Assumption Violations in Instrumental Variables Regression: Weak instruments in subvector testing under heteroskedasticity and assessing the exclusion restriction

Anna Esenther

Trinity College

University of Oxford

A dissertation submitted for the MSc by Research in Statistics

September 2022

I declare that all work contained herein is my own.

Anna Esenther

Contents

Acknowledgments

I would like to thank my supervisor, Dr Frank Windmeijer, for his dedication, enthusiasm, and patience as this dissertation came together. I am the grateful beneficiary of many hours of his instruction and advice and his investments are greatly appreciated. I would additionally like to acknowledge the Rhodes Trust for funding my studies and Dr John Eric Humphries for volunteering to provide feedback on a draft.

Abstract

Instrumental variables (IV) is a commonly used regression technique to estimate treatment parameters when the treatment is not random, i.e., it is endogenous. It relies on key assumptions and this dissertation explores strategies to deploy when those assumptions may not hold. Chapter 1 deals with the case of testing a subset of non-random treatment parameters when both the assumptions of strong instrument relevance and homoskedasticity fail. We recommend bootstrapping the Kleibergen-Paap Test or a 2-step LIML-based J-test in this situation. Chapter 2 considers a different assumption of IV regression: the exclusion restriction. We prove that the test devised by Kiviet (2020) does not function as a true test of the exclusion restriction and recommend reformulating it as a test of the endogeneity of the treatment variable, a test that is only valid when the exclusion restriction holds. Thus while the problems considered in Chapter 1 - weak instruments and heteroskedasticity - can be addressed by using the tests we recommend in the subvector case, the exclusion restriction, considered in Chapter 2, is not easy to relax or test. It is a foundational assumption of IV regression without a simple workaround, although past work offers adjustments if a certain type of subsample is available (van Kippersluis and Rietveld, 2018) and our endogeneity test can provide limited information in assessing whether the exclusion restriction holds for a particular instrument, even though a true test of endogeneity remains impossible.

Table 1: Key Results from Chapter 1

Figure Number	Key Result
1.2	The asymptotic AR-based tests are slightly oversized but implementing the bootstrap improves the size properties.
1.3	Under weak instruments, GKM has higher power than either the asymptotic or bootstrapped AR test, but all three tests converge under strong instruments.
1.6	Basmann rejection rates are higher than their Sargan counterparts.
1.9	The bootstapped statistics suffer a modest power loss compared to GKM.
1.11	The asymptotic KP and J2L tests are robust to heteroskedasticity and therefore perform better in the presence of heteroskedasticity than do the non-robust AR-based statistics.
1.12	All of the bootstrapped statistics are approximately appropriately sized, especially compared to GKM, but the bootstrapped KP and J2L are somewhat less undersized under strong instruments than the bootstrapped AR.
1.15	The power of each test improves as the identification of β strengthens.

Chapter 1

Robust Subvector Testing under Weak Instruments and Heteroskedasticity

1.1 Introduction

Instrumental variables (IV) estimation in linear models is a technique widespread among applied researchers in economics and statistics. It addresses the problem of *endogeneity*, in which treatment variables are confounded by unobserved characteristics, through the use of an *instrument*, a variable outside the regression model that is related to the endogenous regressor but is not itself endogenous. Lottery systems make ideal instruments because winning a lottery is related to the regressor of interest but the randomisation ensures lottery winners are not systematically different from lottery losers. Gray-Lobe, Pathak, and Walters (2021), for example, studied the effects of attending public preschool in Boston, a city that uses a randomised tie-breaking element in its system for assigning students

with different preferences to different schools. Controlling for preferences, offers are functionally allocated by lottery and can then be used as an instrument for preschool attendance. This is a good instrument because receiving an offer is strongly related to preschool enrolment, but because offers are awarded randomly, conditional on preferences, receiving an offer is not related to the error term of the regression.

However, many instruments used in practice do not exhibit the same high correlation with the treatment variable as is the case with preschool assignment and enrolment. Weak instruments are only weakly related to the variable of interest, which can lead to biased estimates and imprecise inference when using standard IV procedures. There are several methods for addressing this problem, most famously the Anderson Rubin (AR) test, which is robust to weak instruments, meaning that even if the correlation between the instruments and treatment variable is low, the test will still control size (Anderson and Rubin, 1949).

We are interested here in the case of subvector testing, in which we wish to conduct a hypothesis test involving only a subset of the endogenous treatment variables in the model. The coefficients on the endogenous treatment variables not under test are then considered nuisance parameters. We show that testing a hypothesis about a subset of parameters is equivalent to conducting an overidentification test on the restricted model in which the null hypothesis is imposed. Because Windmeijer (2021) has shown that overidentification tests and underidentification tests can use the same test statistics, we can adapt statistics conceived of originally for use in either underidentification or overidentification tests for our setting.

The standard method in subvector testing is to use some form of the AR statistic. Several modifications to subvector AR testing have previously been proposed

to improve its performance under weak instruments or to make it robust to a particular form of heteroskedasticity (Andrews, 2017; Guggenberger et al., 2021). A major drawback of the subvector AR statistic, however, is that the version of the statistic used in the subvector literature is not robust to heteroskedasticity of a general form. Noting that the subvector AR test is simply a LIML-based overidentification test, we propose using heteroskedasticity-robust LIML-based overidentification tests instead, namely the Kleibergen-Paap (KP) test and the 2-step LIML J-test (J2L), in order to create a fully robust subvector test, which has not previously been done. Additionally, we bootstrap these tests for better performance. We find that bootstrapping KP and J2L improves the size of subvector hypothesis tests compared to any of the asymptotic tests or to bootstrapping the non-heteroskedasticity-robust AR statistic.

The chapter proceeds as follows: Section 1.2 introduces the model and estimation strategy in the full-vector case, Sections 1.3 and 1.4 cover the subvector case under homoskedasticity and heteroskedasticity, respectively, Section 1.5 examines previous approaches to subvector testing, Sections 1.6 and 1.7 describe the bootstrap procedures and present the simulation results for the homoskedastic and heteroskedastic cases, respectively, and Section 1.8 concludes. Throughout, random vectors are represented by lower case letters and random matrices are represented by uppercase letters.

1.2 Instrumental Variables Model and Estimation

1.2.1 Model Setup

We have sample of size n of independent and identically distributed (iid) data $\{y_i, x_i', z_i'\}_{i=1}^n$. The standard IV setup involves the following set of two equations:

$$y = X\beta + u \tag{1.1}$$

$$X = Z\Pi_X + V_X \tag{1.2}$$

Here, y is an $n \times 1$ vector of outcomes, X is an $n \times k_X$ matrix of potentially endogenous regressors - meaning $E[x_i u_i]$ may not equal zero, Z is an $n \times k_Z$ matrix of exogenous instruments with $k_Z \geq k_X$, β is a $k_X \times 1$ vector of coefficients, Π_X is a $k_Z \times k_X$ matrix of coefficients indicating the instrument strength, and $\{u_i, v_{xi}\}$ are iid errors with mean zero. Note that by convention, v_{xi}, x_i,and z_i are all column vectors. Equation 1.1 is referred to as the structural equation and Equation 1.2 is referred to as the first stage. We are interested in the coefficient vector β. We assume the errors are conditionally homoskedastic:

$$\begin{pmatrix} u_i \\ v_{xi} \end{pmatrix} | z_i \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uV}' \\ \sigma_{uV} & \Sigma_{VV} \end{pmatrix} \right)$$

Here the variance conditional on the instruments does not depend on the instruments, an assumption we will later relax when we move to heteroskedasticity.

In order to conduct IV estimation, two conditions must hold:

Assumption 1: $\text{rank}(E[z_i x_i']) = k_X$

Assumption 2: $E[z_i u_i] = 0$

The first assumption is the *relevance condition*, which states that all instruments have some independent correlation with X or, equivalently, the expected value of the matrix $z_i x_i'$ is of full rank, and the second assumption is the *exogeneity condition*, which states that the instruments are not related to the error term. While Assumption 1 requires that the matrix $E[z_i x_i']$ be of full column rank, applied researchers often use weak instruments such that $\text{rank}(E[z_i x_i'])$ is close to being less than k_X. Rank tests evaluate whether this is the case for a particular dataset by testing the null that $\text{rank}(\Pi_X) = k_X - 1$ in Equation 1.2 against the alternative that $\text{rank}(\Pi_X) = k_X$. If the instruments are indeed weak, the researcher should choose methods robust to weak instruments.

Assumption 2, the exogeneity condition, is necessary because endogeneity in the structural equation is the problem the instrumental variables method seeks to solve. If $E[x_i u_i] \neq 0$, then X is endogenous and OLS estimates are inconsistent:

$$\begin{aligned}
\hat{\beta}_{OLS} =& (X'X)^{-1}X'y \\
=& (X'X)^{-1}X'(X\beta + u) \\
=& \beta + (X'X)^{-1}X'u
\end{aligned}$$

$$\begin{aligned}
\text{plim}[\hat{\beta}_{OLS}] =& \beta + \text{plim}[(X'X/n)^{-1}(X'u/n)] \\
=& \beta + E[x_i x_i']^{-1}E[x_i u_i] \\
\neq& \beta
\end{aligned}$$

Thus, the correlation between X and u causes $\hat{\beta}_{OLS}$ to suffer from inconsistency and therefore also from biasedness. The method of instrumental variables regression is meant to account for the endogeneity of X and so the instruments chosen

for this purpose must themselves not be endogenous or else they are invalid.

Note that the model may also include exogenous control variables, but here these are taken to be partialled out, without loss of generality.

1.2.2 Strong and Weak Instrument Asymptotics

In Equation 1.2, Π_X is modelled as fixed, which indicates we are operating under strong instrument asymptotics. As the sample size grows, standard errors shrink and the F statistic for $H_0 : \Pi_X = 0$ goes to infinity if $\Pi_X \neq 0$. However, we are primarily interested in the case of weak instruments, for which fixed Π_X is not a good model because even with a large sample size, the low correlation between X and Z means that we need more information to confidently reject the null that $\Pi_X = 0$. Staiger and Stock (1997) therefore recommend modelling Π_X as local to zero with the assumption

$$\Pi_X = \Pi_{Xn} = C/\sqrt{n}, \tag{1.3}$$

where C is a constant $k_Z \times k_X$ matrix. This is the setting of weak instrument asymptotics.

Stock, Wright, and Yogo (2002) consider the case where $k_X = 1$ and define the concentration parameter, a measure of the strength of the instruments, as

$$\mu^2 = \pi_x' Z' Z \pi_x / \sigma_v^2.$$

The F-statistic is then an estimator for $\mu^2/k_Z + 1$. If we allow k_X to be greater than one, then we have a matrix V rather than a vector v to represent the error in the first stage equation and the concentration parameter becomes

$$\mu^2 = \Sigma_{VV}^{-1/2\prime} \Pi_X' Z' Z \Pi_X \Sigma_{VV}^{-1/2}.$$

The minimum eigenvalue of the matrix μ^2 diverges when the instruments are strong with fixed Π_X in large samples, but remains finite under weak instrument asymptotics.

Rather than testing against the alternative that the rank of Π_X is local to zero, it is recommended to test for a near rank reduction of one - that is, to test that the rank of Π_X is in the $1/\sqrt{n}$ neighbourhood of $k_X - 1$. Sanderson and Windmeijer (2016) consider the case of two endogenous variables and define the weak instrument asymptotics as

$$\pi_1 = \delta\pi_2 + c/\sqrt{n},$$

where π_1 and π_2 are the coefficients in the first stage regressions of each of the two endogenous variables.

With weak instruments, the standard null limiting distributions for β_{2SLS} do not apply, which can cause problems for testing. The tests of $H_0 : \beta = \beta_0$ in subvector settings that we consider in this paper, described below, are robust to weak instruments, which means they retain the correct asymptotic distributions even when Π_X is local to zero and so maintain correct asymptotic size regardless of instrument strength.

The standard rule of thumb is to consider instruments weak if the first-stage F-statistic is less than ten (Staiger and Stock, 1997). To examine the salience of the weak instrument problem in empirical research, Andrews, Stock, and Sun (2019) surveyed a sample of 17 papers that used an instrumental variables approach and were published in the American Economic Review from 2014 to 2018. They found that weak instruments are frequently used in practice, with the F-statistics of many specifications less than or close to 10, suggesting that weak instrument-

robust inference is an important topic with implications for empirical research.

Another indication of the need for appropriate weak instrument methodology is that the common practice of screening instruments based on first-stage F-statistics has been shown to lead to undesirable outcomes. Dropping specifications with F-statistics under 10 can lead to size distortions (Andrews et al., 2019) and so simply avoiding weak instrument inference is not a satisfactory course of action.

1.2.3 Two Stage Least Squares Estimation

If $k_Z < k_X$, the model is unidentified and estimation is not feasible. If $k_Z = k_X$, the model is just identified and we can set $Z'\hat{u} = Z'(y - X\hat{\beta}) = 0$ in the sample, as a sample equivalent of the population condition $E[z_i u_i] = 0$. The estimator that solves this equation is

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

because

$$\begin{aligned} Z'(y - X\hat{\beta}_{IV}) &= Z'(y - X(Z'X)^{-1}Z'y) \\ &= Z'y - Z'X(Z'X)^{-1}Z'y) \\ &= Z'y - Z'y \\ &= 0. \end{aligned}$$

However, when $k_Z > k_X$, the overidentified case, it is not possible to set all sample moment conditions equal to zero. Under conditional homoskedasticity of u, the efficient estimator is the two-stage least squares (2SLS) estimator:

$$\hat{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y,$$

which is equivalent to the result from first obtaining $\hat{\Pi}_X$ by regressing X on Z in Equation 1.2, then calculating $\hat{X} = Z\hat{\Pi}_X$, and finally substituting X for $\hat{X}$ in Equation 1.1 and regressing y on $\hat{X}$.

Under strong instrument asymptotics - with fixed Π_X - and Assumptions 1 and 2, the limiting distribution of the 2SLS estimator is

$$\sqrt{n}\left(\hat{\beta}_{2SLS} - \beta\right) \xrightarrow{d} N\left(0, \sigma_u^2 Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX}\right),$$

where σ_u^2 is the variance of the structural error and $Q_{AB} = E[a_i b_i']$. Note that this method relies upon the relevance assumption and yields biased estimates if the instruments are weak.

1.2.4 Limited Information Maximum Likelihood Estimation

An alternative to the 2SLS estimator is the Limited Information Maximum Likelihood (LIML) estimator. In the just identified case, LIML and 2SLS are equivalent: $\hat{\beta}_L = \hat{\beta}_{2SLS} = \hat{\beta}_{IV}$. In the overidentified case, they are asymptotically equivalent under strong instrument asymptotics. However, when there are many weak instruments, 2SLS is biased but LIML is asymptotically unbiased (Davies et al., 2014). Despite this advantage, 2SLS is more commonly used in practice.

We define P_Z as the projection matrix $P_Z = Z(Z'Z)^{-1}Z'$ and $M_Z = I_n - P_Z$. Let $R = (y, X)$, $\hat{\Omega} = R'M_Z R/(n - k_Z - k_X)$, and $b = \begin{pmatrix} 1 & -\beta' \end{pmatrix}'$. Then the LIML estimator of β for the system defined by Equations 1.1 and 1.2 is

$$\begin{aligned}
\hat{\beta}_L &= \arg\min_{\beta} \frac{(y - X\beta)' P_Z (y - X\beta)}{(y - X\beta)' M_Z (y - X\beta)/(n - k_Z - k_X)} \\
&= \arg\min_{b} \frac{b' R' P_Z R b}{b' R' M_Z R b/(n - k_Z - k_X)} \\
&= \arg\min_{b} \frac{b' R' P_Z R b}{b' \hat{\Omega} b} \\
&= \arg\min_{b} \frac{b' R' P_Z R b}{b' R' R b/(n - k_X)}.
\end{aligned}$$

The first equivalence holds by simply rewriting in terms of b, the second equivalence holds by substituting in $\hat{\Omega} = R' M_Z R/(n - k_Z - k_X)$, and the third equivalence changes the variance estimator, which also changes the degrees of freedom. The choice of variance matrix does not affect the minimisation problem and therefore also does not affect the LIML estimate of β, but the choice will affect the value of the overidentification test statistic (see below).

To solve the minimisation problem, one can first find the minimum eigenvalue of the matrix

$$\hat{\Omega}^{-1}(R' P_Z R).$$

The eigenvector associated with this smallest eigenvalue is $\hat{b} = (\hat{b}_1 \ \hat{b}_2')'$ and the LIML estimator is $\hat{\beta}_L = -\hat{b}_2/\hat{b}_1$. Next calculate $\hat{u}_L = y - X\hat{\beta}_L$. Then $\hat{\Pi}_{X_L} = (Z' M_{\hat{u}_L} Z)^{-1}(Z' M_{\hat{u}_L} X)$.

1.2.5 Testing for Overidentification

This paper employs tests for overidentification, which are used to test Assumption 2, the exogeneity condition, in the situation where there are more instruments than regressors. Valid instruments are exogenous; their only effect on y is through X and so they have no independent contribution to y once X has been controlled

for. Overidentification tests are therefore equivalent to testing the null hypothesis that $\gamma = 0$ in the equation $y = X\beta + Z_o\gamma + u$, where Z_o are a set of $k_Z - k_X$ overidentifying instruments. A rejection in an overidentification test indicates that at least one instrument is endogenous.

There are two main forms of the general test for overidentifying restrictions. The first is the Sargan version (Sargan, 1958):

$$S(\hat{\beta}_{2SLS}) = \frac{\hat{u}'_{2SLS} P_Z \hat{u}_{2SLS}}{\hat{u}'_{2SLS}\hat{u}_{2SLS}/(n - k_X)},$$

where $\hat{u}_{2SLS} = y - X\hat{\beta}_{2SLS}$, and the second is the Basmann version (Basmann, 1960):

$$B(\hat{\beta}_{2SLS}) = \frac{\hat{u}'_{2SLS} P_Z \hat{u}_{2SLS}}{\hat{u}'_{2SLS} M_Z \hat{u}_{2SLS}/(n - k_Z - k_X)}.$$

Asymptotically, both $S(\hat{\beta}_{2SLS})$ and $B(\hat{\beta}_{2SLS})$ have a $\chi^2_{k_Z - k_X}$distribution under valid moment conditions. The key difference is that Basmann-type variances subtract out the part of the error that is explained by the instruments, which improves the power.

