i=1}^{n} Z_i^2 X_i)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i^2 u_i \xrightarrow{d} N(0, \sigma_u^2 E[Z_i^4] E[Z_i^2 X_i]^{-2}) \tag{40}$$

so we have

$$t_{nonlinear} = \frac{\hat{\beta}_{nonlinear} - \beta_0}{\sqrt{\hat{\sigma}_u^2 (\sum_{i=1}^{n} Z_i^4)(\sum_{i=1}^{n} Z_i^2 X_i)^{-2}}} \xrightarrow{d} N(0,1)$$

$$\hat{\sigma}_u^2 = \frac{1}{n}(Y - X\hat{\beta}_{Nonlinear})'(Y - X\hat{\beta}_{Nonlinear})$$

Figure 19 and 20[12] show the median bias of $\hat{\beta}_{nonlinear}$ and size of t test based on $t_{nonlinear}$. Reassuringly, when δ is large, $\hat{\beta}_{nonlinear}$ is approximately unbiased and the size of t test based on $t_{nonlinear}$ is approximately 5%. An interesting observation is that both the estimator and test are more robust to weak instruments, in the sense that when $\delta = 0.1$, a situation where the instrument is already weak enough (as shown in Figure 15 and 16, there is substantial size distortion and bias), $\hat{\beta}_{nonlinear}$ is still approximately unbiased and t test using $t_{nonlinear}$ still has correct size. Of course, when the instrument becomes even weaker, these methods will not give correct estimate and inference. Thus, we can conclude that correctly specifying the first stage will give valid first stage F statistic, $\hat{\beta}$ and t test when δ is large, and seems to be more robust to small δ compared to conventional TSLS estimate.

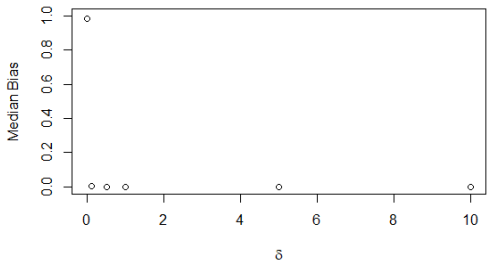

Figure 19: Median bias of $\hat{\beta}_{nonlinear}$

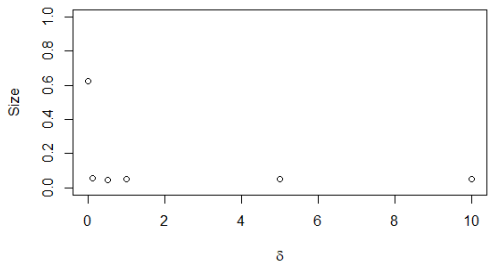

Figure 20: Size of t test using $t_{nonlinear}$

However, in practice we do not know whether true first stage is quadratic or cubic or some other complicated forms. It will have a high cost if we use run OLS on $X_i = Z_i^2\delta + w_i$ but in fact the model is misspecified. For example, if the $E[X_i|Z_i] = Z_i\delta$, $E[F_{nonlinear}]$ and $\hat{\beta}_{nonlinear}$ will behave as in Figure 21, 22, and 23, which does a bad job compared with using

[12]In Figure 19 and 20 only, I add a situation where $\delta = 0$, in order to show that correctly specifying the first stage still cannot solve the fundamental problem of unidentification

conventional TSLS.

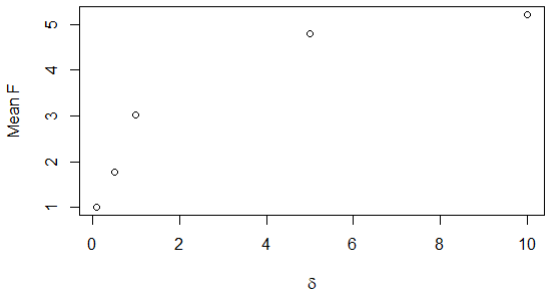

Figure 21: $E[F_{nonlinear}]$ when the true model is linear

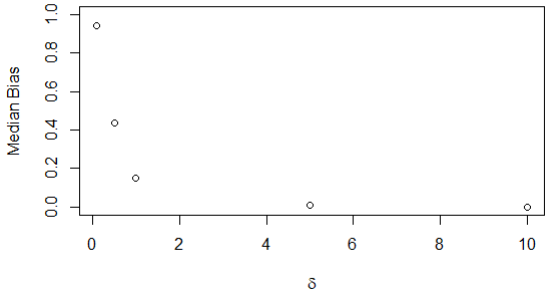

Figure 22: Median bias of $\hat{\beta}_{nonlinear}$ when the true model is linear

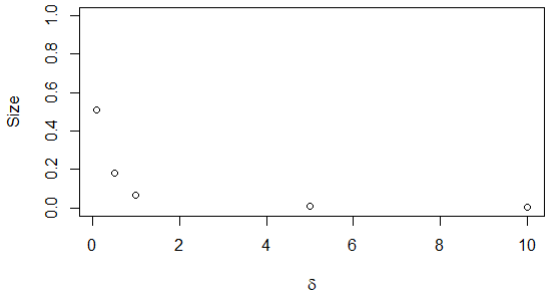

Figure 23: Size of t test when the true model is linear

To tackle misspecification we may want to recall the essence of TSLS estimation. In fact, it is shown in Equation 43 (proof is given in the Appendix) that TSLS replaces endogenous

X with exogenous $E[X|Z]$. So our goal is to correctly estimate $E[X|Z]$[13].

$$\beta_{TSLS} = E[\hat{X}^2]^{-1}E[\hat{X}Y], \hat{X} = E[X|Z] \tag{41}$$
$$= E[E[X|Z]^2]^{-1}E[E[X|Z]Y] \tag{42}$$
$$= \beta \tag{43}$$

Naturally, we want to use a nonparametric first stage, namely creating a dummy variable for each unique value of the instrument. For example, if Z has 10 unique values, we need to create 9 dummy variables (Omitting one to avoid multicollinearity). Each dummy variable takes value 1 if the original instrument takes this unique value, 0 otherwise. Using a nonparametric first stage will give $E[X|Z]$ automatically, no matter what specific functional form the true first stage is. However, an immediate issue of this approach is that Z may take a lot of distinct values, so the first stage will ultimately have a large number of regressors, creating a serious many instruments bias. In fact, the instrument used in the main IV regression of Acemoglu et al. (2001) has virtually as many distinct values as the number of observations. Though there are instruments that only take a few distinct values, for example the quarter of birth instrument used in J. D. Angrist and Krueger (1991), which only has 4 distinct values.

A simple simulation shows the problem of many instrument bias. I cannot use the previous simulation design to generate Z because Z will have as many distinct value as observations, so I designate that Z takes 20 distinct value[14], the total number of observations is 500. The true first stage is given by $X_i = \log Z_i + w_i, E[w_i|Z_i] = 0$, the first stage that I actually run is $X_i = \alpha + \pi'(I_{\{Z_i=1\}}, I_{\{Z_i=1.25\}}, ..., .I_{\{Z_i=5.5\}}) + v_i, E[Z_i v_i] = 0$[15]. $\delta = \{0.05, 0.1, 0.3, 0.5, 0.7, 1\}$, number of iterations is 2000, error variance is 1 and correlation is 0.99.

Figure 24 and 25 compare the size of t test and median bias of $\hat{\beta}$ when using traditional linear first stage and fully saturated first stage. It is clear that there is a substantial many instruments bias.

Jackknife Instrumental Variable Estimator (JIVE) and LIML introduced in Equation 15 are two popular alternatives when facing many instruments bias

$$\hat{\beta}_{JIVE} = \frac{\sum_i \sum_{j \neq i} P_{ij} X_i Y_j}{\sum_i \sum_{j \neq i} P_{ij} X_i X_j}$$

where P_{ij} is the (i, j) element of $P_z = Z(Z'Z)^{-1}Z'$. Figure 26 and 27 show the median

[13] Chen et al. (2020) also examines the problem of potential nonlinear relationship between X and Z in the first stage, and proposes to use machine learning methods to estimate $E[X|Z]$

[14] The values are {1.00,1.25,1.50,1.75,2.00,2.25,2.50,2.75,3.00,3.25,3.50,3.75,4.00,4.25,4.50,4.75,5.00,5.25,5.50,5.75 }. Then I repeat such sequence 25 times to generate a total observation of 500.

[15] $E[v_i|Z_i] = 0$ is true by construction

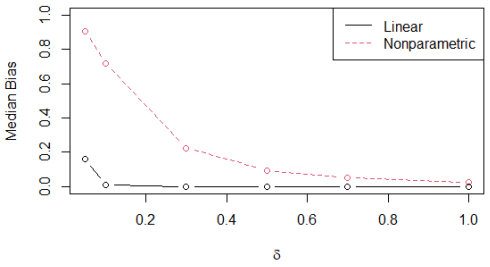

Figure 24: Median Bias

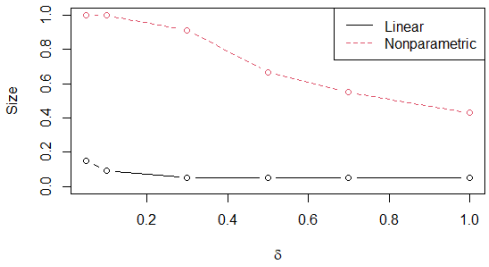

Figure 25: Size of t test

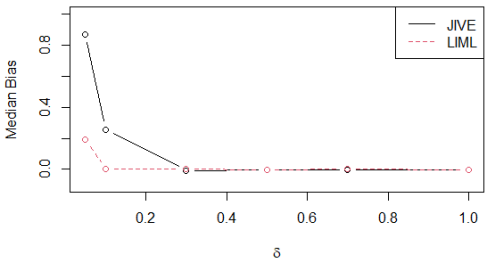

Figure 26: Median bias

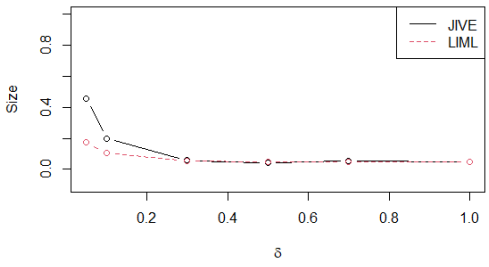

Figure 27: Size of Wald test

bias of both JIVE and LIML estimator, and the size of Wald test based on them. Clearly, both perform better than running naive TSLS in the many instruments context, with LIML doing better than JIVE. This is a little surprising as JIVE is known to be more robust than LIML in the many instruments context, see Mikusheva and Sun (2022). An explanation is that when $\delta = 0.05$ or 0.1, the instrument is already too weak to allow any consistent estimate, so none of the estimate is reliable. It could also be possible that the nonlinear first stage contributes to the result.

Thus, in this simulation, it seems that using a nonparametric first stage together with robust estimators like JIVE or LIML does not do better than running linear first stage with TSLS in terms of bias or size. Researchers sometimes also care about interpretation of the parameter estimate. If we add covariates and allow heterogeneous treatment effect, nonparametric first stage combined with JIVE may be a better choice as it allows us to interpret the parameter estimate as a positively weighted average of Local Average Treatment effect (LATE)[16]. Thus, which method to use in practice depends on which goal the researcher wants to achieve. It might be a good practice to try all these methods and let the reader decide which one is most useful.

To sum up, $E[X|Z]$ being nonlinear in Z can cause problem in estimation and inference to traditional methods. When we try to estimate $E[X|Z]$, machine learning methods may help. A nonparametric first stage together with JIVE/LIML may also solve this problem, but it may also do worse (when there are too many covariate bins). Probably it is a good idea to try all these methods and let the reader decide which one is most useful.

8 Extension to Heteroskedastic Models

I spent most of the paper summarizing results under the assumption of homoskedasticity, mainly for simplicity. However, in practice it is extremely rare to see applied researchers calculating any estimator or test statistic under homoskedasticity. In fact, only one paper in my *AER* sample reports first stage F statistic under homoskedasticity. Therefore, in this section I will introduce the extension of aforementioned estimators and test statistic to heteroskedastic case. In fact, in just identified models, such extension is straightforward.

Let us consider the first stage F statistic. We have shown in Equation (11) that in just-identified model, the first stage F statistic has a limiting distribution of $\chi^2_1((\frac{C}{\sigma_v})^2)$ under homoskedasticity. Under heterskedasticity,

$$\hat{F}_{\text{robust}} = ((Z'Z)^{-1}Z'X)'((Z'Z)^{-1}\sum_{i=1}^{n} Z_i^2\hat{v}_i^2(Z'Z)^{-1})^{-1}(Z'Z)^{-1}Z'X \tag{44}$$

[16]Unfortunately, the interpretation of LIML is tricky, as it is shown in Kolesar (2013) that LIML may be outside the convex hull of LATE

$$\xrightarrow{d} \chi^2_1((\frac{E[Z_i^2]C}{\sqrt{E[Z_i^2 v_i^2]}})^2) \tag{45}$$

$$\hat{v}_i = X_i - Z_i\hat{\pi} \tag{46}$$

Obviously if $E[Z_i^2 v_i^2] = E[Z_i^2]\sigma_v^2$, we are back to the homoskedastic case, assuming $E[Z_i^2] = 1$. As pointed out by I. Andrews et al. (2019), applied researchers normally use the heteroskedasticity robust first stage F, but refer to the "rule of thumb" 10, which is derived under homoskedasticity. Thus, new critical value for the weak instruments test is required to accommodate the heteroskedastic case. Olea and Pflueger (2013) derived the "Effective first stage F", or "F_{Eff}" statistic based on controlling Nagar bias, without assuming homoskedasticity.

More formally, consider the reduced form regression as in Equation 17 with normalized instruments, namely $\frac{1}{n}Z'Z = I_k$

$$\begin{pmatrix} \frac{1}{\sqrt{n}}Z'v_1 \\ \frac{1}{\sqrt{n}}Z'v_2 \end{pmatrix} \xrightarrow{d} N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} W_1 & W_{12} \\ W_{12}' & W_2 \end{pmatrix}) \tag{47}$$

$$F_{Eff} = \frac{X'ZZ'X}{n\text{tr}(\hat{W}_2)}$$

In just-identified model,

$$\text{tr}(\hat{W}_2) = \frac{1}{n}\sum_{i=1}^{n} Z_i^2\hat{v}_i^2$$

, so $F_{Eff} = \hat{F}_{\text{robust}}$. In just identified setting, to ensure the Nagar bias is less than 10%, the effective (robust) first stage F statistic should be at least 23.1, much larger than 10, the threshold for having 10% Nagar bias under homoskedastic model and also the rule of thumb. In fact, I. Andrews (2018) shows in the appendix that there are cases where the homoskedastic first stage F statistic is 500 but the size of t test is still larger than 15%, though the result is obtained with heavily over-identified models. For our purpose, a good example of the importance of using heteroskedastic robust First stage F statistic is the setting in Section 7, where the true first stage is nonlinear in Z. Consider

$$X_i = Z_i\pi + v_i$$
$$X_i = Z_i^2\delta + w_i$$
$$E[e_i|Z_i] = E[Z_i v_i] = 0$$
$$Z \sim N(0,1)$$

In this setting, $\pi = 0$, so $v_i = X_i$. Even though we assume w_i is conditionally homoskedastic, namely $E[w_i^2|Z_i] = \sigma_w^2$, it is not possible for v_i to be conditional homoskedastic, because

$E[v_i^2|Z_i] = Var[v_i|Z_i] + E^2[v_i|Z_i] = \sigma_w^2 + (Z_i^2\delta)^2$, which is a function of Z_i. Figure 28 shows the expected value of robust and non-robust first stage F statistic. We know that the correct answer should be 1, but the non-robust first stage F statistic obviously is too large, which could be misleading.

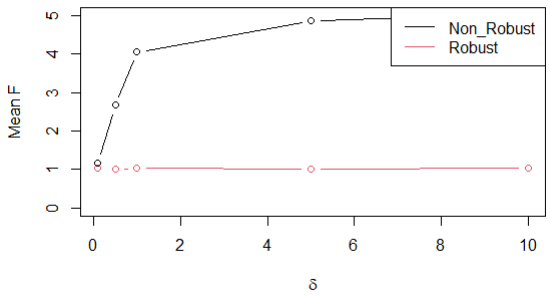

Figure 28: Robust F vs Non-Robust F

Next consider the AR statistic under $H_0 : \beta = \beta_0$

$$AR(\beta_0) = (y - X\beta_0)'Z(Z'Z)^{-1}((Z'Z)^{-1}\sum_{i=1}^{n}\hat{u}_i^2 Z_i^2 (Z'Z)^{-1})^{-1}(Z'Z)^{-1}Z'(y - X\beta_0) \quad (48)$$

$$= \frac{\frac{1}{n}u'ZZ'u}{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 Z_i^2} \quad (49)$$

$$\xrightarrow{d} \frac{\psi_{zu}^2}{E[Z_i^2 u_i^2]} \sim \chi_1^2 \quad (50)$$

$$\hat{u}_i = y_i - X_i\beta_0 \quad (51)$$

Thus AR still has a χ_1^2 distribution under the null, so critical value does not change. The only change applied researchers need to make in practice is to compute the AR statistic with heteroskedastic error, as the original AR statistic does not have χ_1^2 as limiting distribution

$$AR_{\text{homo}}(\beta_0) = \frac{(y - X\beta_0)'P_z(y - X\beta_0)}{\frac{1}{n-1}(y - X\beta_0)'M_z(y - X\beta_0)}$$
$$\xrightarrow{d} \frac{\psi_{zu}^2}{\sigma_u^2 E[Z_i^2]}$$

9 Empirical Example

The most famous empirical example considered in the weak instruments literature is the J. D. Angrist and Krueger (1991) study that uses quarter of birth as an instrument for years of schooling to study return to education. However, the IV model in J. D. Angrist

and Krueger (1991) is over-identified, which does not suit the purpose of this paper. Thus, I pick Acemoglu et al. (2001), another famous IV application but with just-identified model. It is shown in the original paper that the instrument is relatively strong, so I artificially create an instrument that is weakly correlated with the endogenous regressor, and apply the estimation and inference methods introduced in previous sections.

Acemoglu et al. (2001) estimates the effect of institution quality (measured by expropriation risk) on economic growth, exploiting European settlers' mortality rate as an instrument. They estimate the effect using the following equation:

$$logy_i = \alpha_1 + \beta R_i + e_i$$
$$R_i = \alpha_2 + \pi logM_i + v_i$$

where y_i is GDP growth rate, R_i is protection against expropriation, M_i is mortality rate for European settlers.

The relevance of the instrument comes from the fact that European settlers adopted more extractive institutions at places where their mortality is high, and the past institutions can affect current institutions.

To examine the consequences of weak instruments, I mimic the idea of Bound et al. (1995) and generate irrelevant instruments, in this case I let $logM_i \sim U[0,1]$ and estimate the same equation using the new irrelevant instruments.