When conducting tests of overidentifying restrictions, we can also use the LIML estimates in place of the 2SLS estimates, which produces the following Sargan and Basmann test statistics:

$$S(\hat{\beta}_L) = \frac{\hat{u}'_L P_Z \hat{u}_L}{\hat{u}'_L \hat{u}_L/(n - k_X)}$$

$$B(\hat{\beta}_L) = \frac{\hat{u}'_L P_Z \hat{u}_L}{\hat{u}'_L M_Z \hat{u}_L/(n - k_Z - k_X)}$$

Note that these are the same as the matrices described above for which the min-

imum eigenvalue corresponds to an eigenvector that can be used to calculate the LIML estimate.

1.2.6 Anderson-Rubin Test

The Anderson-Rubin (AR) statistic is used for hypothesis testing under weak instrument asymptotics (Anderson and Rubin, 1949). When compared to a critical value, it tests the hypothesis $H_0 : \beta = \beta_0$ in the system defined by Equations 1.1 and 1.2 and does not depend on the value of Π_X, making it robust to arbitrarily weak instruments (Stock, Wright, and Yogo, 2002). The AR statistic is also commonly used as an overidentification test in the context of subvector testing (see Section 1.3).

A Sargan form of the statistic in the full vector case, where $u_0 = y - X\beta_0$, is

$$ARS(\beta_0) = \frac{u_0' P_Z u_0}{u_0' u_0 / n}$$

and a Basmann form is

$$AR(\beta_0) = \frac{u_0' P_Z u_0}{u_0' M_Z u_0 / (n - k_Z)},$$

both of which converge to a $\chi^2_{k_Z}$ distribution under the null. Note that $\hat{\beta}_L$ minimizes $ARS(\beta)$ and $AR(\beta)$. To test the null $H_0 : \beta = \beta_0$, we compare $ARS(\beta_0)$ or $AR(\beta_0)$ to the appropriate $\chi^2_{k_Z}$ critical value. The full-vector AR test does not involve any estimation and even under weak identification, its size is controlled. In addition to its robustness to weak instruments, the AR test can also be made robust to heteroskedasticity in the following way:

$$ARS_r(\beta_0) = u_0' Z (Z' H_{u_0} Z)^{-1} Z' u_0$$

and

$$AR_r(\beta_0) = u_0' Z (Z' H_{e_0} Z)^{-1} Z' u_0,$$

where $e_0 = M_Z u_0$ and $H_a = \text{diag}(a^2)$, so $Z' H_{u_0} Z = \sum_{i=1}^n u_{0i}^2 z_i z_i'$, and $Z' H_{e_0} Z = \sum_{i=1}^n e_{0i}^2 z_i z_i'$. $ARS_r(\beta_0)$ and $AR_r(\beta_0)$ are robust to heteroskedasticity because the terms $Z' H_{u_0} Z$ and $Z' H_{e_0} Z$ allow for different values of the error term to be associated with each z_i, whereas the non-robust $ARS(\beta_0)$ and $AR(\beta_0)$ assume that u_0 is constant across Z. In practice, the homoskedasticity-dependent AR statistic is typically the version used.

1.2.7 Underidentification and Overidentification

Subvector hypothesis tests are overidentification tests (see Section 1.3), so we are interested in considering overidentification test statistics, including AR. However, statistics originally intended for underidentification tests can also act as overidentification statistics when applied in a different context.

Windmeijer (2021) shows that an underidentification test in the model $y = X\beta + u$ is equivalent to an overidentification test in the auxiliary model $x_1 = X_2\delta + \varepsilon$ using the same instruments Z as the main model, where $X = [x_1 X_2]$. The first-stage regression $X = Z\Pi_X + V_X$ can then be partitioned into two parts:

$$x_1 = Z\pi_1 + v_1$$
$$X_2 = Z\Pi_2 + V_2,$$

where $\Pi_X = [\pi_1 \ \Pi_2]$ and $V_X = [v_1 \ V_2]$. We are simply separating out one of the variables in X: x_1 includes one variable and X_2 includes the other $k_X - 1$

variables in X. We continue to use the full Z matrix and π_1, Π_2, v_1, and V_2 simply represent the appropriate portions of the first-stage parameters and errors after partitioning.Note that $\text{rank}(\Pi_X) = \text{rank}((E(z_i z_i'))^{-1}E(z_i x_i')) = \text{rank}(E(z_i x_i'))$, which equals k_x if the full rank condition, Assumption 1, holds. The primary model is then underidentified if $\text{rank}(\Pi_X) < k_X$ and so an underidentification test tests $H_0 : \text{rank}(\Pi_X) = k_X - 1$. If this null is true, following Windmeijer (2021), there exists a δ^* such that $\Pi\delta^* = 0$, which is simply an algebraic representation of the null that Π is rank deficient. After partitioning, we can write $\Pi\delta^*$ as $[\pi_1 \ \Pi_2] \begin{bmatrix} \delta_1^* \\ \delta_2^* \end{bmatrix}$, so $\pi_1\delta_1^* + \Pi_2\delta_2^* = 0$ and $\pi_1 = \Pi_2(-\delta_2^*/\delta_1^*) = \Pi_2\delta$. Plugging $\pi_1 = \Pi_2\delta$ into the first-stage equation for x_1 yields

$$\begin{aligned} x_1 &= Z\Pi_2\delta + v_1 \\ &= X_2\delta + v_1 - V_2\delta \\ &= X_2\delta + \varepsilon. \end{aligned}$$

Under the null of reduced rank, $E(z_i\varepsilon_i) = E(z_i(v_{1i} - v_{2i}\delta)) = 0$, which is clearly an exogeneity condition of the same form as Assumption 2. Tests of the exogeneity condition are tests for overidentification. Hence, testing the null of reduced rank in the primary model is equivalent to testing the exogeneity condition in the auxiliary model and so the same test statistics can be used for both overidentification and underidentification tests, depending on the context.

If the model is not underidentified and $\text{rank}(\Pi_X) = k_X$, then we should reject $H_0 : \text{rank}(\Pi_X) = k_X - 1$ at a level greater than or equal to α. The higher the rejection rate in this case, the higher the power. Alternatively, if $\text{rank}(\Pi_X) = k_X - 1$, then the null is true and we should reject at a level equal to α if the test is ap-

propriately sized. Notably, if rank(Π_X) is local to zero, as in the weak instrument asymptotics given in Equation 1.3, then we will reject at a level less than or equal to α (Gospodinov et al., 2017). Thus, rank tests under weak instruments are conservative.

Similarly, an overidentification test - that is, a test of $H_0 : E(z_i u_i) = 0$ - should reject at level equal to α when the model is fully identified and rank(Π_X) = k_X, then reject at progressively lower levels than α as the rank of Π_X decreases. This is shown in the simulation in Figure 1.1. A Basmann overidentification test rejects at a level close to α when Π_X is of full rank and the instruments are strong. In this case, $k_X = 3$ so rank(Π_X) = 3 in the fully identified case and as the instruments grow in strength (c increases), the rejection rate approaches $\alpha = 0.05$ from below. When rank(Π_X) = 2, that is, with a rank reduction of one, the rejection rate is severely diminished even under very strong instruments. And in the case of a rank reduction of two, when rank(Π_X) = 1, the rejection rate is functionally zero. As expected, rank reductions thus lead to more conservative tests.

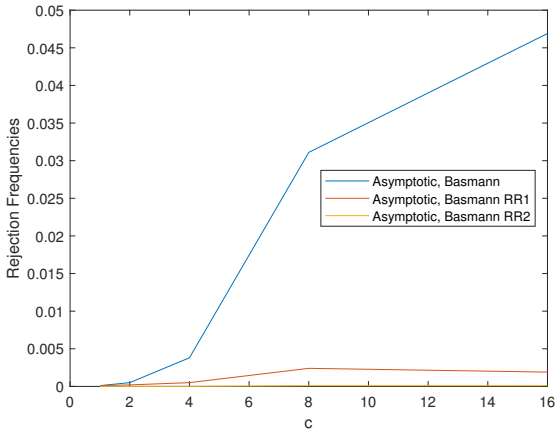
Additionally, we can observe the results of weak instruments, that is, when the instruments have low but nonzero correlations with X. In the simulation, the matrix Π_X is generated as

$$\Pi_X = (c/\sqrt{k_Z n} \begin{pmatrix} -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix})'$$

when it is of full rank or

$$\Pi_X = (c/\sqrt{k_Z n} \begin{pmatrix} -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix})'$$

Figure 1.1: Size of Asymptotic AR Test with Rank Reductions of 0, 1, and 2 of Π_X for $k_X = 3$



Note: Rejection frequencies decrease both as the rank decreases (blue line to red to yellow) and as the instruments weaken (blue line at c=16 compared to c=2).

for a rank reduction of one or

$$\Pi_X = (c/\sqrt{k_Z n} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix})'$$

for a rank reduction of two, where c is a constant that indicates the strength of identification. A higher value of c indicates stronger instruments. Figure 1.1 shows that when c is low and therefore the instruments are weak, the test is very conservative, even when Π_X is of full rank. Although the relevance condition is technically satisfied under weak but nonzero identification, we can think of weak instruments as reducing the rank of Π_X.

Recall the simple model given in Equations 1.1 and 1.2:

$$y = X\beta + u$$

$$X = Z\Pi_X + V_X$$

We can write y as

$$y = Z\pi_y + v_y.$$

Imposing the restriction $\pi_y = \Pi_X\beta$,

$$\begin{aligned} y &= Z\Pi_X\beta + v_y \\ &= X\beta + v_y - V_X\beta. \end{aligned}$$

As above, if the restriction $\pi_y = \Pi_X\beta$ is true, $E(z_i(v_{yi} - v_{Xi}\beta)) = 0$. If the restriction is not true and instead $\pi_y = \Pi_X\beta + \xi$, then $y = Z\Pi_X\beta + v_y + Z\xi = X\beta + v_y + Z\xi - V_X\beta$ and $E(z_i(v_{yi} + z_i\xi_i - v_{xi})) = E(z_i(z_i'\xi_i)) \neq 0$. Hence an underidentification test on the rank of (π_y, Π_X) is equivalent to an overidentification test on $E(z_i(y_i - x_i\beta))$.

The maximum rank of (π_y, Π_X) is $k_X + 1$ so we test the null $\text{rank}(\pi_y, \Pi_X) = k_X$. If the restriction $\pi_y = \Pi_X\beta$ is not true and Π_X is fully identified, then $\text{rank}(\pi_y, \Pi_X) = k_X + 1$ and we reject the null $H_0 : \text{rank}(\pi_y, \Pi_X) = k_X$ at a level greater than or equal to α, which reveals the power of the test. If the restriction $\pi_y = \Pi_X\beta$ is not true but Π_X has a rank reduction of one, then $\text{rank}(\pi_y, \Pi_X) = k_X$ and we should reject the null at a level equal to α, the size of the test. However, if the restriction $\pi_y = \Pi_X\beta$ is true and Π_X is fully identified, then we are also in a situation where $\text{rank}(\pi_y, \Pi_X) = k_X$ and we should reject the null at a level equal to α. We therefore cannot distinguish between the case where π_y is in

the column space of Π_X and the case where Π_X is underidentified simply by observing a rejection rate of α. In the case of weak instruments, $\text{rank}(\Pi_X)$ is close to being less than k_X. Thus if the instruments are also exogenous, as assumed, then $\pi_y = \Pi_X \beta$ and $\text{rank}(\pi_y, \Pi_X)$ is close to being less than k_X, so the test rejects at a rate less than α. This is the same key result as above in different notation: rank tests under weak instruments are conservative.

1.2.8 2-Step Generalized Method of Moments Estimation

We now relax the conditional homoskedasticity assumption and consider methods robust to heteroskedasticity. Heteroskedasticity occurs when the variance of the outcome is dependent on a treatment variable. For example, there is heteroskedasticity in the regression of wages on education because there is low variance in wages among workers with low education levels - most tend to have low wages - but there is a higher variance as wages rise. Some highly educated workers have very high salaries, while others have more modest salaries. Hence, the variance is not constant across X.

In the overidentified case, when $k_Z > k_X$, it is not possible to set all sample moment conditions equal to zero, but we can minimise a distance function, which is the idea behind the Generalised Methods of Moments (GMM) estimator (Hansen, 1982):

$$\begin{aligned}\hat{\beta}_{1GMM} &= \arg\min (y - X\beta)' Z W_n^{-1} Z' (y - X\beta) \\ &= (X' Z W_n^{-1} Z' X)^{-1} X' Z W_n^{-1} Z' y\end{aligned}$$

for some weight matrix $W_n \xrightarrow{p} W$ as $n \to \infty$. The notation $\hat{\beta}_{1GMM}$ indicates that minimising the distance function is the first step of the two step GMM estimation

process. Choosing $W_n = Z'Z$ yields the two stage least squares (2SLS) estimator: $\hat{\beta}_{1GMM} = \hat{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$. Then $\hat{u}_1 = y - X\hat{\beta}_{1GMM}$ and, in the second step,

$$\hat{\beta}_{2GMM} = (X'Z(Z'H_{\hat{u}_1}Z)^{-1}Z'X)^{-1}X'Z(Z'H_{\hat{u}_1}Z)^{-1}Z'y$$

where, for the conditional heteroskedasticity setup, $H_{\hat{u}_1} = \text{diag}(\widehat{u}_1^2)$ and so $Z'H_{\hat{u}_1}Z = \sum_{i=1}^{n} \widehat{u}_{1i}^2 z_i z_i'$. Note that this is a robust estimation of $Var(z_i u_i)$ because the term $\widehat{u}_i^2$ is included within the sum and therefore multiplied by $z_i z_i'$ for each individual in the sample, which allows the conditional error variance to vary across Z: $Var\left(u_i|z_i\right) = \sigma^2\left(z_i\right)$. In contrast, the variance $\sigma^2 \sum_{i=1}^{n} z_i z_i'$ assumes a constant value for the conditional variance, $Var\left(u_i|z_i\right) = \sigma_u^2$, and therefore represents the homoskedastic scenario. Alternative variance matrices can be used to allow for general clustering and serial correlation.

The asymptotic distribution of the 2-step GMM estimator is

$$\sqrt{n}\left(\hat{\beta}_{2GMM} - \beta\right) \xrightarrow{d} N\left(0, (Q_{ZX}'\Theta^{-1}Q_{ZX})^{-1}\right),$$

where

$$\Theta = E\left(\sigma^2\left(z_i\right) z_i z_i'\right)$$

and $Q_{ZX} = E[z_i x_i']$.

As previously mentioned, 2SLS is efficient under homoskedasticity, and under heteroskedasticity of an unknown form, 2-step GMM is efficient.

1.2.9 Heteroskedasticity-Robust Test Statistics for Overidentification

1.2.9.1 Kleibergen-Paap Test

If we use LIML estimation rather than 2SLS in the context of overidentification tests, we arrive at the Kleibergen-Paap (KP) test (Kleibergen and Paap, 2006):

$$KPS(\hat{\beta}_L, \hat{\Pi}_{X_L}) = \hat{u}_L' M_{\hat{X}_L} Z_o (Z_o' M_{\hat{X}_L} H_{\hat{u}_L} M_{\hat{X}_L} Z_o)^{-1} Z_o' M_{\hat{X}_L} \hat{u}_L, \tag{1.4}$$

where $\hat{u}_L = y - X\hat{\beta}_L$, $\hat{X}_L = Z\hat{\Pi}_{X_L}$, and Z_o are a set of $k_Z - k_X$ overidentifying instruments. As with GMM estimation, $H_{\hat{u}_L} = \text{diag}(\widehat{u}_L^2)$. Taking $\hat{e}_L = M_Z \hat{u}_L$, the Basmann form of the KP statistic is

$$KP(\hat{\beta}_L, \hat{\Pi}_{X_L}) = \hat{u}_L' M_{\hat{X}_L} Z_o (Z_o' M_{\hat{X}_L} H_{\hat{e}_L} M_{\hat{X}_L} Z_o)^{-1} Z_o' M_{\hat{X}_L} \hat{u}_L. \tag{1.5}$$

Note that the name of the statistic on its own, in this case KP (Equation 1.5), represents the Basmann form (using first stage residuals) and an S is appended to indicate the Sargan form (e.g., in Equation 1.4, using structural residuals). We will continue with that notation throughout the paper.

The KP test was originally conceived of as a rank test, for example to test the null that $\text{rank}(\Pi_X) = k_X - 1$ in Equation 1.2 against the alternative that $\text{rank}(\Pi_X) = k_X$. Hence, it is commonly used as an underidentification test because it tests the null that the matrix of coefficients representing the correlation between the instruments and regressors is of less than full rank. If this is true, the model is underidentified. However, as shown above, underidentifican tests like KP can also be used as overidentification tests, which is the context we are interested in here.