Table 1: Estimated effects of average protection against expropriation on economic growth

	Original		
	OLS	TSLS	Unbiased
Coef	0.52	0.94	0.91
Std Error	(0.05)	(0.18)	
First F		16.85	
TSLS Confidence Interval		[0.63,1.25]	
AR Confidence Interval		[0.71,1.42]	
CW Confidence Interval		[0.65,1.30]	
VtF Confidence Interval		[0.65,1.30]	

We can see from Table 1 that the original equation has a relatively large first stage F statistic, so if we believe the validity of instrument, we should be confident that the true effect is around 1, which is nearly two times as large as the OLS estimate. The unbiased estimate is also quite close to the TSLS estimate, confirming the result in I. Andrews and Armstrong (2017) which shows that the unbiased estimate has the same asymptotic behaviour as the TSLS estimate under strong instrument. As the instrument is relatively strong, we can see that the naive TSLS confidence interval is not much different from the

Table 2: Estimated effects of average protection against expropriation on economic growth

		Simulated	
	OLS	TSLS	Unbiased
Coef	0.52	6.45	1.30
Std Error	(0.05)	(50.98)	
First F		0.01	
TSLS Confidence Interval		[-93.47,106.47]	
AR Confidence Interval		$(-\infty, +\infty)$	
CW Confidence Interval		$(-\infty, 0.45] \cup [0.51, +\infty)$	
VtF Confidence Interval		$(-\infty, +\infty)$	

confidence interval that we derived from other robust procedures.

For the construction of confidence intervals, the TSLS confidence interval is derived analytically using

$$[\hat{\beta}_{TSLS} - 1.96\sqrt{\frac{\widehat{\text{Asy}}\text{Var}(\hat{\beta}_{TSLS})}{n}}, \hat{\beta}_{TSLS} + 1.96\sqrt{\frac{\widehat{\text{Asy}}\text{Var}(\hat{\beta}_{TSLS})}{n}}]$$

The confidence interval for AR, CW, and VtF are computed using grid search. In this case, I specify the grid to be between -10 and 10, with interval 0.01, and collect parameter value in sets $\{\beta : AR(\beta) \leq 3.84\}, \{\beta : CW(\beta) \leq 3.84\}, \{\beta : VtF(\beta) \leq 3.84\}$ respectively.

Table 2 shows the instrument is completely irrelevant to the endogenous regressor, so the TSLS estimate is extremely biased. The unbiased estimate, however, is much more robust to weak instrument and is only slightly biased. As the instrument is completely irrelevant, we should have no ability to identify the parameter of interest, so it makes sense that the robust confidence interval virtually contains the whole real line.

A single point estimate and standard error tells us extremely little about the distribution of the estimate. Thus, we calibrate our simulation to the original data to examine the effect of weak instruments on the distribution of the estimate in this empirical context. The idea is: Consider running the following first stage and reduce form regression

$$Y_i = Z_i\pi\beta + \epsilon_i \tag{52}$$

$$X_i = Z_i\pi + v_i \tag{53}$$

then we obtain parameter estimate $\widehat{\pi\beta}$ and $\hat{\pi}$, and the estimate of their variance-covariance matrix $\begin{pmatrix} \hat{\sigma}_\epsilon^2 & \hat{\sigma}_{\epsilon v} \\ \hat{\sigma}_{\epsilon v} & \hat{\sigma}_v^2 \end{pmatrix}$. Note all reduced form quantities can be consistently estimated under mild assumptions, so no need to worry about weak instruments problem here. Then we view these estimates as population quantity, and draw $N = 1000$ pairs of $(\pi\beta^*, \pi^*)$ from

this bivariate normal distribution

$$N(\begin{pmatrix} \widehat{\pi\beta} \\ \hat{\pi} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}_\epsilon^2 & \hat{\sigma}_{\epsilon v} \\ \hat{\sigma}_{\epsilon v} & \hat{\sigma}_v^2 \end{pmatrix})$$

and compute $\hat{\beta}_{IV}$ and standard error for each iteration.

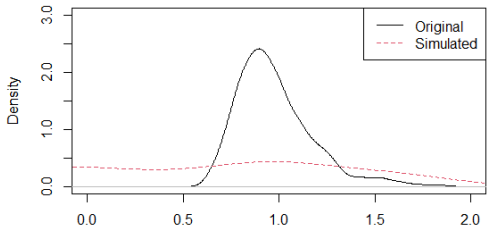

Figure 29: Simulated distribution of TSLS

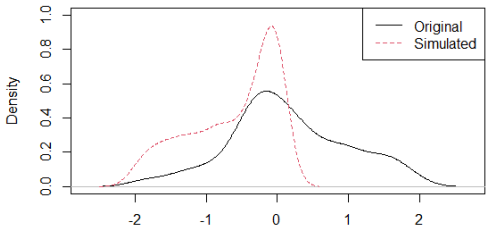

Figure 30: Simulated distribution of t statistic

Figure 29 and 30 show the simulated distribution of TSLS and t statistic for original instrument and simulated instrument respectively. It is clear that the TSLS estimate under simulated instrument is extremely dispersed and is no way close to the parameter estimate 6.45, suggesting that 6.45 is not a reliable estimate. In contrast, the distribution of TSLS using original estimate is approximately centered around 0.9, the original parameter estimate, suggesting a relatively reliable estimate. The t statistic under original estimate also behaves much closer to a standard normal compared to the t statistic under the simulated instrument. Thus, such simulations confirm that the parameter estimate and inference is relatively reliable using original instrument, while the simulated instrument induces severe bias.

We can see from Table 1 and 2 that confidence interval based on robust test has three shapes: bounded interval, real line, and real line except a bounded interval. As we have discussed, confidence interval being a bounded interval or the real line is not surprising, but being the real line except a bounded interval seems counter-intuitive. In fact, if we examine the form of AR statistic

$$AR(\beta_0) = \frac{(y - X\beta_0)' P_z (y - X\beta_0)}{\frac{1}{n-1}(y - X\beta_0)' M_z (y - X\beta_0)}$$

and consider the confidence set $\{\beta : AR(\beta) < \chi_1^2\}$, we have that all β in the confidence set

satisfies the following quadratic inequality

$$\frac{(y - X\beta)' P_z (y - X\beta)}{\frac{1}{n-1}(y - X\beta)' M_z (y - X\beta)} < 3.84$$

$$(nX' P_z X - 3.84 X' M_z X)\beta^2 - 2(nX' P_z y - 3.84 X' M_z y)\beta + ny' P_z y - 3.84 y' M_z y < 0$$

Thus, it is not surprising that AR test can have confidence interval that is the real line except a bounded interval. In our case, it is the CW test that shows such a shape of confidence interval. I am not aware of literature discussing the shape of confidence interval of CW test, but I guess it follows the same pattern as AR. Davidson and MacKinnon (2014) and Zivot et al. (1998) further discuss the shape of AR confidence interval.

10 Directions for Further Research

Although Moreira (2009) shows that AR test is UMPU under just-identification, it does not say anything about tests that are biased, namely tests that have power lower than nominal size α under certain alternatives. Indeed, D. S. Lee et al. (2022) and D. Lee et al. (2023) show that biased tests may have higher power than AR in many alternative values, as shown in Figure 12, 13, and 14, and have shorter expected length of confidence interval, whenever the confidence interval is bounded.[17] Thus, it cases doubt on the recommendation of AR test in just identified model. It seems interesting to explore if other tests can further improve power and produce shorter expected confidence intervals.

Both I. Andrews and Armstrong (2017) and J. Angrist and Kolesár (2023) study if imposing sign restriction on the first stage could give any new results. Indeed, imposing sign restrictions gives unbiased estimator (I. Andrews and Armstrong, 2017), and minimizes median bias (J. Angrist and Kolesár (2023)). J. Angrist and Kolesár (2023) also argue for using t test anyway because it is unlikely to have large degree of endogeneity in practice, though this point is not supported by D. S. Lee et al. (2022). Thus, it seems interesting to explore if there is a general theory that can speak to the reasonablenss of imposing such restrictions. For example, we have shown in Equation 29 that there is a one-to-one relationship between the degree of endogeneity and the parameter value of interest. It will be interesting if we can assess the possible value of β, and if it corresponds to a ρ that is low enough, t test will provide valid inference (though its power may be worse than other procedures, such as "VtF").