1.2.9.2 Hansen's J -Test

The J-test for overidentification based on 2-step GMM estimator is as follows (Hansen, 1982):

$$JS(\hat{\beta}_{2SLS}, \hat{\beta}_{2GMM}) = \hat{u}_2' Z (Z' H_{\hat{u}_1} Z)^{-1} Z' \hat{u}_2.$$

Here $\hat{u}_1$ represents the errors from the 2SLS regression, which acts as the first step of the 2-step GMM estimator, and $\hat{u}_2$ represent the errors from the second step. Recall that $Z' H_{\hat{u}_1} Z$ is the heteroskedasticity-robust variance estimator

$$Z' H_{\hat{u}_1} Z = \sum_{i=1}^{n} \widehat{u}_{1i}^2 z_i z_i'$$

and that the S in JS indicates that it is the Sargan form of the statistic. To calculate the Basmann form, we use $\hat{e}_1 = M_Z \hat{u}_1$ in the equation for the second-step estimator:

$$\hat{\beta}_{2GMM} = (X' Z (Z' H_{\hat{e}_1} Z)^{-1} Z' X)^{-1} X' Z (Z' H_{\hat{e}_1} Z)^{-1} Z' y.$$

Then

$$J(\hat{\beta}_{2SLS}, \hat{\beta}_{2GMM}) = \hat{u}_2' Z (Z' H_{\hat{e}_1} Z)^{-1} Z' \hat{u}_2'.$$

Note that the $\hat{u}_2$ in Basmann form are not equivalent to the $\hat{u}_2$ in the Sargan form because the different variance matrices produce different second step estimates, which in turn produce different errors.

Windmeijer (2021) shows that the J-test is equivalent to a robust score test based on a one-step GMM estimator such as 2SLS. He also proposes an adaptation of the J-test in order to create an invariant LIML-based overidentification statistic.

The adapted version, J2L, involves two steps. First, find the LIML estimates $\hat{\beta}_L$ and $\hat{\Pi}_{X_L}$ and calculate the error $\hat{u}_L = y - X\hat{\beta}_L$. The second step is to produce a new estimate of β:

$$\hat{\beta}_{2L} = (\hat{\Pi}'_{X_L} Z'Z(Z'H_{\hat{u}_L}Z)^{-1}Z'X)^{-1}\hat{\Pi}'_{X_L} Z'Z(Z'H_{\hat{u}_L}Z)^{-1}Z'y.$$

Then the second step error $\hat{u}_{2L} = y - X\hat{\beta}_{2L}$ is calculated and the J2L statistic is

$$J2LS(\hat{\beta}_L, \hat{\Pi}_{X_L}) = \hat{u}'_{2L} Z(Z'H_{\hat{u}_{2L}}Z)^{-1}Z'\hat{u}_{2L}.$$

Alternatively, we can use $\hat{e}_L = M_Z\hat{u}_L$ and $\hat{e}_{2L} = M_Z\hat{u}_{2L}$ to calculate the Basmann version:

$$J2L(\hat{\beta}_L, \hat{\Pi}_{X_L}) = \hat{u}'_{2L} Z(Z'H_{\hat{e}_{2L}}Z)^{-1}Z'\hat{u}_{2L}.$$

Note that KP and J2L rely on the same information, namely $\hat{\beta}_L$ and $\hat{\Pi}_{X_L}$.

1.3 Subvector Testing Under Homoskedasticity

If there are endogenous variables we are not interested in testing, they cannot be partialled out and instead enter the model as nuisance parameters. The subvector setup is therefore as follows:

$$y = X\beta + W\gamma + u \tag{1.6}$$

$$X = Z\Pi_X + V_X \tag{1.7}$$

$$W = Z\Pi_W + V_W \tag{1.8}$$

W is an $n \times k_W$ matrix of endogenous regressors that are not of interest for inference and so γ is a vector of nuisance parameters. If the null hypothesis $\beta = \beta_0$ is true, we have $y_0 = y - X\beta_0 = W\gamma + u$ and $E(Z_i u_i) = 0$. If the null is not true, $y_0 = W\gamma + u + X(\beta - \beta_0)$ and $E(Z_i(u_i + X_i(\beta - \beta_0)))$ does not equal zero by Assumption 1, the relevance condition. Thus, we need to perform an overidentification test.

In the subvector setting, we have two parameters describing the strength of the instruments: Π_X and Π_W. If either one is local to zero, i.e. the instruments are weakly related to either X or W, then the power of the test is reduced. If the nuisance parameters in particular are weakly identified and so $\text{rank}(\Pi_W) < k_W$, then the size of the test is also affected and the overidentification tests will reject less frequently than they should under the null.

1.3.1 Subvector Anderson-Rubin Test

When there are nuisance parameters to estimate, the size of the AR test is not controlled.

In the subvector setup defined by Equations 1.6, 1.7, and 1.8, we can calculate the LIML estimate of the nuisance parameter γ in the restricted version of the model, imposing $H_0 : \beta = \beta_0$ and defining $y_0 = y - X\beta_0$, and we can then plug the estimate into the subvector version of the AR statistic (see Guggenberger et al. 2012):

$$ARS(\beta_0, \hat{\gamma}_L) = \frac{(y_0 - W\hat{\gamma}_L)' P_Z (y_0 - W\hat{\gamma}_L)}{(y_0 - W\hat{\gamma}_L)'(y_0 - W\hat{\gamma}_L)/(n - k_W)}$$

or

$$AR(\beta_0, \hat{\gamma}_L) = \frac{(y_0 - W\hat{\gamma}_L)' P_Z (y_0 - W\hat{\gamma}_L)}{(y_0 - W\hat{\gamma}_L) M_Z (y_0 - W\hat{\gamma}_L)/(n - k_Z - k_W)}.$$

When there are nuisance parameters to estimate, the AR statistics asymptotically have a $\chi^2_{k_Z - k_W}$ distribution under strong identification but the size of the AR test is not controlled under weak identification. $ARS(\beta_0, \hat{\gamma}_L)$ and $AR(\beta_0, \hat{\gamma}_L)$ are also not robust to heteroskedasticity. To deal with the size problem under weak identification, Guggenberger, Kleibergen, and Mavroeidis (2019) propose data-dependent conditioning (see below). To deal with the presence of heteroskedasticity, we consider alternative statistics that are heteroskedasticity-robust (see Section 1.4).

1.3.2 Data-Dependent Adjustments

Guggenberger, Kleibergen, Mavroeidis, and Chen (2012) propose a plug-in method by which they estimate the nuisance parameters γ by LIML and then calculate the AR statistic, as above. This method has correct asymptotic size, but low power and is only valid under homoskedasticity. Guggenberger, Kleibergen, and Mavroeidis (2019) (GKM19) improve upon this method with the addition of critical values dependent on the strength of identification. When instruments are weak, hypothesis tests tend to underreject because the low correlation between endogenous treatment variables and instruments leads to situations in which the available evidence is too often too weak to reject the null hypothesis, a problem that does not occur under strong instruments. The method of GKM19 adjusts for this by using smaller critical values when the instruments are weaker, in order to raise the rejection rates to an appropriate level. This results in a test of higher power, but that is still not robust to heteroskedasticity. To address that

problem, Guggenberger, Kleibergen, and Mavroeidis (2021) (GKM21) introduce a procedure that is robust to conditional heteroskedasticity when the covariance matrix has an approximate Kronecker product structure. As a test for this structure rejected its presence in about half of the 118 specifications from highly cited IV papers (Guggenberger, Kleibergen, and Mavroeidis, 2020), this method for addressing heteroskedasticity is not universally valid. Given this fact, GKM21 recommend using it only when the test for approximate Kronecker product structure fails to reject and, where the test rejects the necessary covariance structure, applying the two-step method of Andrews (2017) instead (see Section 1.5).

The AR statistic can be calculated as the minimum eigenvalue of the matrix

$$(R'M_ZR)^{-1}(R'P_ZR),$$

where $R = (y_0, W)$, making it an overidentification test, and the particular data-dependent adjustment made by GKM19 (referred to in the remainder of the paper as GKM) is to condition on the largest eigenvalue of this matrix. In contrast, Grohmann and Windmeijer (2021) propose conditioning on the second smallest eigenvalue instead, a method which hereafter will be referred to as GKM2. In each case, the adjusted critical value is calculated with the help of the tables of values provided by GKM19 to aid practical application. There are different tables for different values of $k_Z - k_W$ so in our simulations, which have six instruments, we use the $k_Z - k_W = 5$ table for one nuisance parameter and the $k_Z - k_W = 4$ table for two nuisance parameters. Then the critical values used in our simulations are calculated through linear interpolation.

1.4 Subvector Testing Under Heteroskedasticity

1.4.1 Kleibergen-Paap Test

The KP statistic fills the need for a heteroskedasticity-robust subvector overidentification test. It can be written either as

$$KPS(\beta_0, \hat{\gamma}_L, \hat{\Pi}_{W_L}) = \hat{u}'_{0L} M_{\hat{W}_L} Z_o (Z'_o M_{\hat{W}_L} H_{\hat{u}_{0L}} M_{\hat{W}_L} Z)^{-1}_o Z'_o M_{\hat{W}_L} \hat{u}_{0L}$$

or

$$KP(\beta_0, \hat{\gamma}_L, \hat{\Pi}_{W_L}) = \hat{u}'_{0L} M_{\hat{W}_L} Z_o (Z'_o M_{\hat{W}_L} H_{\hat{e}_{0L}} M_{\hat{W}_L} Z)^{-1}_o Z'_o M_{\hat{W}_L} \hat{u}_{0L},$$

where $\hat{u}_{0L} = y_0 - W\hat{\gamma}_L$, $\hat{W}_L = Z\hat{\Pi}_{W_L}$, $\hat{e}_{0L} = M_Z \hat{u}_{0L}$, and Z_o is any subset of size $k_Z - k_W$ of the instruments Z. The subvector KP statistic has the advantage over both the standard subset AR statistic and the data-dependent GKM of being robust to heteroskedasticity of an unknown form. Even GKM21 relies on a particular form of heteroskedasticity - i.e. with an approximate Kronecker product covariance matrix - and when that form is not valid, reverts to a non-heteroskedasticity-robust method. Thus the KP test provides a method to conduct subvector testing without any pretesting that is always robust to conditional heteroskedasticity.

1.4.2 2-Step LIML-Based J-Test

Similarly, J2L also fills the role of a heteroskedasticity-robust subvector overidentification test and therefore is an alternative to KP. Both rely on the same information, namely $\beta_0, \hat{\gamma}_L$, and $\hat{\Pi}_{W_L}$, and so the results of KP and J2L can be expected to be quite similar. To calculate the invariant heteroskedasticity-robust J2L statistic, first conduct LIML estimation of the restricted subvector model and

calculate $\hat{u}_{0L} = y_0 - W\hat{\gamma}_L$. Then the second stage estimate of γ is

$$\hat{\gamma}_{2L} = (\hat{\Pi}'_{W_L} Z'Z(Z'H_{\hat{u}_{0L}}Z)^{-1}Z'W)^{-1}\hat{\Pi}'_{W_L} Z'Z(Z'H_{\hat{u}_{0L}}Z)^{-1}Z'y,$$

where $Z'H_{\hat{u}_{0L}}Z = \sum_{i=1}^{n} \widehat{u}_{0Li}^2 z_i z_i'$. Then $\hat{u}_{02L} = y_0 - W\hat{\gamma}_{2L}$ and the Sargan form of the J2L statistic can be written as

$$J2LS(\beta_0, \hat{\gamma}_L, \hat{\Pi}_{W_L}) = \hat{u}'_{02L} Z(Z'H_{\hat{u}_{02L}}Z)^{-1}Z'\hat{u}_{02L}.$$

For the Basmann form, $\hat{e}_{0L} = M_Z\hat{u}_{0L}$ so the second stage estimate is

$$\hat{\gamma}_{2L} = (\hat{\Pi}'_{W_L} Z'Z(Z'H_{\hat{e}_{0L}}Z)^{-1}Z'W)^{-1}\hat{\Pi}'_{W_L} Z'Z(Z'H_{\hat{e}_{0L}}Z)^{-1}Z'y$$

and $\hat{e}_{02L} = M_Z\hat{u}_{02L}$. Then the Basmann J2L statistic is

$$J2L(\beta_0, \hat{\gamma}_L, \hat{\Pi}_{W_L}) = \hat{u}'_{02L} Z(Z'H_{\hat{e}_{02L}}Z)^{-1}Z'\hat{u}_{02L}.$$

1.5 Previous Methods in Subvector Testing and Inference

Early work in subvector testing employed a projection approach. Dufour and Taamouti (2005) describe a method for producing projection-based confidence intervals for a subvector of parameters, using the AR statistic. Their method is conservative, although Chaudhuri and Zivot (2011) later improved the power. This less conservative method involves calculating a confidence set for nuisance parameters, then finding the infimum of a test statistic over the nuisance parameter with the added restraint that it must belong to the confidence set. Similarly

to Chaudhuri and Zivot (2011), Andrews (2017) starts by finding a confidence set for the nuisance parameter. Then the null hypothesis, in terms of the parameter of interest, is rejected if the test statistic is greater than the critical value at a data-dependent significance level, adapted to account for the strength of the instruments, for all values of the nuisance parameter within the confidence set. The second step in this two-step process can be inverted to produce a confidence set for the parameter of interest.

Another option is to first estimate the nuisance parameters and then use the resulting point estimates in the calculation of the test statistic. This is the approach taken in the methods described in Section 1.3.2, including GKM and GKM2. In particular, these plug-in methods involve estimating the nuisance parameters by LIML and calculating the AR statistic.

Bootstrapping has also been used to conduct inference in IV settings. The idea of bootstrapping is to resample with replacement from the sample data and calculate the test statistic of interest on each replication, in order to produce, after a large number of replications, a distribution of the test statistic. For a test of size 0.05, one would then compare the test statistic from the original sample to the 95th percentile of the bootstrapped distribution rather than to the 95th percentile of the asymptotic distribution.

Davidson and MacKinnon (2010) bootstrap several statistics in the context of IV regression, including the AR statistic. Specifically, they adapt the heteroskedasticity-robust wild bootstrap first proposed by Wu (1986), which involves multiplying the residuals in each iteration by a random variable. Liu (1988) notes that it is desirable for the random variable to have a distribution with mean zero, variance one, and third moment one. The distribution suggested by Mammen (1993) has the desired three moments but also an undesirable fourth moment of 2 and Davidson

and Flachaire (2008) show that it is outperformed in practice by the Rademacher distribution, which is equal to either 1 or -1, each with probability 0.5, and therefore its first four moments are 0, 1, 0, 1. Thus, Davidson and MacKinnon (2010) bootstrap the AR statistic, among others, with the wild bootstrap implemented with the Rademacher distribution.

Wang and Doko Tchatoka (2018) bootstrap the AR statistic in the context of subvector hypothesis testing under the assumption of conditional homoskedasticity. They consider the case of two endogenous regressors, one under test and the other not, and find that the bootstrap works well if the nuisance parameter is strongly identified, but that the bootstrap is inconsistent under weak identification of the nuisance parameter. Specifically, they show that the subset bootstrap fails to properly replicate both (1) the strength of the instruments and (2) the level of endogeneity in the original sample. The first problem results from the fact that there is an approximation error when calculating the bootstrapped identification strength of γ that is negligible under strong identification but relevant under weak identification and the second problem results from the fact that the bootstrapped endogeneity parameter has an extra term when compared to the equation for the original sample endogeneity parameter, so the bootstrap does not well-replicate the appropriate level of endogeneity in the model. The consequence of these problems is that the bootstrapped statistics do not converge to the appropriate null limiting distributions. In their simulations, however, the LIML bootstrap is still size-controlled, whereas the other estimation methods like the 2SLS bootstrap overreject severely. Although Wang and Doko Tchatoka (2018) do not comment on the reason for this result, it can be understood within the context of the result from Section 1.2.7 that rank tests under weak instruments are conservative. While the LIML-based test is a rank test, the other methods

like the 2SLS bootstrap are not.

Wang and Doko Tchatoka (2018) also propose a Bonferroni-based size correction to the asymptotic AR test. Their method involves creating a confidence set for the nuisance parameter, as in Chaudhuri and Zivot (2011) and Andrews (2017), rather than a point estimate as in GKM. However, their subsequent use of data-dependent critical values is similar to the GKM approach and they also add a negative size correction factor into their critical value calculation to reduce the conservativeness of the test, under the constraint that it must still control for size. That is, their size correction factor raises rejection rates but not too much as to overreject. This proposed asymptotic adjustment is not particularly useful in our setup because it relies on the assumption of a single nuisance parameter and so we are more interested in their bootstrap method, which performs well when implemented with LIML.

Their bootstrap algorithm estimates Π_X and Π_W by OLS and γ by one of four estimation methods, including LIML, under the null. Davidson and MacKinnon (2015) also examine several data generating processes (DGPs) similar to this strategy, in the context of overidentification tests under homoskedasticity. For example, in one of their methods, they estimate Π_W by OLS and γ by 2SLS and in another they estimate both by LIML. They find that bootstrapping improves the performance of overidentification tests with moderately weak instruments, but comes with a modest cost to power. MacKinnon (2013) argues that bootstrapping is particularly useful for improving p-values and confidence intervals for heteroskedasticity-robust test statistics, a setting in which the wild bootstrap outperforms the pairs bootstrap. Although heteroskedasticity-robust test statistics, including the tests considered in this chapter, are asymptotically valid, bootstrapping provides a refinement by converging to the asymptotic distribution

more quickly.

If typical bootstrapping cannot be used, for example in cases of non-smooth functions, sometimes m-out-of-n bootstrapping or subsampling are used instead. Both the m-out-of-n bootstrap and subsampling use a sample size smaller than n, with replacement in the former and without replacement in the latter. These techniques can have their own problems, however. For example, Andrews and Guggenberger (2009) show that subsampling tests in the context of weak instruments may be inappropriately sized when testing exogenous controls. This is not a problem here for two reasons. First, we are considering hypothesis tests on the endogenous variables that are instrumented for rather than exogenous controls and Andrews and Guggenberger (2006) show that subsampling is approximately appropriately sized in this case. Second, subsampling and m-out-of-n bootstrapping are unnecessary in our context, given the well-behaved nature of the statistics we consider. In cases like ours where subsampling or the m-out-of-n bootstrap is unnecessary, the standard bootstrap is preferred because it provides second-order refinements, a result proved by Moreira et al. (2009), who also show the bootstrapped score statistic is valid under weak instruments, although with less refinement than under strong instruments.

1.6 Simulation Results under Homoskedasticity

1.6.1 Procedure

1.6.1.1 One Nuisance Parameter

For our simulations, we use a sample size of $n = 250$, have $k_Z = 6$ instruments, and set $k_X = 1$. We test the null hypothesis $H_0 : \beta = \beta_0$ at the 5% level. In the

case of one nuisance parameter, $k_W = 1$, we set the nuisance parameter $\gamma = 0.5$ and use the covariance matrix from GKM19:

$$\Sigma = \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.3 \\ 0.8 & 0.3 & 1 \end{pmatrix}.$$

We take

$$\pi_x = (4/\sqrt{k_Z n}) \begin{pmatrix} 1, & 1, & 1, & 1, & 1, & 1 \end{pmatrix}'$$

and

$$\pi_w = (\pi_\gamma/\sqrt{k_Z n}) \begin{pmatrix} 1, & -1, & 1, & 1, & 1, & 1 \end{pmatrix}'$$

and we vary π_γ through the values 1, 2, 4, 8, and 16 to represent a range from very weak instruments ($\pi_\gamma = 1$) to very strong instruments ($\pi_\gamma = 16$). For each of the 10,000 repetitions in the Monte Carlo simulation, we generate k_Z random instruments as well as a random error matrix V from the multivariate normal distribution with mean zero and covariance Σ. We then calculate x, w, y, and $y_0 = y - x\beta_0$ from our generated instruments and errors as well as π_x and π_w. We directly compare the GKM test statistic to the data-dependent critical values tabulated in Guggenberger et al. (2019). We estimate the restricted equation $y_0 = w\gamma + u$ by LIML in order to calculate the AR, KP, and J2L statistics, each in both their Basmann and Sargan forms, and to use the LIML estimates to calculate the restricted residuals. We then estimate the unrestricted equation $y = x\beta + w\gamma + u$ by LIML and calculate the unrestricted residuals (see Appendix 1.9).

For each of $B = 399$ bootstraps, we randomly resample with replacement n groups of instruments and residuals from the original sample of size n. The

wild bootstrap is implemented by multiplying each residual by either -1 or 1 each with probability 0.5, although the implementation of the wild bootstrap, which makes the bootstrap heteroskedasticity-robust, should not make a difference in the homoskedastic case. Using the new restricted residual values, we calculate w_R^*, the bootstrapped values of w, and y_{0R}^*, the bootstrapped values of y_0. Note that Wang and Doko Tchatoka (2018) calculate bootstrapped values of x and y and then calculate $y_0^* = y^* - x^*\beta_0$, but this is unnecessary when we can resample y_0 directly. Next, we estimate the restricted equation $y_{0R}^* = w_R^*\gamma + u$ by LIML using the bootstrapped instruments Z^* and calculate the bootstrapped restricted AR, KP, and J2L statistics in both their Basmann and Sargan forms. A similar procedure is repeated using the unrestricted residual values to calculate w_{UR}^* and y_{0UR}^*, estimate $y_{0UR}^* = w_{UR}^*\gamma + u$ by LIML, and calculate the corresponding bootstrapped unrestricted test statistics. The results for the unrestricted bootstrap are reported in the appendix. P-values for the asymptotic, restricted, and unrestricted versions of the AR, KP, and J2L tests are calculated by comparing the values from the restricted Monte Carlo simulations to the 95th percentile of the $\chi^2_{k_Z-k_W}$distribution, the bootstrapped restricted distribution, and the bootstrapped unrestricted distribution respectively.

In steps, the procedure for the restricted bootstrap is as follows:

1. Generate Z and V.

2. Calculate x, w, y and y_0.

3. Estimate the restricted equation imposing the null $\beta = \beta_0$ by LIML and calculate the test statistics.

4. Calculate the LIML residuals:

$$\hat{u}_L = y_0 - w\hat{\gamma}_L$$
$$\hat{V}_{wL} = w - Z\hat{\pi}_{wL}$$

5. For each of the $b = 1, ..., B$ bootstraps, jointly resample with replacement n times from the rows of $\hat{u}_L$, $\hat{V}_{wL}$, and Z. Then multiply the errors $\{\hat{u}_{Li}, \hat{V}_{wLi}\}$ by a random variable equal to -1 or 1 each with 50% probability. This step will create the bootstrapped terms u^*_{Lb}, V^*_{wLb}, and Z^*_b.

6. Calculate w^*_b and y^*_{0b}:

$$w^*_b = Z^*_b\hat{\pi}_{wL} + V^*_{wLb}$$
$$y^*_{0b} = w^*_b\hat{\gamma}_L + u^*_{Lb}$$

7. Estimate each of the $b = 1, ..., B$ bootstrapped restricted equations by LIML, recentring w^*_b and y^*_{0b}, and calculate the test statistics.

8. Calculate the 95th percentile of the bootstrapped distribution of each test statistic. Reject the hypothesis if the test statistic calculated in step 3 exceeds this 95th percentile threshold.

1.6.1.2 Two Nuisance Parameters

A similar procedure is used in the case of $k_W = 2$, with the following differences:

$$\gamma = (1, -1)\,,$$

$$\sum = \begin{pmatrix} 1 & 0.5 & 0.4 & 0.8 \\ 0.5 & 1 & 0.3 & 0.1 \\ 0.4 & 0.3 & 1 & 0.2 \\ 0.8 & 0.1 & 0.2 & 1 \end{pmatrix},$$

and

$$\Pi_W = (\pi_\gamma / \sqrt{k_Z n}) \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}'.$$

The variance matrix in this case is taken from Grohmann (2021).

1.6.2 Simulation Results

In total there were 10 asymptotic tests: AR, KP, J2L, GKM, GKM2, ARS, KPS, J2LS, GKMS, and GKM2S. Recall that GKM involves conditioning on the largest eigenvalue, whereas GKM2 involves conditioning on the second smallest eigenvalue. They are therefore identical in the case of $k_W = 1$, which has 2 eigenvalues. The lines labelled GKM in the $k_W = 1$ graphs thus also represent GKM2.

In addition to the asymptotic tests, there are also both restricted and unrestricted bootstrapped tests of AR, KP, and J2L. Bootstrapping the Sargan versions ARS, KPS, and J2LS produces numerically identical results in the case of AR and ARS and similar results in the case of KP and KPS and J2L and J2LS.

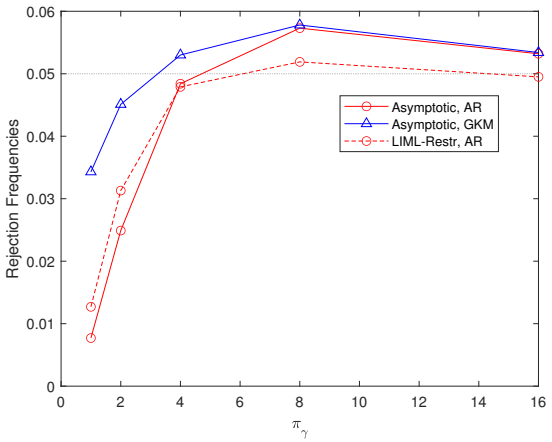
A perfect test, in our case of testing $H_0 : \beta = 0$, would have a constant rejection frequency of 0.05 when $\beta = 0$, regardless of the strength of the instruments, given by π_γ. However, as we see in the size graphs below, it is common for tests to be undersized, i.e. have a rejection frequency less than 0.05 when instruments are very weak, i.e. $\pi_\gamma = 1$ or $\pi_\gamma = 2$. The ideal test would also increase rapidly in rejection frequency as β moves away from 0, indicating that it has the power to

detect when the null hypothesis is untrue. Flat power graphs indicate that the test is not well able to distinguish cases where the null hypothesis holds from cases where it does not.

1.6.2.1 Basmann-Type Variance, $k_W = 1$

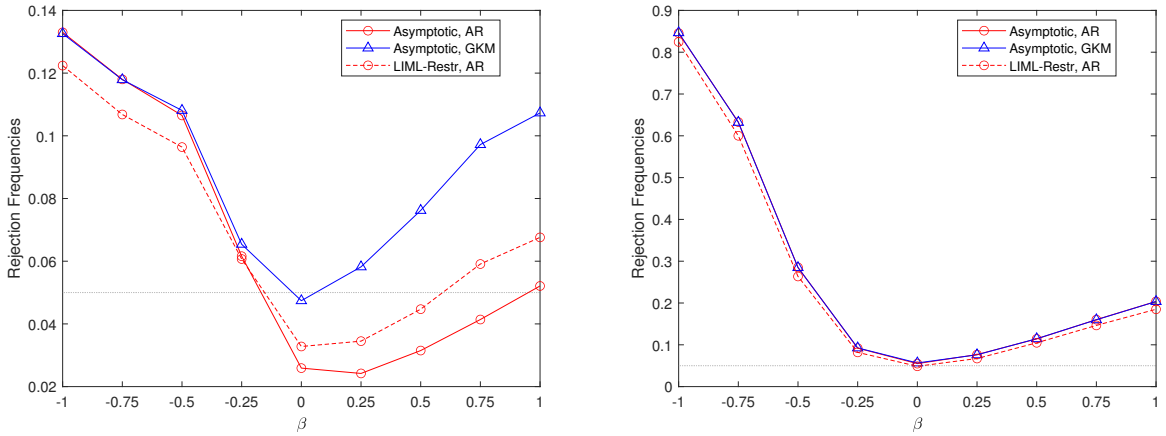
First, we compare previously available asymptotic methods, namely the AR and GKM tests, to a restricted bootstrap of the AR test. Recall that the GKM test is simply the asymptotic AR test with data-dependent conditional critical values. In Figure 1.2, a larger value for π_γ represents stronger instruments and so we see that, as expected, under strong instruments, the size of asymptotic AR and GKM converge, both overrejecting slightly, and that the LIML-based restricted bootstrap of the AR statistic improves upon both these previously used asymptotic tests. Under very weak instruments, i.e. $\pi_\gamma = 1$ or $\pi_\gamma = 2$, we see that GKM has better size properties than the asymptotic AR - namely, it underrejects less severely - because of its conditioning on the larger of the two eigenvalues, and in fact it even performs better than the LIML-based restricted AR bootstrap in terms of size under weak instruments. Additionally, Figure 1.3 shows that the GKM test also has better power properties than the other two under weak instruments ($\pi_\gamma = 2$), but all three tests converge under strong instruments ($\pi_\gamma = 16$). The power curves are asymmetric, with higher power for negative values of β than for postive values of β, a trend that appears in all power graphs reported in this chapter. Van de Sijpe and Windmeijer (2022) show analytically that the power curve of the AR test is asymmetric for the standard $k_x = 1$ case. Here we find that this asymmetry carries over to the subset AR test, as was also found by GKM.

Figure 1.2: Size of Asymptotic AR, GKM, and Restricted Bootstrap AR Tests, k_W=1, under homoskedasticity



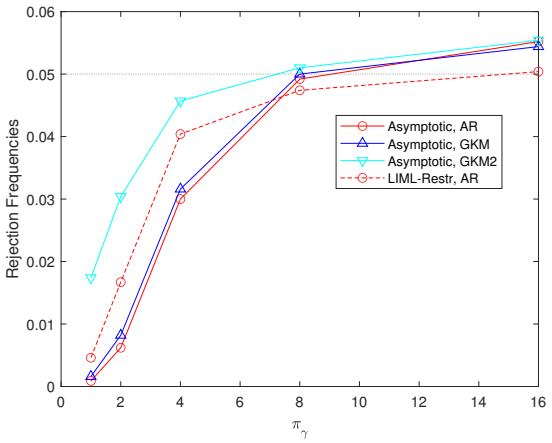
Note: The asymptotic AR-based tests are slightly oversized but implementing the bootstrap improves the size properties.

Figure 1.3: Power of Asymptotic AR, GKM, and Restricted Bootstrap AR Tests, k_W=1, under homoskedasticity, with weak instruments (left, $\pi_\gamma = 2$) and strong instruments (right, $\pi_\gamma = 16$)



Note: Under weak instruments, GKM has higher power than either the asymptotic or bootstrapped AR test, but all three tests converge under strong instruments.

Figure 1.4: Size of Asymptotic AR, GKM, and Restricted Bootstrap AR Tests, k_W=2, under homoskedasticity

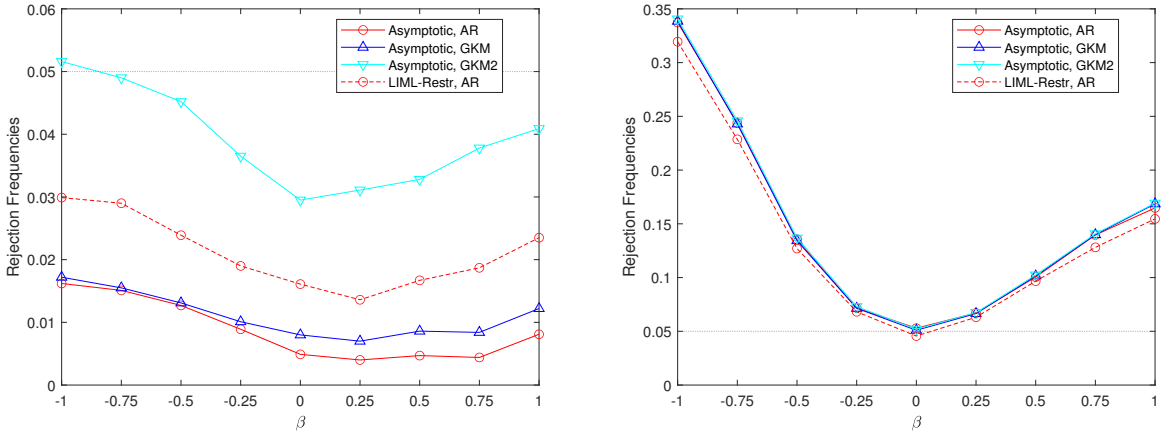


Note: With two nuisance parameters, GKM2 is less undersized under weak instruments than GKM, but is slightly oversized under strong instruments, in contrast to the appropriately sized AR bootstrap when $\pi_\gamma = 16$.

1.6.2.2 $k_W = 2$

When there are two nuisance parameters, GKM2, which was equivalent to GKM in the $k_W = 1$ case, again has better size properties than the LIML-based restricted AR bootstrap under weak instruments and worse size properties under strong instruments (see Figure 1.4). GKM2 also dominates the AR bootstrap in terms of power under weak instruments (Figure 1.5). The LIML-based restricted AR bootstrap outperforms the asymptotic AR and GKM tests under both weak and strong instruments, however, in terms of size (Figure 1.4) and also outperforms both in terms of power under weak instruments (Figure 1.5). As usual, the tests show a high degree of convergence under strong instruments. Under weak instruments, the largest eigenvalue of the AR matrix is quite large and so conditioning on it, the strategy of GKM, does not have as much benefit as conditioning on the second smallest eigenvalue, the strategy of GKM2, leading to a clear preference for GKM2.

Figure 1.5: Power of Asymptotic AR, GKM, and Restricted Bootstrap AR Tests, k_W=2, under homoskedasticity, with weak instruments (left, $\pi_\gamma = 2$) and strong instruments (right, $\pi_\gamma = 16$)



Note: Under weak instruments, the tests all perform poorly. A good test should have a rejection rate of 0.05 at $\beta = 0$ and rejection rates higher than 0.05 away from $\beta = 0$. The tests shown on the lefthand side of the figure are all undersized and greatly lacking in power, although GKM2 outperforms the others. On the righthand side of the figure, under strong instruments, the tests converge and all are appropriately sized and have better power properties.

1.6.2.3 Sargan-Type Variance

Sargan-type variances do not subtract out the part of the error explained by the instruments and so have lower power than the corresponding Basmann-variance statistics. However, the Sargan-variance statistics are also less likely to overreject. Figure 1.6 shows that asymptotic KP and J2L statistics overreject in their Basmann forms, whereas AR, GKM, and GKM2 are closer to the correct size, but this distinction does not exist with the Sargan forms, which are all quite similar. Note that while the Basmann-variance statistics may overreject, the Sargan-variance statistics may underreject (see Figure 5, right: $k_W = 1$ and $\pi_\gamma = 16$; $k_W = 2$ and $\pi_\gamma = 8$). Relatedly, Figure 1.7 shows that the Basmann-variance versions of the asymptotic statistics have higher power than the Sargan-variance versions and that while the KP and J2L statistics have notably higher power than the others

in Basmann form, the differences between statistics diminish in the Sargan form.