[17] Dufour (1997) shows that confidence interval that has correct coverage under potentially weak instruments must have infinite expected length, so comparing the unconditional expected length of confidence interval among valid tests is not useful, as all of them are infinite.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91*(5), 1369–1401. https://doi.org/10.1257/aer.91.5.1369

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63. https://doi.org/10.1214/aoms/1177730090

Andrews, D. W., & Guggenberger, P. (2009). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, *26*(2), 426–468. https://doi.org/10.1017/s0266466609100051

Andrews, D. W., Moreira, M. J., & Stock, J. H. (2007). Performance of conditional wald tests in iv regression with weak instruments. *Journal of Econometrics*, *139*(1), 116–132. https://doi.org/10.1016/j.jeconom.2006.06.007

Andrews, I. (2018). Valid two-step identification-robust confidence sets for gmm. *The Review of Economics and Statistics*, *100*(2), 337–348. https://doi.org/10.1162/rest_a_00682

Andrews, I., & Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign: Unbiased iv estimation. *Quantitative Economics*, *8*(2), 479–503. https://doi.org/10.3982/qe700

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angrist, J., & Kolesár, M. (2023). One instrument to rule them all: The bias and coverage of just-ID IV. *Journal of Econometrics*, 105398. https://doi.org/10.1016/j.jeconom.2022.12.012

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Angrist, J. D., Pathak, P. A., & Zarate, R. A. (2023). Choice and consequence: Assessing mismatch at chicago exam schools. *Journal of Public Economics*, *223*, 104892. https://doi.org/10.1016/j.jpubeco.2023.104892

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657. https://doi.org/10.2307/2951662

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous ex-

planatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443. https://doi.org/10.2307/2291055

Chen, J., Chen, D. L., & Lewis, G. (2020). Mostly harmless machine learning: Learning optimal instruments in linear iv models. https://doi.org/10.48550/ARXIV.2011.06158

Chernozhukov, V., Hansen, C., & Jansson, M. (2009). Admissible invariant similar tests for instrumental variables regression. *Econometric Theory*, *25*(3), 806–818. https://doi.org/10.1017/s0266466608090312

Davidson, R., & MacKinnon, J. G. (2014). Confidence sets based on inverting anderson–rubin tests. *The Econometrics Journal*, *17*(2), S39–S58. https://doi.org/10.1111/ectj.12015

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, *65*(6), 1365. https://doi.org/10.2307/2171740

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica*, *70*(1), 163–189. https://doi.org/10.1111/1468-0262.00272

Hausman, J., Stock, J. H., & Yogo, M. (2005). Asymptotic properties of the hahn–hausman test for weak-instruments. *Economics Letters*, *89*(3), 333–342. https://doi.org/10.1016/j.econlet.2005.06.007

Hirano, K., & Porter, J. R. (2014). Location properties of point estimators in linear instrumental variables and related models. *Econometric Reviews*, *34*(6–10), 720–733. https://doi.org/10.1080/07474938.2014.956573

Keane, M., & Neal, T. (2023). Instrument strength in IV estimation and inference: A guide to theory and practice. *Journal of Econometrics*, *235*(2), 1625–1653. https://doi.org/10.1016/j.jeconom.2022.12.009

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, *70*(5), 1781–1803. https://doi.org/10.1111/1468-0262.00353

Kolesar, M. (2013). *Estimation in an instrumental variables model with treatment effect heterogeneity* (tech. rep.).

Lee, D., McCrary, J., Moreira, M., Porter, J., & Yap, L. (2023, November). *What to do when you can't use "1.96" confidence intervals for iv*. https://doi.org/10.3386/w31893

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2022). Valid t-ratio inference for iv. *American Economic Review*, *112*(10), 3260–3290. https://doi.org/10.1257/aer.20211063

Lehmann, E., & Romano, J. (2022). *Testing statistical hypotheses*. Springer International Publishing. https://books.google.com/books?id=ZNJ2EAAAQBAJ

Mikusheva, A., & Sun, L. (2022). Inference with many weak instruments. *The Review of Economic Studies*, *89*(5), 2663–2686. https://doi.org/https://doi.org/10.1093/restud/rdab097

Mills, B., Moreira, M. J., & Vilela, L. P. (2014). Tests based on t-statistics for iv regression with weak instruments. *Journal of Econometrics*, *182*(2), 351–363. https://doi.org/10.1016/j.jeconom.2014.03.012

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, *71*(4), 1027–1048. https://doi.org/10.1111/1468-0262.00438

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, *152*(2), 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012

Olea, J. L. M., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business and Economic Statistics*, *31*(3), 358–369. https://doi.org/10.1080/00401706.2013.806694

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557. https://doi.org/10.2307/2171753

Stock, J. H., & Yogo, M. (2005, June). Testing for weak instruments in linear IV regression. In *Identification and inference for econometric models* (pp. 80–108). Cambridge University Press. https://doi.org/10.1017/cbo9780511614491.006

Zivot, E., Startz, R., & Nelson, C. R. (1998). Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review*, *39*(4), 1119. https://doi.org/10.2307/2527355

Appendix

Proof of the distribution of IV estimator when the instrument is irrelevant in Equation 3

$$
\begin{aligned}
\hat{\beta} - \beta &= \frac{Z'u}{Z'X} \\
&= \frac{Z'u}{Z'v} \\
&= \frac{\frac{1}{\sqrt{n}}Z'u}{\frac{1}{\sqrt{n}}Z'v} \\
&\xrightarrow{d} \frac{\psi_{zu}}{\psi_{zv}} \\
\begin{pmatrix}\psi_{zu}\\ \psi_{zv}\end{pmatrix} &\sim N(\begin{pmatrix}0\\0\end{pmatrix}, E[z_i^2]\begin{pmatrix}\sigma_u^2 & \sigma_{uv}\\ \sigma_{uv} & \sigma_v^2\end{pmatrix}), E[z_i^2] = 1 \\
\psi_{zu} &= E[\psi_{zu}|\psi_{zv}] + \eta \\
&= \frac{\sigma_{uv}}{\sigma_v^2}\psi_{zv} + \eta, \psi_{zv} \perp \eta \\
Var(\psi_{zu}) &= Var(\frac{\sigma_{uv}}{\sigma_v^2}\psi_{zv}) + Var(\eta) \\
\sigma_u^2 &= \frac{\sigma_{uv}^2}{\sigma_v^4}\sigma_v^2 + Var(\eta) \\
Var(\eta) &= \sigma_u^2(1-\rho^2), \rho = \frac{\sigma_{uv}}{\sigma_u\sigma_v} \\
\Rightarrow \hat{\beta} - \beta &\xrightarrow{d} \frac{\sigma_{uv}}{\sigma_v^2} + \frac{\eta}{\psi_{zv}} \\
\begin{pmatrix}\eta\\ \psi_{zv}\end{pmatrix} &\sim N(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_u^2(1-\rho^2) & 0\\ 0 & \sigma_v^2\end{pmatrix})
\end{aligned}
$$

Proof of Equation 6 the distribution of t^2 statistic under weak instrument asymptotics

$$
\begin{aligned}
t^2 &= (\frac{\sqrt{n}(\hat{\beta}_{IV} - \beta_0)}{\sqrt{\hat{\text{Asy}\text{Var}}(\hat{\beta}_{IV})}})^2 \\
&= (\frac{Z'u/Z'X}{\sqrt{\hat{\sigma}_u^2(Z'X)^{-2}Z'Z}})^2 \\
&= (\frac{Z'u}{\sqrt{\hat{\sigma}_u^2 Z'Z}} \cdot \frac{|Z'X|}{Z'X})^2 \\
\hat{\sigma}_u^2 &= \frac{1}{n}(y - X\hat{\beta}_{IV})'(y - X\hat{\beta}_{IV})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}[X(\beta-\hat{\beta}_{IV})+u]'[X(\beta-\hat{\beta}_{IV})+u] \\
&= \frac{1}{n}[(\beta-\hat{\beta}_{IV})^2X'X + 2X(\beta-\hat{\beta}_{IV})u + u'u] \\
&= \frac{1}{n}(\beta-\hat{\beta}_{IV})^2(Z\pi+v)'(Z\pi+v) + 2\frac{1}{n}(\beta-\hat{\beta}_{IV})(Z\pi+v)u + \frac{1}{n}u'u \\
&= \frac{1}{n}(\frac{Z'u}{Z'X})^2(Z\frac{C}{\sqrt{n}}+v)'(Z\frac{C}{\sqrt{n}}+v) - 2\frac{1}{n}\frac{Z'u}{Z'X}(Z\frac{C}{\sqrt{n}}+v)u + \frac{1}{n}u'u \\
&\xrightarrow{d} \sigma_u^2[(\frac{z_u}{C/\sigma_v+z_v})^2 - 2\rho\frac{z_u}{C/\sigma_v+z_v}+1] \\
t^2 &= (\frac{Z'u}{\sqrt{\hat{\sigma}_u^2 Z'Z}}\cdot\frac{|Z'X|}{Z'X})^2 \\
&\xrightarrow{d} \frac{z_u^2}{(\frac{z_u}{C/\sigma_v+z_v})^2 - 2\rho\frac{z_u}{C/\sigma_v+z_v}+1}
\end{aligned}
$$

Proof of $\sqrt{n}(\hat{\beta}-\frac{1}{\hat{c}}) \xrightarrow{p} 0$ in Equation 14

$$
\begin{aligned}
\hat{\beta} &= (X'P_ZX)^{-1}X'P_ZY \\
\hat{c} &= (Y'P_ZY)^{-1}Y'P_ZX \\
\sqrt{n}(\hat{\beta}-\frac{1}{\hat{c}}) &= \sqrt{n}(\beta+(X'P_ZX)^{-1}X'P_Zu-\beta-(Y'P_ZX)^{-1}Y'P_Zu) \\
&= \sqrt{n}((X'P_ZX)^{-1}X'P_Zu-(Y'P_ZX)^{-1}Y'P_Zu) \\
&= \frac{n^{-\frac{3}{2}}[X'P_ZuY'P_ZX - Y'P_ZuX'P_ZX]}{n^{-2}X'P_ZXY'P_ZX} \\
&= \frac{①}{②} \\
② &= n^{-2}X'P_ZX(X\beta+u)'P_ZX \\
&= \beta(\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X)^2 + \frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X\frac{1}{n}u'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X \\
&\xrightarrow{p} \beta E[X_iZ_i']E[Z_iZ_i']^{-1}E[Z_iX_i] \\
① &= [\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{\sqrt{n}}Z'u][\beta\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X + \frac{1}{n}u'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X)] \\
&- [\beta\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{\sqrt{n}}Z'u + \frac{1}{n}u'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{\sqrt{n}}Z'u][\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X] \\
&= [\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{\sqrt{n}}Z'u][\frac{1}{n}u'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X] \\
&- [\frac{1}{n}u'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{\sqrt{n}}Z'u][\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X] \\
&\xrightarrow{p} 0 \\
\sqrt{n}(\hat{\beta}-\frac{1}{\hat{c}}) &\xrightarrow{p} 0, \quad \text{By CMT for } \xrightarrow{p}
\end{aligned}
$$

Proof of $\hat{\beta}_{LIML} = \hat{\beta}_{2SLS}$ in just identified model

The idea is to exploit the fact that $\kappa \geq 1$, so we assume $\kappa = 1$, and see if the equation holds. If the equation holds, it means that when the model is just-identified, the smallest root of the equation is $\kappa = 1$, so $\kappa = 1$, and $\hat{\beta}_{LIML} = \hat{\beta}_{2SLS}$

$$\begin{aligned}
\det(\bar{Y}'\bar{Y} - \bar{Y}'M_z\bar{Y}) &= \det(\bar{Y}'P_z\bar{Y}) \\
&= \det(\begin{bmatrix} Y'P_zY & Y'P_zX \\ X'P_zY & X'P_zX \end{bmatrix}) \\
&= Y'P_zYX'P_zX - (Y'P_zX)^2 \\
&= (Y'Z)^2(X'Z)^2(Z'Z)^{-2} - (Y'Z)^2(X'Z)^2(Z'Z)^{-1} \\
&= 0
\end{aligned}$$

The reason we can have the second to last equality is that the model is just identified, so $Z'Z, Z'X, Z'Y$ are all scalars.