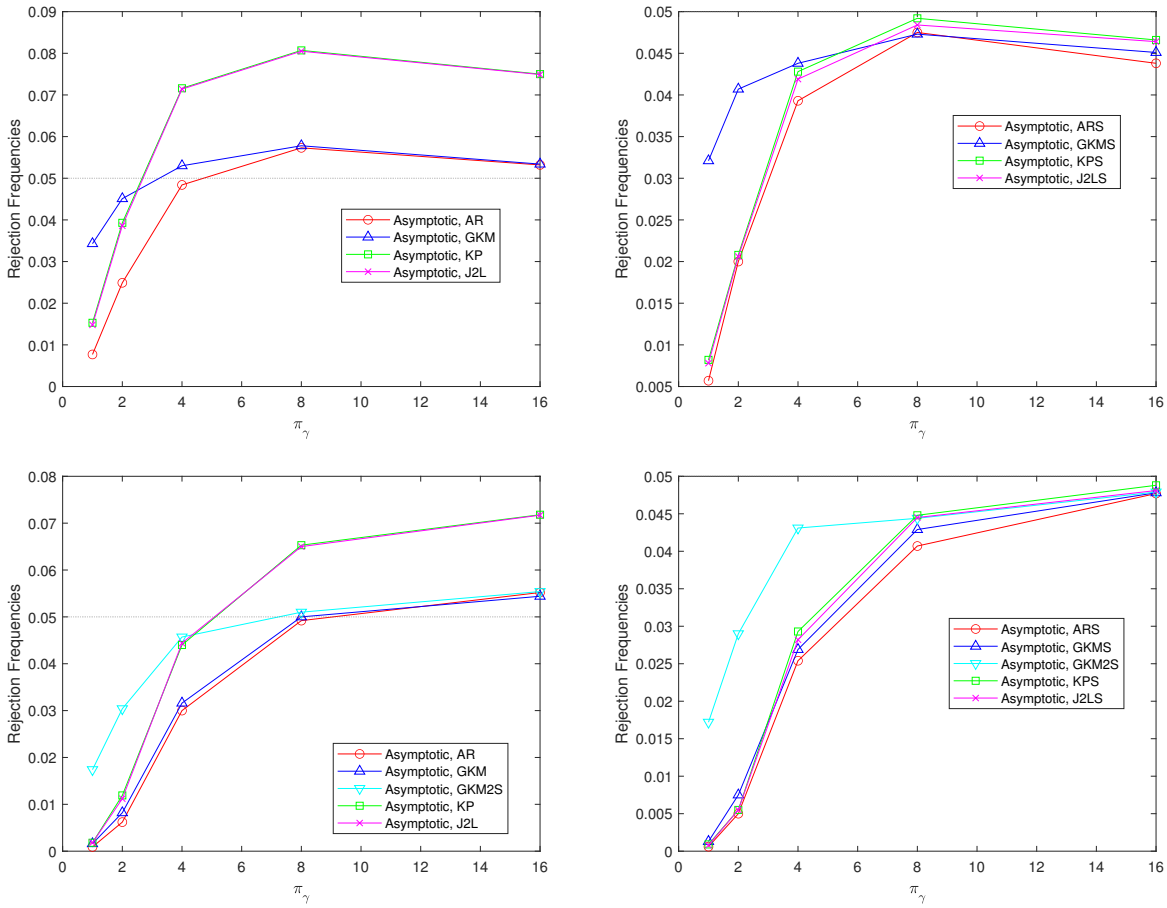
While the Basmann-form asymptotic statistics, especially KP and J2L, have higher rejection rates than their Sargan-form counterparts, leading to improved power but potentially distorted size, the bootstrapped Basmann-form and Sargan-form statistics are nearly indistinguishable. Figures 1.8 and 1.9 show that the use of either variance results in tests of similar size and power, respectively. Interestingly, the restricted bootstrap of the KP and J2L statistics with Sargan-type variances even shows a slightly greater degree of overrejection than the corresponding statistics with Basmann-type variances in the size diagrams when $k_W = 1$ and $\pi_\gamma = 8$ (see Figure 1.8, upper panel). This slight deviation runs in a contrary direction to the pattern established with the asymptotic statistics, but overall the Basmann and Sargan restricted bootstraps have strikingly similar results. Note also that the bootstrapped statistics suffer from some power loss compared to GKM (see Figure 1.9).

1.7 Simulation Results under Heteroskedasticity

1.7.1 Procedure

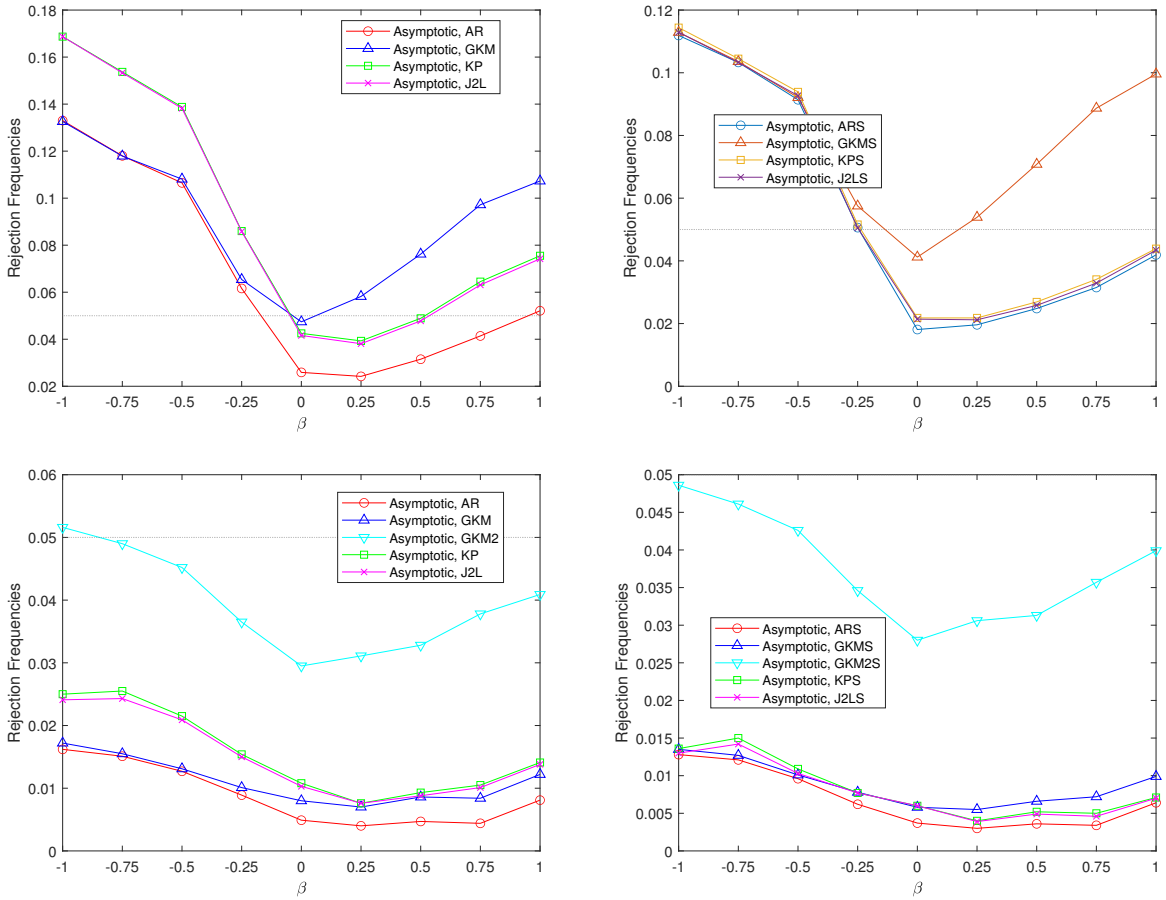
The procedure for the heteroskedastic bootstrap is the same as the homoskedastic version, but with a different procedure for generating the errors. First, generate an instrument matrix Z with n rows and k_Z columns as well as an error matrix V with n rows and k_W+2 columns. The first column of V will be the structural error u, the second column will be v_X and the remaining columns will be V_W. Then, the variance of u is made heteroskedastic according to the following equation:

Figure 1.6: Size of Asymptotic AR, GKM, GKM2, KP, and J2L Tests, $k_W = 1$ (top), $k_W = 2$ (bottom), Basmann variance (left), Sargan variance (right), under homoskedasticity



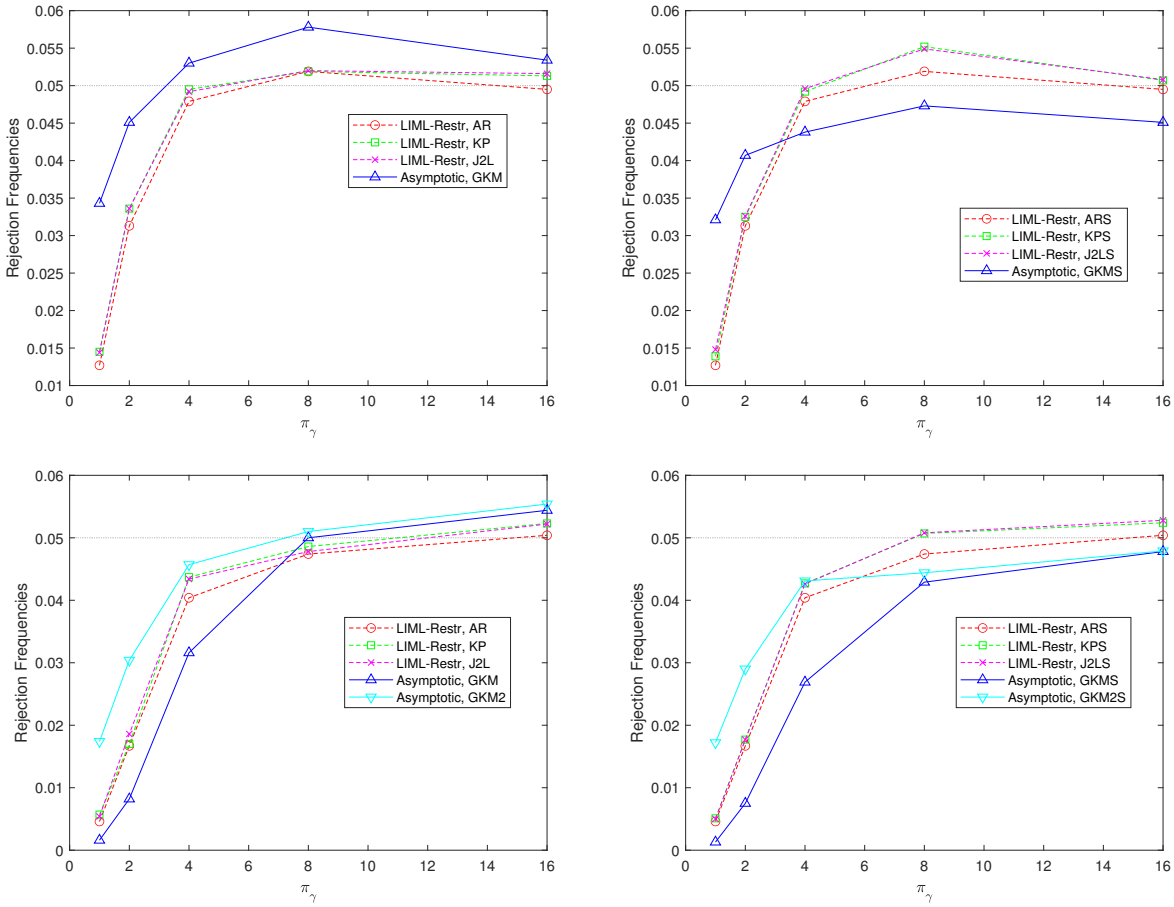
Note: With a Basmann variance (left), the asymptotic KP and J2L tests are oversized, whereas with the Sargan variance (right) the rejection rates of these two tests are comparable to the other asymptotic tests. In general, the Basmann rejection rates are higher than their Sargan counterparts.

Figure 1.7: Power of Asymptotic AR, GKM, GKM2, KP, and J2L Tests, $k_W = 1$ (top), $k_W = 2$ (bottom), Basmann variance (left), Sargan variance (right), under homoskedasticity and weak instruments ($\pi_\gamma = 2$)



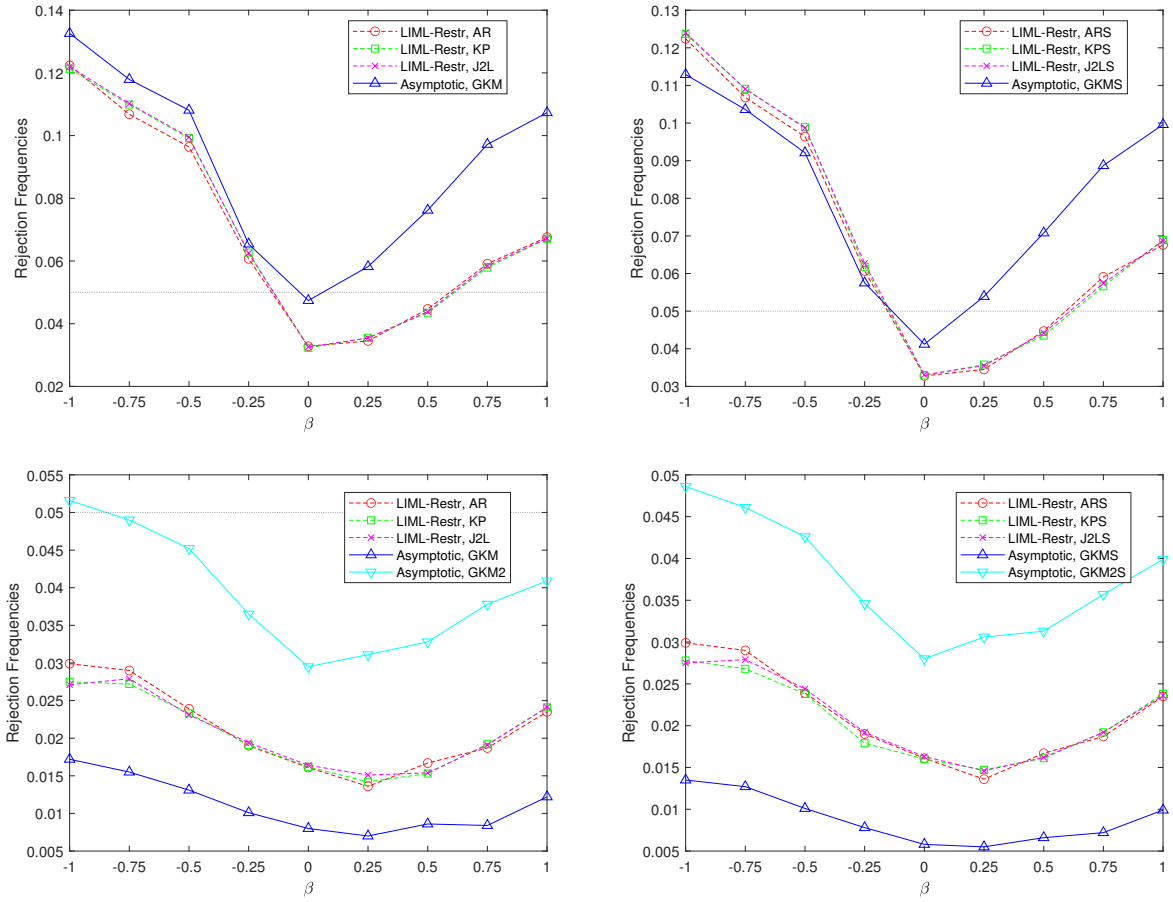
Note: In the top of the figure, with one nuisance parameter, GKM outperforms the other asymptotic tests, although KP and J2L have higher power against negative hypothesized values of β when using the Basmann variance (left). In the bottom of the figure, with two nuisance parameters, GKM2 outperforms the other asymptotic tests, although none of them have power in this weak instrument setting. Rejection rates with the Basmann variance (left) are higher than with the Sargan variance (right).

Figure 1.8: Size of Restricted Bootstrap AR, KP, and J2L Tests Compared to GKM, $k_W = 1$ (top), $k_W = 2$ (bottom), Basmann variance (left), Sargan variance (right), under homoskedasticity



Note: Bootstrapped Basmann (left) and Sargan (right) statistics are quite similar, in contrast to the asymptotic versions.

Figure 1.9: Power of Restricted Bootstrap AR, KP, and J2L Tests Compared to GKM, $k_W = 1$ (top), $k_W = 2$ (bottom), Basmann variance (left), Sargan variance (right), under homoskedasticity and weak instruments ($\pi_\gamma = 2$)



Note: The bootstapped statistics suffer a modest power loss compared to GKM.

$$u_H = \text{diag}\left(\frac{\sqrt{n}exp(0.7z_1)}{exp(0.7z_1)'exp(0.7z_1)}\right)u,$$

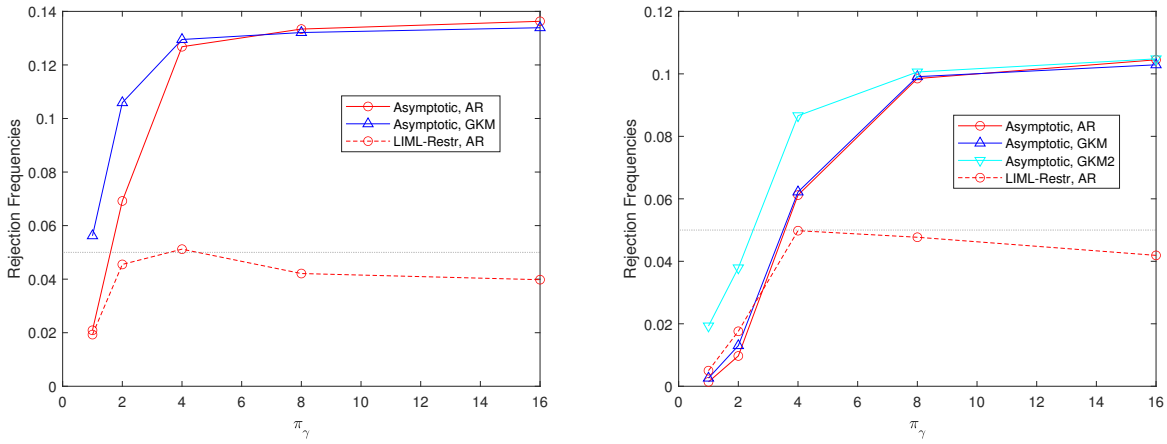
where diag(a) is the diagonalisation of vector a and z_1 is the first column of Z. Then update V as $V = V * upperchol(\sum)$, where $upperchol(A)$ represents the upper triangular matrix in the Cholesky decomposition of A and $\sum$ is the covariance matrix as defined in Section 1.6. We take u, v_X, and V_W as the first, second, and remaining columns, respectively, of the updated matrix V.

1.7.2 Simulation Results

1.7.2.1 Basmann-Type Variance

The AR-based tests assume conditional homoskedasticity and so do not perform well asymptotically in a heteroskedastic setting. Figure 1.10 shows that the bootstrapped AR test improves upon the asymptotic AR and GKM tests. As in the homoskedastic case, GKM is nearly identical to asymptotic AR in the $k_W = 2$ case because conditioning on the largest eigenvalue does not make much difference when that eigenvalue is very large. GKM2, however, is distinct from AR and GKM. With $k_W = 2$, there are three eigenvalues: the smallest is the AR statistic, the second smallest is conditioned on in GKM2, and the largest is conditioned on in GKM. This is in contrast to the two eigenvalues present in the $k_W = 1$ case, in which the smaller is the AR statistic and the larger is conditioned on by GKM/GKM2. Note that while GKM2 generally outperforms GKM/AR when $k_W = 2$, it can be seen in Figure 1.10 that the overrejection is more severe when $\pi_\gamma = 4$ and in fact GKM also overrejects somewhat more than asymptotic AR when $k_W = 1$ and $\pi_\gamma = 2$. The most notable feature in Figure 1.10, however, is the stark overrejection of the nonrobust asymptotic methods in the presence

Figure 1.10: Size of Asymptotic AR, GKM, and Restricted Bootstrap AR Tests, $k_W = 1$ (left), $k_W = 2$ (right), under heteroskedasticity

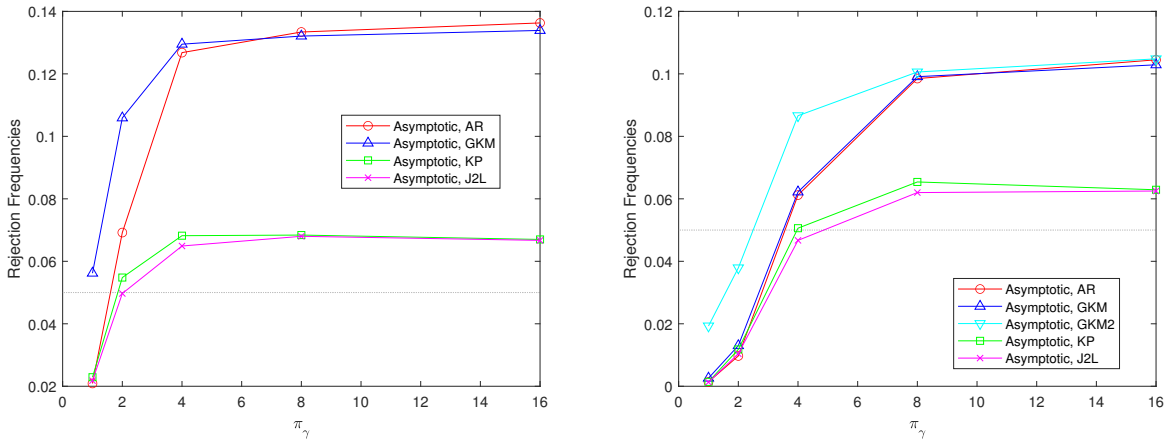


Note: The asymptotic AR-based statistics perform poorly under heteroskedasticity, but the bootstrapped AR maintains approximately the correct size.

of heteroskedasticity even under very strong instruments and the fact that bootstrapping greatly improves the size control.

The reason that asymptotic AR and GKM/GKM2 do not perform well in Figure 1.10 is that they are not robust to heteroskedasticity. When we examine the asymptotic performance of KP and J2L we find they are strikingly better, in the sense of having a smaller degree of overrejection, because they are robust to heteroskedasticity. The asymptotic KP and J2L tests do still overreject, however, even under strong instruments and so there is room for improvement. The bootstrapped tests provide this refinement. In Figure 1.12, we see that GKM has rejection rates over 12% under strong instruments when $k_W = 1$, whereas the bootstrapped methods are closer to 4-5%. The bootstrapped KP and J2L statistics are particularly consistent and do not suffer the same slight drop in rejection rates as the bootstrapped AR does as the instruments grow stronger. The same patterns are present when $k_W = 2$: GKM and GKM2 overreject severely, the bootstraps control size well, and the AR bootstrap underrejects slightly under

Figure 1.11: Size of Asymptotic AR, GKM, GKM2, KP, and J2L Tests, $k_W = 1$ (left), $k_W = 2$ (right), under heteroskedasticity



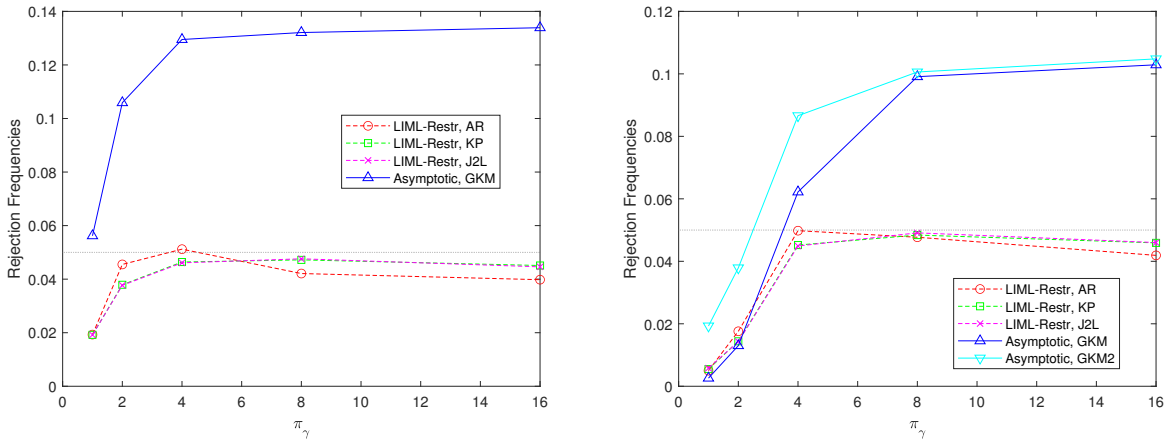
Note: The asymptotic KP and J2L tests are robust to heteroskedasticity and therefore perform better in the presence of heteroskedasticity than do the non-robust AR-based statistics.

especially strong instruments.

1.7.2.2 Sargan-Type Variance

Comparing the two types of variance estimators, we see that switching from Basmann to Sargan again causes a drop in size, as in the homoskedastic case. The asymptotic KP and J2L statistics overreject somewhat in their Basmann forms, but underreject somewhat in their Sargan forms. The asymptotic AR and GKM statistics always overreject, but to a lesser degree in their Sargan forms (see Figure 1.13). Interestingly, GKM2 does much better in its Sargan form than its Basmann form. As the latter, it overrejects severely, but as the former it underrejects only slightly, with overall remarkably well-controlled size. In contrast, the bootstraps always perform well, regardless of which variance matrix is used (see Figure 1.14).

Figure 1.12: Size of Restricted Bootstrap AR, KP, and J2L Tests Compared to GKM, $k_W = 1$ (left), $k_W = 2$ (right), under heteroskedasticity



Note: All of the bootstrapped statistics are approximately appropriately sized, especially compared to GKM, but the bootstrapped KP and J2L are somewhat less undersized under strong instruments than the bootstrapped AR.

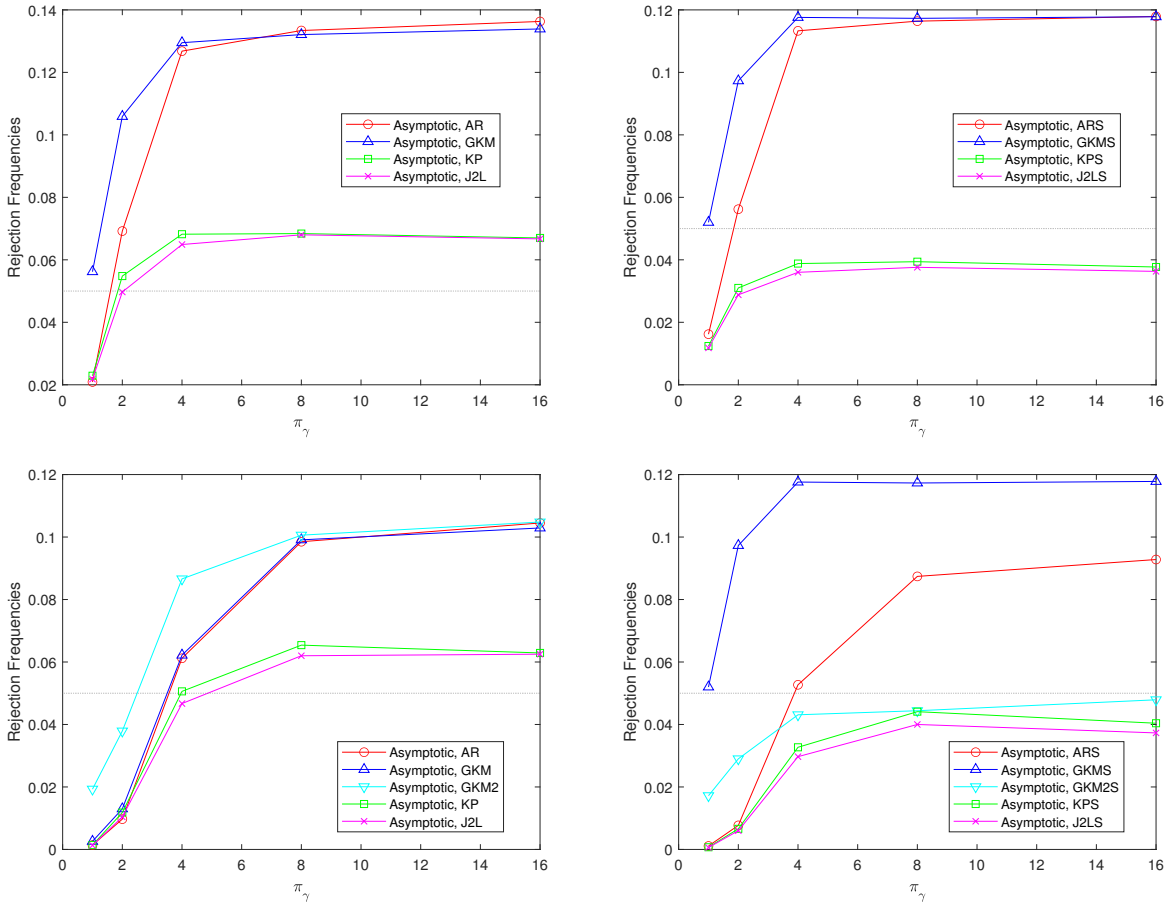
1.7.2.3 Identification of Beta

In our simulations,

$$\pi_X = (\pi_\beta / \sqrt{k_Z n})\,(1, 1, 1, 1, 1, 1)'$$

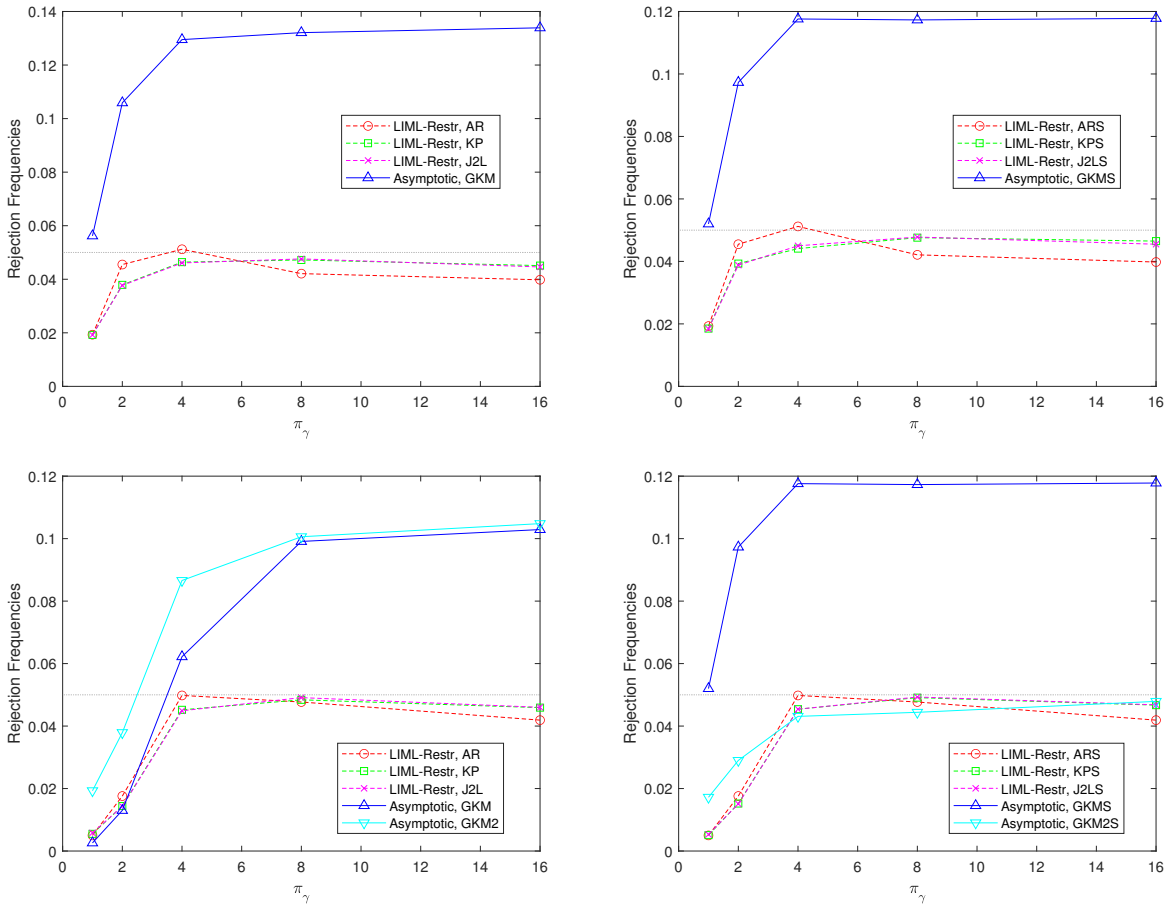
and we have so far set π_β to the constant value of 4, which represents moderately weak identification of β. We now consider the results of varying π_β from very weak to very strong instruments. Figure 1.15 plots the power of both the asymptotic and bootstrapped tests for $\pi_\beta = 1$, $\pi_\beta = 4$, and $\pi_\beta = 16$ and we can see that the power improves as the identification of β strengthens. The size, however, is largely unaffected. Note how the rejection rate of GKM when $\beta = 0$ hovers around 10% regardless of the strength of identification of β, whereas the heteroskedasticity-robust methods tend to have approximately correct size in all situations.

Figure 1.13: Size of Asymptotic AR, GKM, GKM2, KP, and J2L Tests, $k_W = 1$ (top), $k_W = 2$ (bottom), Basmann variance (left), Sargan variance (right), under heteroskedasticity



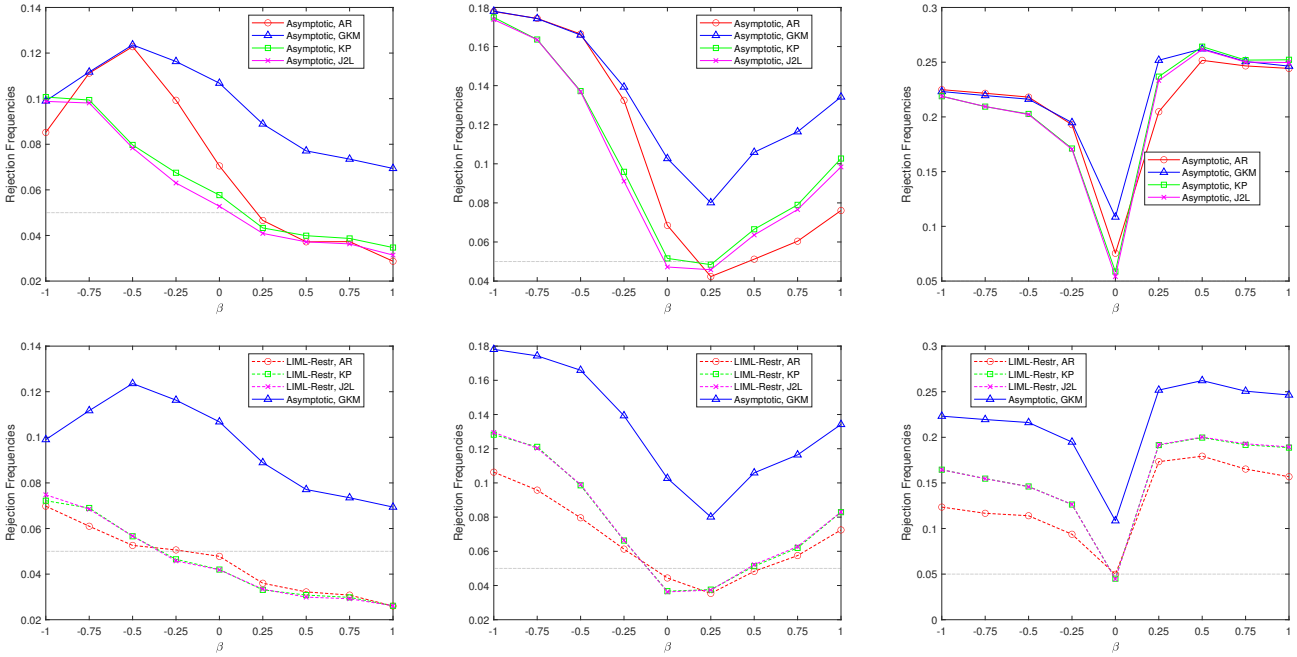
Note: Under heteroskedasticity, the robust asymptotic statistics (KP, J2L) are closer to the appropriate size than the non-robust asymptoptic statistics. In general, Sargan variance rejection rates (right) are lower than Basmann variance rejection rates (left).

Figure 1.14: Size of Restricted Bootstrap AR, KP, and J2L Tests Compared to GKM, $k_W = 1$ (top), $k_W = 2$ (bottom), Basmann variance (left), Sargan variance (right), under heteroskedasticity



Note: The bootstrapped statistics are generally well-sized, seen here compared to the oversized GKM.

Figure 1.15: Power of Asymptotic (top) and Restricted Bootstrap (bottom) Tests, $\pi_\beta = 1$ (left), $\pi_\beta = 4$ (middle), $\pi_\beta = 16$ (right), $k_W = 1$, under heteroskedasticity



Note: The power of each test improves as the identification of β strengthens (left to right).