Proof of the $LM \sim \chi^2_1$ distribution

The idea is to first show $\bar{S}$ and $\bar{T}$ are independent, which is equivalent to show S and T are independent, then show $LM|T = t \sim \chi^2_1$, thus $LM \sim \chi^2_1$.

$$S = Z'(y_1 - \beta_0 y_2), T = \frac{1}{|\Omega|}Z'[(\beta_0\sigma_2^2 - \sigma_{12})y_1 + (\sigma_1^2 - \beta_0\sigma_{12})y_2]$$

Both S and T are normally distributed, to show they are independent, it is sufficient to show $Cov(S,T) = 0$

$$\begin{aligned}
Cov(S,T) &= \frac{1}{|\Omega|}Cov(Z'(y_1 - \beta_0 y_2), Z'[(\beta_0\sigma_2^2 - \sigma_{12})y_1 + (\sigma_1^2 - \beta_0\sigma_{12})y_2]) \\
&= \frac{1}{|\Omega|}\sum_{i=1}^{n} Z_i^2[(\beta_0\sigma_2^2 - \sigma_{12})Var(y_{1i}) - \beta_0(\sigma_1^2 - \beta_0\sigma_{12})Var(y_{2i}) \\
&+ (\sigma_1^2 - \beta_0\sigma_{12} - \beta_0(\beta_0\sigma_2^2 - \sigma_{12}))Cov(y_{1i}, y_{2i})] \\
&= \frac{1}{|\Omega|}\sum_{i=1}^{n} Z_i^2[(\beta_0\sigma_2^2 - \sigma_{12})\sigma_1^2 - \beta_0(\sigma_1^2 - \beta_0\sigma_{12})\sigma_2^2 \\
&+ (\sigma_1^2 - \beta_0\sigma_{12} - \beta_0(\beta_0\sigma_2^2 - \sigma_{12}))\sigma_{12}] \\
&= 0
\end{aligned}$$

Thus, S and T are independent, which implies $\bar{S}$ and $\bar{T}$ are independent.

As $\bar{S}$ and $\bar{T}$ are independent, we can consider the conditional distribution $LM|T=t$

$$\begin{aligned} LM|(T=t) &= \bar{S}'\bar{t}(\bar{t}'\bar{t})^{-1}\bar{t}'\bar{S} \\ \bar{t}'\bar{S} &\sim N(0, \bar{t}'\bar{t}) \\ (\bar{t}'\bar{t})^{-\frac{1}{2}}\bar{t}'\bar{S} &\sim N(0,1) \\ \bar{S}'\bar{t}(\bar{t}'\bar{t})\bar{t}'\bar{S} &\sim \chi_1^2 \end{aligned}$$

Thus, $LM \sim \chi_1^2$

Proof of Equivalence between Kleibergen's K statistic and LM statistic

Note K statistic is derived without distributional assumption, so in the proof all the error variance in the LM statistic will be replace by their consistent estimate under the null.

$$\begin{aligned} \bar{S}'\bar{T} &= ((Z'Z)^{-\frac{1}{2}}Z'Yb_0(b_0'\hat{\Omega}b_0)^{-\frac{1}{2}})'(Z'Z)^{-\frac{1}{2}}Z'Y\hat{\Omega}^{-1}A_0(A_0\hat{\Omega}^{-1}A_0)^{-\frac{1}{2}}) \\ &= (b_0'\hat{\Omega}b_0)^{-1}|\hat{\Omega}|^{\frac{1}{2}}b_0'Y'Z(Z'Z)^{-1}Z'Y\hat{\Omega}^{-1}A_0 \\ &= (b_0'\hat{\Omega}b_0)^{-1}|\hat{\Omega}|^{-\frac{1}{2}}(y-X\beta_0)'P_z[(\beta_0\hat{\sigma}_2^2-\hat{\sigma}_{12})y+(\hat{\sigma}_1^2-\beta_0\sigma_{12})X] \\ &= (b_0'\hat{\Omega}b_0)^{-1}|\hat{\Omega}|^{-\frac{1}{2}}(y-X\beta_0)'P_z\cdot(1) \\ (1) &= (\beta_0\hat{\sigma}_2^2-\hat{\sigma}_{12})y+(\hat{\sigma}_1^2-\beta_0\hat{\sigma}_{12})X \\ &= (\beta_0\hat{\sigma}_2^2-\hat{\sigma}_{12})[y-X\beta_0+\frac{\hat{\sigma}_1^2-2\beta_0\hat{\sigma}_{12}+\beta_0^2\hat{\sigma}_2^2}{\beta_0\hat{\sigma}_2^2-\hat{\sigma}_{12}}X] \\ S_{uv} &= \frac{1}{n}(y-X\beta_0)'M_zX \\ &= \frac{1}{n}(v_1-\beta_0v_2)'M_zv_2 \\ &= \frac{1}{n}v_1'M_zv_2-\frac{1}{n}\beta_0v_2'M_zv_2 \\ &= \hat{\sigma}_{12}-\beta_0\hat{\sigma}_2^2 \\ S_{uu} &= \frac{1}{n}(y-X\beta_0)'M_z(y-X\beta_0) \\ &= \frac{1}{n}(v_1-\beta_0v_2)'M_z(v_1-\beta_0v_2) \\ &= \frac{1}{n}v_1'M_zv_1-2\beta_0\frac{1}{n}v_1'M_zv_2+\beta_0^2\frac{1}{n}v_2'M_zv_2 \\ &= \hat{\sigma}_1^2-2\beta_0\hat{\sigma}_{12}+\beta_0^2\hat{\sigma}_2^2 \\ &= b_0'\hat{\Omega}b_0 \\ (1) &= -S_{uv}[y-X\beta_0-\frac{S_{uu}}{S_{uv}}X] \\ &= S_{uu}[X-(y-X\beta_0)\frac{S_{uv}}{S_{uu}}] \end{aligned}$$

$$\bar{S}'\bar{T} = |\hat{\Omega}|^{-\frac{1}{2}}(y - X\beta_0)'P_z[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}]$$

$$\begin{aligned}
\bar{T}'\bar{T} &= ((Z'Z)^{-\frac{1}{2}}Z'Y\hat{\Omega}^{-1}A_0(A_0'\hat{\Omega}^{-1}A_0)^{-\frac{1}{2}})'((Z'Z)^{-\frac{1}{2}}Z'Y\hat{\Omega}^{-1}A_0(A_0'\hat{\Omega}^{-1}A_0)^{-\frac{1}{2}}) \\
&= (A_0'\hat{\Omega}^{-1}A_0)^{-1}A_0'\hat{\Omega}^{-1}Y'Z(Z'Z)^{-1}Z'Y\hat{\Omega}^{-1}A_0 \\
&= |\hat{\Omega}|(b_0'\hat{\Omega}^{-1}b_0)^{-1}\cdot(1)'\cdot P_z\cdot(1) \\
&= |\hat{\Omega}|S_{uu}[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}]'P_z[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}] \\
\frac{(\bar{S}'\bar{T})^2}{\bar{T}'\bar{T}} &= \frac{1}{S_{uu}}\frac{(y - X\beta_0)'(P_z[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}])(P_z[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}])'(y - X\beta_0)}{(P_z[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}])'(P_z[X - (y - X\beta_0\frac{S_{uv}}{S_{uu}})])} \\
&= \frac{(y - X\beta_0)'P_{\tilde{Z}}(y - X\beta_0)}{\frac{1}{n}(y - X\beta_0)'M_z(y - X\beta_0)} \\
&= K \\
\tilde{Z} &= P_z[X - (y - X\beta_0)\frac{S_{uv}}{S_{uu}}]
\end{aligned}$$

We have shown that LM statistic has a limiting distribution of $\chi^2(1)$ (replacing reduced form error variance with consistent estimate) and that LM and K statistic are equivalent, so we K statistic also has a limiting distribution of $\chi^2(1)$. Kleibergen (2002) gives an independent proof of K statistic having limiting distribution of $\chi^2(1)$ in the appendix.

Proof of the LR statistic as in Equation 25

$$\begin{aligned}
f(y_1, y_2; \beta, \Pi) &= (2\Pi)^{-n}|\Omega|^{-\frac{n}{2}}\exp(-\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)'\Omega^{-1}(Y_i - \mu)) \\
L &= \log f(y_1, y_2; \beta, \Pi) \\
&= \log((2\pi)^{-n}|\Omega|^{-\frac{n}{2}}) - \frac{1}{2}\sum_{i=1}^{n}[Y_i'\Omega^{-1}Y_i - 2(Z_i'\Pi A)'\Omega^{-1}Y_i + (Z_i'\Pi A)'\Omega^{-1}Z_i'\Pi A] \\
&= C - \frac{1}{2}\sum_{i=1}^{n}Y_I'\Omega^{-1}Y_i + \sum_{i=1}^{n}(Z_i'\Pi A)'\Omega^{-1}Y_i - \frac{1}{2}\sum_{i=1}^{n}(Z_i\Pi A)'\Omega^{-1}Z_i'\Pi A \\
&= C - \frac{1}{2}\mathrm{tr}(Y\Omega^{-1}Y') + \Pi'Z'Y\Omega^{-1}A - \frac{1}{2}\Pi'Z'Z\Pi A'\Omega^{-1}A \\
\frac{\partial L}{\partial \Pi}\Big|_{\hat{\Pi}} &= Z'Y\Omega^{-1}A - Z'Z\hat{\Pi}A'\Omega^{-1}A = 0 \\
\hat{\Pi} &= (Z'Z)^{-1}Z'Y\Omega^{-1}A(A'\Omega^{-1}A)^{-1} \\
L &= C - \frac{1}{2}\mathrm{tr}(Y\Omega^{-1}Y') + \hat{\Pi}'Z'Y\Omega^{-1}A - \frac{1}{2}\hat{\Pi}'Z'Z\hat{\Pi}A'\Omega^{-1}A \\
&= C - \frac{1}{2}\mathrm{tr}(Y\Omega^{-1}Y') + \frac{1}{2}\frac{A'\Omega^{-1}Y'P_zY\Omega^{-1}A}{A'\Omega^{-1}A}
\end{aligned}$$

$$
\begin{aligned}
\hat{\Pi}A' &= (Z'Z)^{-1}Z'Y\Omega^{-1}A(A'\Omega^{-1}A)^{-1}A' \\
&= (Z'Z)^{-1}Z'Y\frac{1}{|\Omega|}\begin{bmatrix}\beta\sigma_2^2-\beta\sigma_{12} & \beta\sigma_2^2-\sigma_{12}\\ \beta\sigma_1^2-\beta^2\sigma_{12} & \beta_1^2-\beta\sigma_{12}\end{bmatrix}(A'\Omega^{-1}A)^{-1} \\
I-b(b'\Omega b)^{-1}b'\Omega &= (b'\Omega b)^{-1}(b'\Omega bI-bb'\Omega) \\
&= (b'\Omega b)^{-1}\begin{bmatrix}\beta\sigma_2^2-\beta\sigma_{12} & \beta\sigma_2^2-\sigma_{12}\\ \beta\sigma_1^2-\beta^2\sigma_{12} & \beta_1^2-\beta\sigma_{12}\end{bmatrix} \\
\hat{\Pi}A' &= (Z'Z)^{-1}Z'Y[I-b(b'\Omega b)^{-1}b'\Omega] \\
L &= \hat{\Pi}'Z'Y\Omega^{-1}A-\frac{1}{2}\hat{\Pi}'Z'Z\hat{\Pi}A'\Omega^{-1}A' \\
&= \sum_{i=1}^{n}Y_i'\Omega^{-1}(\hat{\Pi}A')'Z_i-\frac{1}{2}\sum_{i=1}^{n}Z_i'\hat{\Pi}A'\Omega^{-1}(\hat{\Pi}A')'Z_i \\
&= \sum_{i=1}^{n}Y_i'\Omega^{-1}((Z'Z)^{-1}Z'Y[I-b(b'\Omega^{-1}b)b'\Omega])'Z_i \\
&\quad-\frac{1}{2}\sum_{i=1}^{n}Z_i'(Z'Z)^{-1}Z'Y[I-b(b'\Omega b)^{-1}b'\Omega]\Omega^{-1}[(Z'Z)^{-1}Z'Y(I-b(b'\Omega^{-1}b'\Omega)]'Z_i \\
&= ①+② \\
① &= \sum_{i=1}^{n}Y_i'\Omega^{-1}Y'Z(Z'Z)^{-1}Z_i-\sum_{i=1}^{n}Y_i'b(b'\Omega^{-1}b)^{-1}b'Y'Z(Z'Z)^{-1}Z_i \\
&= \sum_{i=1}^{n}\mathrm{tr}(Y_i'\Omega^{-1}Y'Z(Z'Z)^{-1}Z_i)-\sum_{i=1}^{n}\mathrm{tr}(Y_ib(b'\Omega^{-1}b)^{-1}b'Y'Z(Z'Z)^{-1}Z_i) \\
&= \sum_{i=1}^{n}\mathrm{tr}(\Omega^{-1}Y'Z(Z'Z)^{-1}Z_iY_i')-\sum_{i=1}^{n}\mathrm{tr}(b(b'\Omega^{-1}b')b'Y'Z(Z'Z)^{-1}Z_iY_i' \\
&= \mathrm{tr}(\Omega^{-1}Y'Z(Z'Z)^{-1}Z'Y)-\mathrm{tr}((b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z'Yb) \\
② &= -\frac{1}{2}\sum_{i=1}^{n}Z_i'(Z'Z)^{-1}Z'Y\Omega^{-1}Y'Z(Z'Z)^{-1}Z_i \\
&\quad-\frac{1}{2}\sum_{i=1}^{n}Z_i'(Z'Z)^{-1}Z'Yb(b'\Omega b)^{-1}b'\Omega\Omega^{-1}\Omega b(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z_i \\
&\quad+\sum_{i=1}^{n}Z_i'(Z'Z)^{-1}Z'Y\Omega^{-1}\Omega b(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z_i \\
&= ③+④+⑤ \\
③ &= -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}(Z_i'(Z'Z)^{-1}Z'Y\Omega^{-1}Y'Z(Z'Z)^{-1}Z_i) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}((Z'Z)^{-1}Z'Y\Omega^{-1}Y'Z(Z'Z)^{-1}Z_iZ_i')
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\mathrm{tr}((Z'Z)^{-1}Z'Y\Omega^{-1}Y'Z) \\
&= -\frac{1}{2}\mathrm{tr}(\Omega^{-1}Y'Z(Z'Z)^{-1}Z'Y) \\
④ &= -\frac{1}{2}\sum_{i=1}^{n} Z_i'(Z'Z)^{-1}Z'Yb(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z_i \\
&= -\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}(Z_i'(Z'Z)^{-1}Z'Yb(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z_i) \\
&= -\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}((Z'Z)^{-1}Z'Yb(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z_iZ_i') \\
&= -\frac{1}{2}\mathrm{tr}((b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z'Yb) \\
⑤ &= \sum_{i=1}^{n} \mathrm{tr}((Z'Z)^{-1}Z'Yb(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z_iZ_i') \\
&= \mathrm{tr}(Z'Yb(b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}) \\
&= \mathrm{tr}((b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z'Yb) \\
② &= -\frac{1}{2}\mathrm{tr}(\Omega^{-1}Y'Z(Z'Z)^{-1}Z'Y) + \frac{1}{2}\mathrm{tr}((b'\Omega^{-1}b)^{-1}b'Y'Z(Z'Z)^{-1}Z'Yb) \\
L &= ① + ② \\
&= \frac{1}{2}\mathrm{tr}(\Omega^{-1}Y'Z(Z'Z)^{-1}Z'Y) - \frac{1}{2}\mathrm{tr}((b'\Omega b)^{-1}b'Y'Z(Z'Z)^{-1}Z'Yb) \\
&= \frac{1}{2}\mathrm{tr}(\Omega^{-1}Y'P_zY) - \frac{1}{2}\frac{b'Y'P_zYb}{b'\Omega b} \\
LR &= 2(L_{\text{unrestrict}} - L_{\text{restrict}}) \\
&= \frac{b_0'YP_zYb_0}{b_0'\Omega b_0} - \lambda_{\min} \\
&= \bar{S}'\bar{S} - \lambda_{\min}
\end{aligned}
$$

where $\lambda_{\min}$ is the smallest root of equation $|Y'P_zY - \lambda\Omega| = 0$

or is the smallest eigenvalue of $\Omega^{-\frac{1}{2}}Y'P_zY\Omega^{-\frac{1}{2}}$

$$
\begin{aligned}
(\bar{S},\bar{T}) &= (Z'Z)^{-\frac{1}{2}}Z'Y\Omega^{-\frac{1}{2}}J \\
J &= [\Omega^{\frac{1}{2}}b_0(b_0'\Omega b_0)^{-\frac{1}{2}}, \Omega^{-\frac{1}{2}}A_0(A_0'\Omega^{-1}A_0)] \quad \text{is orthogonal} \\
(\bar{S},\bar{T})'(\bar{S},\bar{T}) &= J'\Omega^{-\frac{1}{2}}Y'P_zY\Omega^{-\frac{1}{2}}J \\
\Omega^{-\frac{1}{2}}Y'P_zYP_z\Omega^{-\frac{1}{2}}x &= \lambda x \\
\implies J'\Omega^{-\frac{1}{2}}Y'P_zY\Omega^{-\frac{1}{2}}J &= J'J\lambda x \\
&= \lambda x
\end{aligned}
$$

$$
\begin{aligned}
&\implies (\bar{S},\bar{T})'(\bar{S},\bar{T}) \text{ has the same eigenvalue as } \Omega^{-\frac{1}{2}}Y'P_zY\Omega^{-\frac{1}{2}} \\
&0 = |(\bar{S},\bar{T})'(\bar{S},\bar{T}) - \lambda I)| \\
&0 = \lambda^2 - (\bar{S}'\bar{S} + \bar{T}'\bar{T})\lambda + \bar{S}'\bar{S}\bar{T}'\bar{T} - (\bar{S}'\bar{T})^2 \\
&\lambda_{\min} = \frac{1}{2}[\bar{S}'\bar{S} + \bar{T}'\bar{T} - \sqrt{(\bar{S}'\bar{S} + \bar{T}'\bar{T})^2 - 4(\bar{S}'\bar{S}\bar{T}'\bar{T} - (\bar{S}'\bar{T})^2)}]
\end{aligned}
$$