1.8 Conclusion

This paper has considered the problem of performing hypothesis tests on a subset of the endogenous parameters in an instrumental variables regression under weak instruments and we have proposed bootstrapping the KP or J2L statistic, two choices which use the same information and therefore produce nearly identical results. In contrast to previous approaches, the method of bootstrapping the KP or J2L statistic is robust to unknown forms of heteroskedasticity, a key improvement in making subvector testing more practically useful as many empirical regressions of interest to applied researchers are not homoskedastic and thus the methods reliant on the assumption of homoskedasticity are not appropriate. In our simulations, for example, we show that although GKM is the preferred method based on power, it is greartly oversized under heteroskedasticity. Additionally, bootstrapping provides a size advantage over asymptotic approaches generally and because LIML estimates are used in the calculation of the KP and J2L statistics, they are rank tests and therefore conservative. While the asymptotic KP and J2L tests sometimes overreject, particularly under strong instruments, bootstrapping pushes the rejection rates to a level closer to alpha. Another advantage of our method is that it is applicable for general k_W, in contrast to other methods that only allow for a single nuisance parameter.

Bootstrapping the KP or J2L statistic is a versatile method for subvector testing with desirable size properties that is robust both to weak instruments and heteroskedasticity and applicable in situations with multiple nuisance parameters.

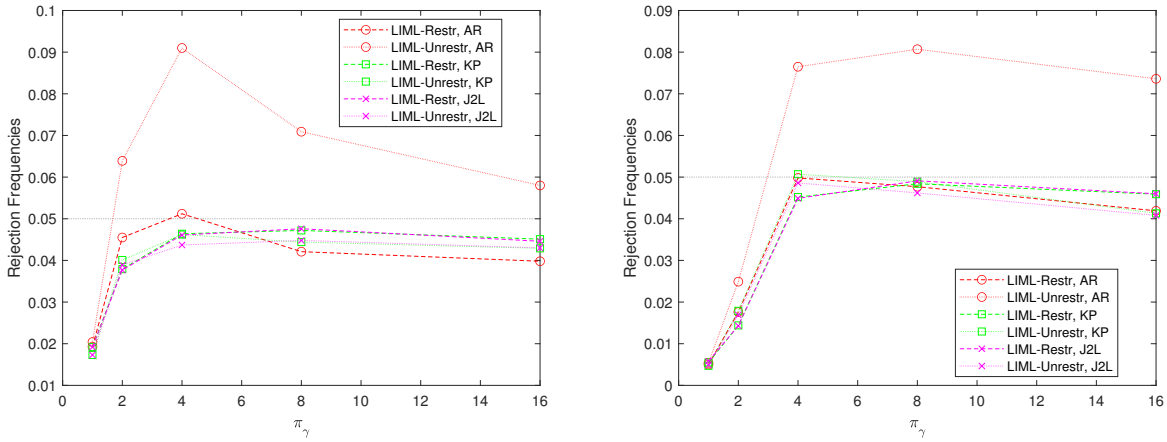
1.9 Appendix

1.9.1 Unrestricted Bootstrap

The restricted bootstrap statistics we have examined thus far have come from estimating the restricted equation $y_0 = W\gamma + u$, with the null hypothesis $H_0 : \beta = \beta_0$ imposed. Alternatively, we could use the unrestricted estimates, obtained from estimating $y = X\beta + W\gamma + u$, in the bootstrap. Overall, the results are rather similar, except for the AR statistic. Both the restricted and unrestricted KP and J2L statistics control size well but slightly underreject and the restricted AR statistic is in the same cluster of slight underrejection. The unrestricted AR statistic however, overrejects considerably (see Figure 1.16) .

The reason the unrestricted bootstrap is not included in the main text is that we are performing overidentification tests on the restriction $E(Z_i u_i) = 0$ for the restricted equation $y_0 = y - X\beta = W\gamma + u$. As explained in Section 1.3, if the null $\beta = \beta_0$ is true, then the restricted model is overidentified and the test fails to reject. Thus it makes sense to use the estimates from the restricted equation in the test statistics as opposed to the corresponding unrestricted estimates because it is the restricted equation we are testing.

Figure 1.16: Size of Restricted and Unrestricted Bootstrap AR, KP, and J2L Tests, $k_W = 1$ (left), $k_W = 2$ (right), under heteroskedasticity



Note: KP and J2L perform similarly in their restricted and unrestricted bootstrap forms. AR, however, noticeably overrejects in its unrestricted form but not in its restricted form.

Chapter 2

Assessing the Exclusion Restriction and Testing Endogeneity

2.1 Introduction

Instrumental variables regression relies on two assumptions: the relevance condition, which states that the instrument has a nonzero correlation with the endogenous variable, and the exogeneity condition, which states that the instrument has a zero correlation with the error. In the first chapter, we considered the case of a near violation of the relevance condition and described procedures for handling weak instruments in the context of subvector testing under heteroskedasticity. In this chapter, we consider invalid instruments, which violate the endogeneity condition. In particular, we are concerned with the exclusion restriction, which states that the instrument does not independently contribute to the outcome variable, over and above its influence on the outcome through the endogenous treatment

variable.

The exclusion restriction is generally held to be impossible to directly test. This is problematic because it implies that instrumental variables regression, a popular technique in fields such as statistics, economics, and epidemiology, depends on an untestable assumption, casting doubt upon the conclusions of a large number of papers. However, there are some methods available for gaining information about the exclusion restriction. The standard method is to conduct a Sargan test, the procedure for which is described in the previous chapter (Sargan, 1958). It is important to recognise that the Sargan test is a test of overidentifying restrictions, which means that the number of instruments (k_Z) must be strictly greater than the number of endogenous variables (k_X) and the test still relies on the untestable assumption that some set of k_X instruments are valid. Windmeijer (2019) provides an intuitive interpretation of the Sargan test. In the case of a regression with one endogenous variable and two instruments, z_1 and z_2, one twice estimates the coefficient on the endogenous variable by just-identified 2SLS (i.e., IV), producing estimates $\hat{\beta}_{IV1}$ and $\hat{\beta}_{IV2}$, when using z_1 or z_2 respectively. The Sargan test rejects the null that all instruments are exogenous when $\hat{\beta}_{IV1}$ and $\hat{\beta}_{IV2}$ are significantly different and fails to reject when they are similar. However, while different IV estimates of β from different instruments are indeed good evidence that at least one of the instruments fails the exclusion restriction, it is not the case that similar IV estimates imply valid instruments. This is the reason that the Sargan test relies on the assumption of k_X valid instruments and only works in overidentified settings.

Another option suggested by Masten and Poirier (2021) is to calculate a falsification adaptive set, which is a range of values for β that are nonfalsified allowing for different model specifications. Similarly to Windmeijer (2019), the bounds

of the set correspond to different estimates for β calculated by using only one instrument at a time in an IV regression. However, Masten and Poirier (2021) also include the unused instruments as regressors with the rationale that the goal is to estimate what β would be if the exclusion restriction did not hold for some of the instruments. When the exclusion restriction is violated, the instrument has a direct effect on the outcome variable and should therefore be included as a regressor in the model. This method still requires the unverifiable validity of at least one instrument, but presents a range of estimates to provide information in case any of the instruments are invalid and therefore acts as a type of sensitivity analysis.

The plausibly exogenous method, originally proposed by Conley et al. (2012), provides a way to adjust the estimate of the effect of X based on assumed or estimated information about γ in the model $y = X\beta + Z\gamma + u$. In particular, if there is a subsample for which Z necessarily has no relationship with X, then an estimate for γ can be calculated in that subsample and then used to estimate β in the full sample (van Kippersluis and Rietveld, 2018).

Rather than attempt to account for possible violations of the exclusion restriction, Kiviet (2020) proposes a method to account for the possible endogeneity of X. While the Sargan test, the falsification adaptive set method, and the plausibly exogenous method all consider the model $y = X\beta + Z\gamma + u$ and are concerned with the possibility that γ may not be zero, Kiviet (2020) considers the model $y = X\beta + u$, which does not include any instruments. This technique of instrument-free inference calculates estimates for β based on assumed values of the correlation between X and u. This correlation is of course a measure of the endogeneity of X. A researcher can thus produce a range of β estimates across a range of endogeneity estimates. This method, known as Kinky Least Squares

(KLS), adjusts the OLS estimate of β to account for the bias introduced by the assumed level of endogeneity. KLS dodges the issue of a potentially violated exclusion restriction by not including any instruments at all and instead performs a sensitivity analysis directly on the endogeneity of X, the problem motivating all instrument-based regression.

But Kiviet (2020) additionally argues that the same logic behind the KLS method of estimating β can be harnessed to create a test of the exclusion restriction, a bold claim given that testing the exclusion restriction in the case of a single instrument has long been thought to be impossible. Recall that the previously mentioned assessments of the exclusion restriction depended on having at least one valid instrument and therefore are unable to evaluate the validity of a single instrument, which is a popular case in practice. In order to conduct his test, Kiviet (2020) returns to the standard model in evaluations of the exclusion restriction: $y = X\beta + Z\gamma + u$. He develops a hypothesis test of the null $H_0 : \gamma = 0$ at each potential value of the correlation between X and u and claims that if the test fails to reject in a range of plausible values for the endogeneity correlation, then the researcher has evidence in favour of instrument validity, a method purported to work even when $k_Z = 1$. However, as acknowledged in Kiviet (2022), this test is misleading. The problem is that, in the case of one instrument, the correlation implied by the IV estimate of β produces a KLS estimate of γ that tends to zero and so the instrument always appears valid at that level of correlation, regardless of the true validity of the instrument.

Independently from Kiviet (2022), we show in this chapter that the KLS estimate of γ calculated at the endogeneity level implied by IV does not just tend to zero asymptotically but is in fact equal to zero in finite samples. This proves that the proposed test of the exclusion restriction is not valid. However, it is still

possible to gain some information about the exclusion restriction even though the test of $H_0 : \gamma = 0$ fails. We suggest a reformulation such that we directly test the endogeneity correlation rather than testing the exclusion restriction per se. That is, we test $H_0 : \rho_{xu} = r_{xu}$, where $\rho_{xu} = \frac{\sigma_{xu}}{\sigma_x \sigma_u}$ is the correlation between X and the error and r_{xu} is a particular value chosen by the researcher. We show that the KLS exclusion restriction test also functions as a test of ρ_{xu} and we additionally propose several alternate methods for testing ρ_{xu} and examine their performance. In each case, the researcher chooses a plausible value for the endogeneity correlation and tests the hypothesis that the correlation is that value. A rejection of the hypothesis then casts doubt upon the validity of the instrument, but one must still be careful not to conclude that an instrument is valid based on a failure to reject the hypothesised correlation. Testing the impossible is, after all, impossible.

2.2 Past Approaches to Assessing the Exclusion Restriction

We consider the model

$$y = X\beta + Z\gamma + u, \tag{2.1}$$

where y is an $n \times 1$ vector of outcomes, X is an $n \times k_X$ matrix of potentially endogenous regressors, and Z is an $n \times k_Z$ matrix of instruments of unknown validity with $k_Z \geq k_X$. We assume for now that the errors are i.i.d. (independent and identically distributed), mesokurtic, and homoskedastic:

$$u_i | z_i \sim (0, \sigma^2)$$

Recall that 2SLS regression relies on the following assumptions:

Assumption 1: $\text{rank}(E[z_i x_i']) = k_X$

Assumption 2: $E[z_i u_i] = 0$

Assumption 1 is the relevance condition and Assumption 2 is the exogeneity condition. The exogeneity condition encompasses the exclusion restriction, the focus of this paper, which states that the instruments do not contribute to y other than through the channel of their effect on X. If the instruments are valid, $\gamma = 0$ in Equation 2.1 and if the exclusion restriction fails, $\gamma \neq 0$. Due to the fact that X may be endogenous, valid instruments are needed to identify the model. However, if $\gamma \neq 0$, Z contributes to y directly, violating the exclusion restriction. The model is then underidentified. When $\gamma = 0$, we can remove Z as regressors and instead use Z as valid instruments in a 2SLS regression of the following model:

$$y = X\beta + u$$

We are especially interested in the just-identified case with one endogenous regressor

$$y = x\beta + z\gamma + u$$

because the standard Sargan test cannot provide any insight in this scenario and so alternative methods are required to gain information about the validity of the single instrument.

2.2.1 Sargan Test for Overidentifying Restrictions

As introduced in the previous chapter, the Sargan test for overidentifying restrictions can be written as follows (Sargan, 1958):

$$S(\hat{\beta}_{2SLS}) = \frac{(y - X\hat{\beta}_{2SLS})' P_Z (y - X\hat{\beta}_{2SLS})}{(y - X\hat{\beta}_{2SLS})'(y - X\hat{\beta}_{2SLS})/(n - k_X)}$$

It is a test for overidentifying restrictions because it can only test the exclusion restriction for a $k_Z - k_X$ subset of instruments in an overidentified setting. Thus the Sargan test requires at least $k_X + 1$ instruments and is only a test of $\gamma = 0$ in the equation $y = X\beta + Z_o\gamma + u$, where Z_o are a set of $k_Z - k_X$ overidentifying instruments, rather than a test of γ in $y = X\beta + Z\gamma + u$, where Z contains all instruments. If there are at least k_X valid and sufficiently strong instruments, an untestable assumption, then the test will be able to identify whether or not the additional instruments are valid. However, as there is no guarantee that the requisite number of instruments are in fact valid, it is difficult to interpret the results of the test. A failure to reject the null cannot be interpreted as proof that all instruments are valid; the same result could arise from a situation where all instruments are invalid.

To see why this is, consider that the Sargan statistic is equivalent to a minimum distance statistic for the IV estimates from individual instruments (Windmeijer, 2019). These IV estimates can be written as $\hat{\beta}_j = (z_j'x)^{-1}z_j'y$ for the case of one endogenous variable. When the distance between estimates becomes large - i.e. estimates using different instruments are different - the test rejects the null and the implication is that at least one instrument is invalid. However, even if all instruments are invalid, they may have biases in the same direction and so the minimum distance would be small and the Sargan test would fail to reject. If it is possible to predict the direction of the bias, a researcher can try to choose instruments with biases in opposite directions to avoid a failure to reject due only to invalid instruments with similar biases.

2.2.2 Falsification Adaptive Set

If the exclusion restriction does not hold, the model specification is incorrect and should be adjusted. If it was known that an instrument contributed independently to the outcome variable, over and above its relationship with the endogenous treatment variable, then it should be included as an additional regressor in the model, which would alter the estimate of β. The idea of the falsification adaptive set is to consider the possibility that, in the case of a single endogenous regressor, there is only one instrument that truly satisfies the exclusion restriction out of multiple possibilities. As we cannot test the exclusion restriction, we do not know which if any of the instruments is valid. Thus we calculate multiple estimates for β assuming that each instrument is the only valid one in turn. This method considers the worst case scenario that still satisfies the necessary requirement for IV inference in a single endogenous variable setting: the existence of one valid instrument.

The procedure for calculating the falsification adaptive set is to calculate $\hat{\beta}_j$, the IV estimate using only instrument z_j and inserting all other instruments into the regression as controls, for each of the instruments $z_1, ..., z_{k_Z}$. They define

$$\hat{\mathcal{J}}_{rel} = \{j \in \{1, ..., k_Z\} : F_j > 10\},$$

which is the set of the indices of the instruments that have first stage F statistics greater than 10 and are therefore considered sufficiently relevant. Then the falsification adaptive set is estimated as

$$\hat{FAS} = [\min_{j \in \hat{\mathcal{J}}_{rel}} \hat{\beta}_j, \max_{j \in \hat{\mathcal{J}}_{rel}} \hat{\beta}_j]$$

It is thus a report of the range of estimates for β under any IV model specification that satisfies the crucial constraint of including at least one valid instrument, the same requirement as the Sargan test, because these are the estimates that are non-falsifiable given that we cannot evaluate the exclusion restriction in the case of a single instrument. There is, however, no guarantee that at least one instrument is valid and so we next consider the case where none of the instruments are valid.

2.2.3 Plausibly Exogenous

2.2.3.1 Method

Conley et al. (2012) propose a method, refined by van Kippersluis and Rietveld (2018), to adjust estimators to account for violations of the exclusion restriction. The idea is to find a way to estimate γ, the coefficient on the instruments, in Equation 2.1. Normally, it is not possible to obtain consistent estimates of β and γ in Equation 2.1 because the model is not identified when both X and Z are endogenous. We need external instruments to account for the endogeneity of X in an instrumental variables regression, but if the instruments are potential contributors to y, they belong in the model and are therefore no longer external. Thus, we need additional information to identify the model.