Proof of $LM = \bar{S}'\bar{T}(\bar{T}'\bar{T})^{-1}\bar{T}'\bar{S}$ as in Equation 24

Note: To get the formula $LM = \bar{S}'\bar{T}(\bar{T}'\bar{T})^{-1}\bar{T}'\bar{S}$, it seems that a slightly modified version of the information matrix is required. I believe that the formula appeared in Moreira (2009) is correct, as when the model is just identified, it reduces to the AR statistic. However, I just cannot get back to this exact formula when using the traditional information matrix. I will make this point explicit in the derivation.

$$
\begin{aligned}
L &= C - \frac{1}{2}\text{tr}(Y\Omega^{-1}Y') + \Pi'Z'Y\Omega^{-1}A - \frac{1}{2}\Pi'Z'Z\Pi A'\Omega^{-1}A \\
&= C - \frac{1}{2}\text{tr}(Y'\Omega^{-1}Y) \\
&+ \frac{1}{|\Omega|}\Pi'Z'[(\beta\sigma_2^2 - \sigma_{12})y + (\sigma_1^2 - \beta\sigma_{12})X] - \frac{1}{2}\frac{1}{|\Omega|}(\beta^2\sigma_2^2 - 2\beta\sigma_{12} + \sigma_1^2)\Pi'Z'Z\Pi \\
\frac{\partial}{\partial\beta}\Big|_{\beta_0,\hat{\Pi}} &= \frac{1}{|\Omega|}\hat{\Pi}'Z'(\sigma_2^2 y - \sigma_{12}X) - \frac{1}{2}\frac{1}{|\Omega|}(2\sigma_2\beta_0 - 2\sigma_{12})\hat{\Pi}'Z'Z\hat{\Pi} \\
&= (A_0'\Omega^{-1}A_0)^{-1}A_0'\Omega^{-1}Y'P_z\frac{1}{|\Omega|}[\sigma_2^2 y - \sigma_{12}X] \\
&- \frac{1}{|\Omega|}(\sigma_2^2\beta_0 - \sigma_{12})(A_0'\Omega^{-1}A_0)^{-2}A_0'\Omega^{-1}y'P_z y\Omega^{-1}A_0 \\
&= \frac{1}{|\Omega|}(b_0'\Omega b_0)^{-1}[(\beta_0\sigma_2^2 - \sigma_{12})y + (\sigma_1^2 - \beta_0\sigma_{12})X]'P_z[\sigma_2^2 y - \sigma_{12}X] \\
&- \frac{1}{|\Omega|}(\sigma_2^2\beta_0 - \sigma_{12})(b_0'\Omega b_0)^{-1}[(\beta_0\sigma_2^2 - \sigma_{12})y + (\sigma_1^2 - \beta_0\sigma_{12})X]'P_z\cdot \\
&[(\beta_0\sigma_2^2 - \sigma_{12})y + (\sigma_1^2 - \beta_0\sigma_{12})X] \\
&= \frac{1}{|\Omega|}(b_0'\Omega^{-1}b_0)^{-1}[(\beta_0\sigma_2^2 - \sigma_{12})y + (\sigma_1^2 - \beta_0\sigma_{12})X]'P_z\cdot \\
&[\sigma_2^2 y - \sigma_{12}X - \frac{\beta_0\sigma_2^2 - \sigma_{12}}{b_0'\Omega b_0}((\beta_0\sigma_2^2 - \sigma_{12})y - (\sigma_1^2 - \beta_0\sigma_{12})X)] \\
&= \frac{1}{|\Omega|(b_0'\Omega b_0)}|\Omega|A_0'\Omega^{-1}Y'P_z\cdot \\
&[\frac{\sigma_2^2 b_0'\Omega b_0 - (\beta_0\sigma_2^2 - \sigma_{12})^2}{b_0'\Omega b_0}y - \frac{\sigma_{12}b_0'\Omega b_0 + (\beta_0\sigma_2^2 - \sigma_{12})(\sigma_1^2 - \beta_0\sigma_{12}))}{b_0'\Omega b_0}X] \\
&= \frac{1}{b_0'\Omega b_0}A_0'\Omega^{-1}Y'P_z\cdot[\frac{|\Omega|}{b_0'\Omega b_0}y - \frac{|\Omega|}{b_0'\Omega b_0}\beta_0 X]
\end{aligned}
$$

$$= |\Omega| \frac{A_0' \Omega^{-1} Y' P_z Y b_0}{(b_0' \Omega b_0)^2}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta^2}\Big|_{\beta_0, \hat{\Pi}} &= -\frac{\sigma_2^2}{|\Omega|} \hat{\Pi}' Z' Z \hat{\Pi} \\
&= -\frac{\sigma_2^2}{|\Omega|} \frac{A_0' \Omega^{-1} Y' P_z Y \Omega^{-1} A_0}{(A_0' \Omega^{-1} A_0)^2} \\
\bar{T}' \bar{S} &= ((Z'Z)^{-1} Z' Y \Omega^{-1} A_0 (A_0' \Omega^{-1} A_0)^{-\frac{1}{2}})' (Z'Z)^{-\frac{1}{2}} Z' Y b_0 (b_0' \Omega b_0)^{-\frac{1}{2}} \\
&= |\Omega|^{\frac{1}{2}} \frac{A_0' \Omega^{-1} Y' P_z Y b_0}{b_0' \Omega b_0} \\
\bar{T}' \bar{T} &= ((Z'Z)^{-\frac{1}{2}} Z' Y \Omega^{-1} A_0 (A_0' \Omega^{-1} A_0)^{-\frac{1}{2}})' (Z'Z)^{-\frac{1}{2}} Z' Y \Omega^{-1} A_0 (A_0' \Omega^{-1} A_0)^{-\frac{1}{2}} \\
&= \frac{A_0' \Omega^{-1} Y' P_z Y \Omega^{-1} A_0}{A_0' \Omega^{-1} A_0} \\
(\frac{\partial}{\partial \beta}\Big|_{\beta_0, \hat{\Pi}})^2 &= (\bar{T}' \bar{S})^2 \cdot \frac{|\Omega|}{(b_0' \Omega b_0)^2} \\
(-\frac{\partial}{\partial \beta^2}\Big|_{\beta_0, \hat{\Pi}})^{-1} &= (\bar{T}' \bar{T})^{-1} \cdot \frac{|\Omega|}{\sigma_2^2} A_0' \Omega^{-1} A_0 \\
(\frac{\partial}{\partial \beta}\Big|_{\beta_0, \hat{\Pi}})^2 (-\frac{\partial}{\partial \beta^2}\Big|_{\beta_0, \hat{\Pi}})^{-1} &= \frac{|\Omega|}{\sigma_2^2 (b_0' \Omega b_0)} \bar{S}' \bar{T} (\bar{T}' \bar{T})^{-1} \bar{T}' \bar{S} \\
|\Omega| - \sigma_2^2 (b_0' \Omega b_0) &= \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 - \sigma_2^2 (\sigma_1^2 - 2\beta_0 \sigma_{12} + \beta_0^2 \sigma_2^2) \\
&= -\beta_0^2 \sigma_2^4 + 2\beta_0 \sigma_2^2 \sigma_{12} - \sigma_{12}^2 \\
&= -(\beta_0 \sigma_2^2 - \sigma_{12})^2 \\
&\neq 0 \quad \text{unless the degree of endogeneity is 0}
\end{aligned}$$

Thus, it seems that the Moreira (2009) implicitly uses a scaled version of the information matrix, but I am not sure.

Proof of the relationship between ρ and ρ_{RF} in Equation 29

$$\begin{aligned}
\rho &= \frac{Cov(u, v_2)}{\sqrt{Var(u) Var(v_2)}} \\
Cov(u, v_2) &= Cov(v_1 - \beta v_2, v_2) \\
&= Cov(v_1, v_2) - \beta Var(v_2) \\
&= \rho_{RF} \sigma_1 \sigma_2 - \beta \sigma_2^2 \\
Var(u) &= Var(v_1 - \beta v_2) \\
&= \sigma_1^2 - 2\beta \rho_{RF} \sigma_1 \sigma_2 + \beta^2 \sigma_2^2
\end{aligned}$$

$$
\begin{aligned}
\rho &= \frac{\rho_{RF}\sigma_1\sigma_2 - \beta\sigma_2^2}{\sqrt{(\sigma_1^2 - 2\beta\rho_{RF}\sigma_1\sigma_2 + \beta^2\sigma_2^2)\sigma_2^2}} \\
&= \frac{(\rho_{RF} - \beta\frac{\sigma_2}{\sigma_1})\sigma_1\sigma_2}{\sqrt{(1 - 2\beta\rho_{RF}\frac{\sigma_2}{\sigma_1} + \beta^2\frac{\sigma_2^2}{\sigma_1^2})\sigma_1^2\sigma_2^2}} \\
&= \frac{\rho_{RF} - \beta\frac{\sigma_2}{\sigma_1}}{\sqrt{1 - 2\beta\rho_{RF}\frac{\sigma_2}{\sigma_1} + \beta^2\frac{\sigma_2^2}{\sigma_1^2}}}
\end{aligned}
$$

Proof of Equation 35, the limiting distribution of $\hat{\beta}$ when the true first stage is quadratic

$$
\begin{aligned}
\hat{\beta} - \beta &= \frac{Z'u}{Z'X} \\
&= \frac{\frac{1}{\sqrt{n}}Z'u}{\frac{1}{\sqrt{n}}Z'X} \\
&\xrightarrow{d} \frac{\phi_{zu}}{\phi_{zx}}, \quad \text{By CMT for } \xrightarrow{d} \\
E[\phi_{zx}] &= E[Z_iX_i] \\
&= E[Z_i(Z_i^2\delta + w_i)] \\
&= 0 \\
Var(\phi_{zx}) &= Var(Z_iX_i) \\
&= E[(Z_iX_i)^2] \\
&= E[(Z_i^3\delta + Z_iw_i)^2] \\
&= E[\delta^2Z_i^6 + 2\delta Z_i^4w_i + Z_i^2w_i^2] \\
&= \delta^2E[Z_i^6] + \sigma_w^2 \\
Cov(\phi_{zu}, \phi_{zx}) &= Cov(Z_iu_i, Z_iw_i) \\
&= E[Z_i^2u_iw_i] \\
&= \sigma_{uw} \\
\implies \begin{pmatrix}\phi_{zu}\\ \phi_{zx}\end{pmatrix} &\sim N(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_u^2 & \sigma_{uw}\\ \sigma_{uw} & \delta^2E[Z_i^6] + \sigma_w^2\end{pmatrix}) \\
\hat{\beta} - \beta &\xrightarrow{d} \frac{\phi_{zu}}{\phi_{zx}} \\
&= \frac{E[\phi_{zu}|\phi_{zx}] + \xi}{\phi_{zx}} \\
&= \frac{\sigma_{uw}}{\delta^2E[Z_i^6] + \sigma_w^2} + \frac{\xi}{\phi_{zx}}
\end{aligned}
$$

$$
\begin{aligned}
Var(\phi_{zu}) &= Var(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6]+\sigma_w^2}\phi_{zx}) + Var(\xi) \\
\sigma_u^2 &= \frac{\sigma_{uw}^2}{\delta^2 E[Z_i^6]+\sigma_w^2} + Var(\eta) \\
Var(\eta) &= \sigma_u^2 - \frac{\sigma_{uw}^2}{\delta^2 E[Z_i^6]+\sigma_w^2} \\
\implies \begin{pmatrix} \xi \\ \phi_{zx} \end{pmatrix} &\sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 - \frac{\sigma_{uw}^2}{\delta^2 E[Z_i^6]+\sigma_w^2} & 0 \\ 0 & \delta^2 E[Z_i^6]+\sigma_w^2 \end{pmatrix})
\end{aligned}
$$

Proof of Equation 37, the distribution of Wald statistic when the true first stage is quadratic under $\beta = \beta_0$

$$
\begin{aligned}
t^2 &= \frac{(\hat{\beta}-\beta_0)^2}{\hat{\sigma}_u^2 (Z'Z)(Z'X)^{-2}} \\
&= \frac{(Z'u)^2}{\hat{\sigma}_u^2 (Z'Z)} \\
&= \frac{(\frac{1}{\sqrt{n}}Z'u)^2}{\hat{\sigma}_u^2 \frac{1}{n}Z'Z} \\
\hat{\sigma}_u^2 &= \frac{1}{n}(Y-X\hat{\beta})'(Y-X\hat{\beta}) \\
&= \frac{1}{n}[X(\beta-\hat{\beta})+u]'[X(\beta-\hat{\beta})+u] \\
&= \frac{1}{n}(\frac{Z'u}{Z'X})^2 X'X - \frac{2}{n}\frac{Z'u}{Z'X}X'u + \frac{1}{n}u'u \\
&= \frac{1}{n}(\frac{Z'u}{Z'X})^2 \sum_{i=1}^n (Z_i^2\delta + w_i)^2 - \frac{2}{n}\frac{Z'u}{Z'X}\sum_{i=1}^n (Z_i^2\delta+w_i)u_i + \frac{1}{n}\sum_{i=1}^n u_i^2 \\
&= ① + ② + ③ \\
① &= (\frac{Z'u}{Z'X})^2 \frac{1}{n}\sum_{i=1}^n [Z_i^4\delta^2 + 2\delta Z_i^2 w_i + w_i^2] \\
&\xrightarrow{d} (\frac{\sigma_{uw}}{\delta^2 E[Z_i^6]+\sigma_w^2} + \frac{\xi}{\phi_{zx}})^2 (\delta^2 E[Z_i^4]+\sigma_w^2) \\
② &= -2\frac{Z'u}{Z'X}\frac{1}{n}[\sum_{i=1}^n \delta Z_i^2 u_i + \sum_{i=1}^n w_i u_i] \\
&\xrightarrow{d} -2\sigma_{uw}(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6]+\sigma_w^2} + \frac{\xi}{\phi_{zx}}) \\
③ &= \frac{1}{n}\sum_{i=1}^n u_i^2 \\
&\xrightarrow{p} \sigma_u^2
\end{aligned}
$$

$$\hat{\sigma}_u^2 \xrightarrow{d} (\frac{\sigma_{uw}}{\delta^2 E[Z_i^6] + \sigma_w^2} + \frac{\xi}{\phi_{zx}})^2(\delta^2 E[Z_i^4] + \sigma_w^2) - 2\sigma_{uw}(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6] + \sigma_w^2} + \frac{\xi}{\phi_{zx}}) + \sigma_u^2$$

by CMT for $\xrightarrow{d}$

$$t^2 \xrightarrow{d} \frac{\phi_{zu}^2}{(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6]+\sigma_w^2} + \frac{\xi}{\phi_{zx}})^2(\delta^2 E[Z_i^4] + \sigma_w^2) - 2\sigma_{uw}(\frac{\sigma_{uw}}{\delta^2 E[Z_i^6]+\sigma_w^2} + \frac{\xi}{\phi_{zx}}) + \sigma_u^2}$$

Proof of Consistency and Asymptotic Normality of $\hat{\beta}_{Nonlinear}$ as in Equation 39 and 40

$$\begin{aligned}
\hat{\beta}_{nonlinear} &= (\sum_{i=1}^n \hat{X}_i X_i) \sum_{i=1}^n \hat{X}_i Y_i \\
&= (\sum_{i=1}^n Z_i^2 \hat{\delta} X_i)^{-1} \sum_{i=1}^n Z_i^2 \hat{\delta} Y_i \\
&= (\sum_{i=1}^n Z_i^2 X_i)^{-1} \sum_{i=1}^n Z_i^2 (\beta X_i + u_i) \\
&= \beta + (\sum_{i=1}^n Z_i^2 X_i)^{-1} \sum_{i=1}^n Z_i^2 u_i \\
&= \beta + (\frac{1}{n}\sum_{i=1}^n Z_i^2 X_i)^{-1} \frac{1}{n} \sum_{i=1}^n Z_i^2 u_i \\
&\xrightarrow{p} \beta \\
\sqrt{n}(\hat{\beta}_{nonlinear} - \beta) &= (\frac{1}{n}\sum_{i=1}^n Z_i^2 X_i)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i^2 u_i \\
&\xrightarrow{d} N(0, \sigma_u^2 E[Z_i^4] E[Z_i^2 X_i]^{-2})
\end{aligned}$$

Next I show using the idea of "Forbidden Regression" also works in this setting

$$\begin{aligned}
\hat{\beta}_{Nonlinear} &= (\sum_{i=1}^n \hat{X}_i^2) \sum_{i=1}^n \hat{X}_i Y_i \\
&= (\sum_{i=1}^n Z_i^4 \hat{\delta}^2)^{-1} \sum_{i=1}^n Z_i^2 \hat{\delta}(X_i \beta + u_i) \\
&= \hat{\delta}^{-1} (\sum_{i=1}^n Z_i^4)^{-1} \sum_{i=1}^n Z_i^2 (\beta X_i + u_i) \\
&= (\sum_{i=1}^n Z_i^2 X_i)^{-1} \sum_{i=1}^n (\beta Z_i^2 X_i + Z_i^2 u_i) \\
&= \beta + (\sum_{i=1}^n Z_i^2 X_i)^{-1} \sum_{i=1}^n Z_i u_i
\end{aligned}$$

$$
\begin{aligned}
&= \beta + (\frac{1}{n}\sum_{i=1}^{n} Z_i^2 X_i)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i^2 u_i \\
&\xrightarrow{p} \beta
\end{aligned}
$$

Proof of Equation 43, the formula for TSLS

$$
\begin{aligned}
\beta_{TSLS} &= E[\hat{X}^2]^{-1} E[\hat{X}Y] \\
&= E[E[X|Z]^2]^{-1} E[E[X|Z]Y] \\
&= E[E[X|Z]^2]^{-1} E[E[X|Z](X\beta + e)] \\
&= E[E[X|Z]^2]^{-1} (E[E[X|Z]X\beta + E[X|Z]e] \\
&= E[E[X|Z]^2]^{-1} (\beta E[E[X|Z]X] + E[E[X|Z]e]) \\
&= \beta E[E[X|Z]^2]^{-1} E[E[X|Z]^2] + E[E[X|Z]^2]^{-1} E[E[X|Z]E[e|Z]] \\
&= \beta
\end{aligned}
$$