In the original conception of the plausibly exogenous method by Conley et al. (2012), one simply assumes information about γ, such as its bounds or its distribution, and uses the assumptions to calculate adjusted estimates for β. If one were to assume only the bounds of γ, confidence intervals for β could be constructed for each value γ_0 within the bounds by estimating by 2SLS the equation

$$y - Z\gamma_0 = X\beta + u.$$

Then the confidence interval for β would be the union of all the confidence intervals assuming each γ_0 within the bounds. Alternatively, one could assume

$$\gamma \sim N(\mu_\gamma, \Omega_\gamma).$$

In that case,

$$\hat{\beta} \overset{approx}{\sim} N(\beta + A\mu_\gamma, V_{2SLS} + A\Omega_\gamma A'),$$

where V_{2SLS} is the standard two stage least squares (2SLS) asymptotic variance and

$$A = (X'P_Z X)^{-1} X'Z.$$

with $P_Z = Z(Z'Z)^{-1}Z'$. This distribution follows from the following logic:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= (X'P_Z X)^{-1} X'P_Z y \\
&= (X'P_Z X)^{-1} X'P_Z (X\beta + Z\gamma + u) \\
&= \beta + (X'P_Z X)^{-1} X'Z\gamma + (X'P_Z X)^{-1} X'P_Z u
\end{aligned}$$

and so

$$\begin{aligned}
\text{plim}[\hat{\beta}_{2SLS}] &= \text{plim}[\beta + (X'P_Z X)^{-1} X'Z\gamma + (X'P_Z X)^{-1} X'P_Z u] \\
&= \beta + \text{plim}[(X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z\gamma] \\
&= \beta + (E(x_i'z_i)E((z_i'z_i)^{-1})E(z_i'x_i))^{-1} E(x_i'z_i)\mu_\gamma \\
&= \beta + E(A)\mu_\gamma
\end{aligned}$$

Thus the 2SLS estimate is consistent when $\mu_\gamma = 0$, but should be adjusted when $\mu_\gamma \neq 0$ by subtracting $A\mu_\gamma$. Furthermore,

$$
\begin{aligned}
Var(\hat{\beta}) &= Var(\beta + (X'P_ZX)^{-1}X'Z\gamma + (X'P_ZX)^{-1}X'P_Zu) \\
&= V_{2SLS} + Var(A\gamma) \\
&= V_{2SLS} + A\Omega_\gamma A'.
\end{aligned}
$$

2.2.3.2 Zero First Stage

While the plausibly exogenous method offered by Conley et al. (2012) describes the theoretical adjustments that should be made to correct for exclusion restriction violations, it relies on information that is not known or estimated but rather assumed, reducing its usefulness to applied researches. Van Kippersluis and Rietveld (2018) improve upon this method by providing a way to estimate γ. Their insight is that one can use a subgroup for which Π_X, the coefficient on the instruments in the first stage regression

$$
X = Z\Pi_X + V_X, \tag{2.2}
$$

is equal to zero in order to estimate γ. It is thus referred to as the zero first stage method. An example discussed in van Kippersluis and Rietveld (2018), inspired by the earlier work of Angrist et al. (2010), seeks to determine whether having more children reduces a mother's employment. Whether or not the first two children are boys can be used as an instrument for the total number of children because parents with two children of one sex are more likely to have a third child, in hopes that it will be the other sex, than parents who already have one boy and one girl. The zero first stage group is mothers who come from countries with a strong preference for sons because when these women have two sons, they are less

motivated to try for a daughter.

If for a subgroup of the sample, Π_X is zero by construction, then

$$\begin{aligned} y &= X\beta + Z\gamma + u \\ &= Z\Pi_X\beta + Z\gamma + u + V_X\beta \\ &= Z\gamma + u + V_X\beta. \end{aligned}$$

Because $E(u + V_X\beta) = 0$, one can regress y on Z in the subsample to obtain an estimate $\hat{\gamma}_{PE}$, which can be used as μ_γ in the plausibly exogenous method to obtain a new estimate for β:

$$\hat{\beta}_{PE} = \hat{\beta}_{2SLS} - A\hat{\gamma}_{PE}$$

While this method provides point estimates for β, it does not provide an obvious way to conduct inference because it does not include a variance estimator. Van Kippersluis and Rietveld (2018) suggest a potential solution based on the guideline that covariates should be similar enough in treatment and control groups that their normalised difference is below 0.25 (Imbens and Rubin, 2015). This can be adapted to the plausibly exogenous setting by assuming the normalized difference in $\hat{\gamma}_{PE}$ between the zero first stage subsample and the rest of the sample is limited to 0.25 in 95% of cases. Given this assumption, van Kippersluis and Rietveld (2018) calculate the variance as

$$\Omega_\gamma = \left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2,$$

where S_0^2 is the standard error of $\hat{\gamma}$ in the zero first stage subsample and S_{-0}^2 is the standard error of $\hat{\gamma}$ in the rest of the sample.

2.3 Kinky Least Squares

2.3.1 KLS Estimator

The plausibly exogenous method accounts for violations of the exclusion restriction by adjusting 2SLS estimators to offset the assumed or estimated endogeneity of the instruments. In contrast, the Kinky Least Squares (KLS) method discards the instruments entirely and adjusts the OLS estimator to account for the endogeneity of the treatment variable directly (Kiviet, 2020). This is done by assuming a value for the correlation between the treatment variable and the error term, ρ_{xu}. Because this value is not known in practice, the KLS estimator is typically calculated across a range of potential values of ρ_{xu}.

In the simplest and most popular case of one endogenous treatment variable,

$$y = x\beta + u,$$

the KLS estimator at a particular value of ρ_{xu} is calculated as

$$\tilde{\beta}\left(\rho_{xu}\right) = \hat{\beta} - \tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu}/\hat{\sigma}_x, \tag{2.3}$$

where $\hat{\beta}$ is the OLS estimator and $\hat{\sigma}_x^2$ is the variance of x. Note that due to the bias introduced by the endogeneity of x, we cannot use the variance of the error term, $\hat{\sigma}_u$, directly but must adjust it according to the degree of endogeneity. Thus Equation 2.3 includes the term $\tilde{\sigma}_u\left(\rho_{xu}\right)$, an adjusted estimate of the standard error of u that depends on ρ_{xu}:

$$\tilde{\sigma}_u\left(\rho_{xu}\right) = \hat{\sigma}_u/(1-\rho_{xu}^2)$$

Recall that ρ_{xu} is the correlation between x and the error:

$$\rho_{xu} = \frac{x'u/(n-1)}{\hat{\sigma}_x \hat{\sigma}_u}$$

Under regularity conditions, including i.i.d. and mesokurtic observations, the limiting distribution of the KLS estimator is normal:

$$n^{1/2}(\tilde{\beta}\,(\rho_{xu}) - \beta) \to N(0, (\sigma_u^2/\sigma_x^2))$$

In practice, ρ_{xu} is unknown and so the researcher must adopt an assumed value, r_{xu}, in order to calculate the KLS estimate $\tilde{\beta}\,(r_{xu})$. The researcher can assume a series of values over a plausible range for the endogeneity, which may be $(-1, 1)$.

2.3.2 Testing the Exclusion Restriction

In addition to introducing the KLS estimator, Kiviet (2020) proposes that one may conduct a test of the exclusion restriction, which is traditionally thought to be impossible. Given that the KLS estimator is calculated by assuming a value for ρ_{xu}, the idea is to test $H_0 : \gamma = 0$ for each value of ρ_{xu} within a range, in the model with one instrument:

$$y = x\beta + z\gamma + u \tag{2.4}$$

Taking $X^* = [x\ z]$, we define

$$\Sigma = E(X_i^* X_i^{*'}) = \begin{pmatrix} \sigma_x^2 & \sigma_{xz} \\ \sigma_{xz} & \sigma_z^2 \end{pmatrix}$$

and

$$\Sigma^{-1} = \begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{12} & \sigma^{22} \end{pmatrix}.$$

Then

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xz} \\ \hat{\sigma}_{xz} & \hat{\sigma}_z^2 \end{pmatrix}$$

and

$$\hat{\Sigma}^{-1} = \begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xz} \\ \hat{\sigma}_{xz} & \hat{\sigma}_z^2 \end{pmatrix}^{-1} = \begin{pmatrix} \hat{\sigma}^{11} & \hat{\sigma}^{12} \\ \hat{\sigma}^{12} & \hat{\sigma}^{22} \end{pmatrix}.$$

In this model, γ is estimated by

$$\tilde{\gamma}\left(\rho_{xu}\right) = \hat{\gamma} - \hat{\sigma}_x \hat{\sigma}^{12} \tilde{\sigma}_u\left(\rho_{xu}\right) \rho_{xu},$$

where $\tilde{\gamma}\left(\rho_{xu}\right)$ is the KLS estimator for γ, which is dependent on ρ_{xu}, and $\hat{\gamma}$ is the OLS estimator for γ. Taking $\hat{\sigma}_u^2$ as the OLS variance estimator, $\tilde{\sigma}_u\left(\rho_{xu}\right) = \sqrt{\frac{\hat{\sigma}_u^2}{1-\rho_{xu}^2 \hat{\sigma}_x^2 \hat{\sigma}^{11}}}$ is the bias-corrected KLS variance estimator. If we wish to attempt the impossible - testing the exclusion restriction - we need a test statistic for the null $H_0 : \gamma = 0$. Kiviet (2020) suggests

$$W\left(\rho_{xu}\right) = \tilde{\gamma}^2 / (n^{-1} \tilde{\sigma}_u\left(\rho_{xu}\right) \tilde{V}_z) \tag{2.5}$$

where

$$\tilde{V}_z = \hat{\sigma}^{22} - \rho_{xu}^2 \left[1 + (1-\rho_{xu}^2)/(1-\rho_{xu}^2 \hat{\sigma}_x^2 \hat{\sigma}^{11})\right] \hat{\sigma}_x^2 (\hat{\sigma}^{12})^2 / (1-\rho_{xu}^2 \hat{\sigma}_x^2 \hat{\sigma}_{11}).$$

$W\left(\rho_{xu}\right)$ is asymptotically distributed as χ_1^2, although using the critical value from $F_{1,n-3}$ may improve performance in small samples. Because the true value ρ_{xu}

is not known, we use an assumed value r_{xu} when computing KLS calculations. Kiviet (2020) recommends testing $H_0 : \gamma = 0$ with $W(r_{xu})$ for every r_{xu} within a range, which may be $(-1, 1)$ or any range thought appropriate by the researcher.

For our simulations, we use a sample size $n = 500$, significance level $\alpha = 0.05$, and a grid of r_{xu} values from -0.99 to 0.99 by steps of 0.01. We generate random errors, regressors, and instruments (u, x, and z, respectively) from the multivariate normal distribution with mean zero and covariance Σ_{uxz}:

$$\Sigma_{uxz} = \begin{pmatrix} 1 & \rho_{xu} & \rho_{zu} \\ \rho_{xu} & 1 & \rho_{xz} \\ \rho_{zu} & \rho_{xz} & 1 \end{pmatrix}$$

We then generate y according to the equation $y_i = x_i\beta + u_i$, where β is a parameter value we set to 0 for these simulations. In later simulations, we introduce heteroskedasticity by generating y as $y_i = x_i\beta + u_i e^{0.7z_i}$. At each point in the r_{xu} grid for which $1 - \rho_{xu}^2\hat{\sigma}_x^2\hat{\sigma}^{11} > 0$, we calculate the test statistic W, as defined in Equation 2.5, and plug it into the chi-squared cumulative distribution function to determine the p-value. The result is a graph of p-values for the test of the exclusion restriction across the range of ρ_{xu} (e.g. Figure 2.1).

Kiviet (2020) suggests that the regions of the graph with high p-values, i.e. where one fails to reject the null, indicate the levels of endogeneity for which the exclusion restriction is valid. If the researcher believes the indicated range of endogeneity to be plausible, then it is considered acceptable to continue with IV estimation, under the assumption of valid instruments.

2.3.3 Failure of the Exclusion Restriction Test

As Kiviet (2022) points out, the graph of p-values produced by the KLS test of exclusion restrictions is misleading. Consider the following:

$$\begin{aligned}
\tilde{\gamma}\left(r_{xu}\right) &= \hat{\gamma} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(r_{xu}\right)r_{xu} \\
&= \hat{\gamma} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu} + \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(r_{xu}\right)r_{xu} \\
&= \hat{\gamma} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu} + \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu}\left(1 - \frac{\hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(r_{xu}\right)r_{xu}}{\hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu}}\right) \\
&= \hat{\gamma} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu} + \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu}\left(1 - \left(\frac{1-\rho_{xu}^2\hat{\sigma}_x^2\hat{\sigma}^{11}}{1-r_{xu}^2\hat{\sigma}_x^2\hat{\sigma}^{11}}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}}\right)
\end{aligned}$$

The KLS estimator $\tilde{\gamma}\left(\rho_{xu}\right) = \hat{\gamma} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(\rho_{xu}\right)\rho_{xu}$ converges to γ and so

$$\tilde{\gamma}\left(r_{xu}\right) \xrightarrow{p} \gamma + \sigma_x\sigma^{12}\sigma_u\rho_{xu}\left(1 - \left(\frac{1-\rho_{xu}^2\sigma_x^2\sigma_{11}}{1-r_{xu}^2\sigma_x^2\sigma_{11}}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}}\right). \tag{2.6}$$

Herein lies the downfall of the exclusion restriction test: it is clear that when we choose r_{xu} such that

$$\gamma = -\sigma_x\sigma^{12}\sigma_u\rho_{xu}\left(1 - \left(\frac{1-\rho_{xu}^2\sigma_x^2\sigma^{11}}{1-r_{xu}^2\sigma_x^2\sigma^{11}}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}}\right) \tag{2.7}$$

then $\tilde{\gamma}\left(r_{xu}\right) \xrightarrow{p} 0$ and so when we attempt to test the exclusion restriction, we fail to reject the null $H_0 : \gamma = 0$, even though the true value of γ is the value given above in Equation 2.7, which may well be different than 0. Hence, we do not have a true test of the exclusion restriction.

Result 1 Consider $\tilde{\gamma}\left(r_{xu}\right) = \hat{\gamma} - \hat{\sigma}_x\hat{\sigma}^{12}\tilde{\sigma}_u\left(r_{xu}\right)r_{xu}$, the KLS estimator for γ in the equation $y = x\beta + z\gamma + u$. There always exists an $r_{xu}^* \in [-1, 1]$ such that $\tilde{\gamma}\left(r_{xu}^*\right) \xrightarrow{p} 0$.

Proof. See Appendix F in Kiviet (2022).

To illustrate, Figure 2.1 provides two examples of output graphs a researcher might encounter if the true parameters were $\rho_{xu} = 0.3$, $\rho_{xz} = 0.5$, and $\rho_{zu} = 0.2$. We simulate this situation using the same data generating process (DGP) twice and get results that imply different conclusions, although the underlying information is identical.

The first graph in Figure 2.1 demonstrates a situation in which the instruments are invalid ($\rho_{xu} = 0.2$), but the KLS exclusion restriction test still fails to reject the null $H_0 : \gamma = 0$ in the neighborhood of $r_{xu} = -0.16$, which merely reflects the value that solves Equation 2.7, without providing any useful information. Note that in particular it does not reflect the true value of the endogeneity, $\rho_{xu} = 0.3$. If this were a true test of the exclusion restriction, the test would reject at all levels of r_{xu} in this case because the instruments are in fact invalid. However, the KLS exclusion restriction test will always have an area of nonrejection around the r_{xu} value that solves Equation 2.7. Hence, it always appears as if the instruments are valid even when they are not. Kiviet (2020) suggests considering whether the implied correlation is plausible and taking that as evidence for or against the validity of the instruments. However, this is not a foolproof strategy. The second graph in Figure 2.1 represents a different run of the test with the same parameter values as used in the first graph. The second time, the implied correlation appears to be approximately 0.2 instead of -0.16 as in the first run. Given that the true correlation is $\rho_{xu} = 0.3$, a researcher might well consider a correlation of 0.2 to be plausible and therefore conclude that the instruments are valid. This, however, would be an incorrect conclusion as the instruments are in fact invalid in this simulation.

Although it is possible that an individual draw may be misleading, this strat-

egy is informative on average. Figure 2.2 illustrates the likelihood that each r_{xu} value is rejected, calculated as a rejection frequency over 10,000 runs of the exclusion restriction test. Note that whereas the graphs in Figure 2.1 represent a single draw from the DGP to simulate what a researcher would observe, the graphs in Figure 2.2 summarise the results of many draws from the DGP to demonstrate the average performance of the test. We see that when the instruments are valid, as in the left graph of Figure 2.2, we observe the expected rejection rate of about $\alpha = 0.05$ when $r_{xu} = \rho_{xu} = 0.3$. But when the instruments are invalid, as in the right graph with $\rho_{zu} = 0.2$, the minimum rejection rate does not occur at $r_{xu} = \rho_{xu} = 0.3$, but rather at about -0.1. Hence on average, a researcher will observe a maximum p-value at a negative r_{xu} value on any particular draw. If the researcher had reason to believe the endogeneity was in the positive direction, a test result indicating that the instruments are valid when $r_{xu} = -0.1$ would cast appropriate doubt on the validity of the instruments.

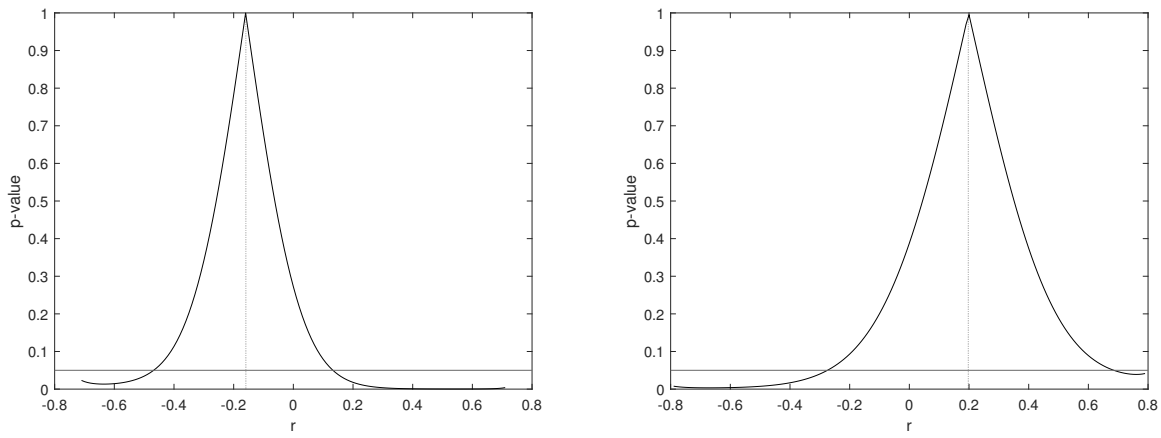
2.3.4 Irrelevant Instruments

In Equation 2.6, we saw choosing an r_{xu} that may not equal ρ_{xu} resulted in a bias:

$$\tilde{\gamma}\left(r_{xu}\right) \xrightarrow{p} \gamma + \sigma_x \sigma^{12} \sigma_u \rho_{xu} \left(1 - \left(\frac{1 - \rho_{xu}^2 \sigma_x^2 \sigma^{11}}{1 - r_{xu}^2 \sigma_x^2 \sigma^{11}}\right)^{1/2} \frac{r_{xu}}{\rho_{xu}}\right).$$

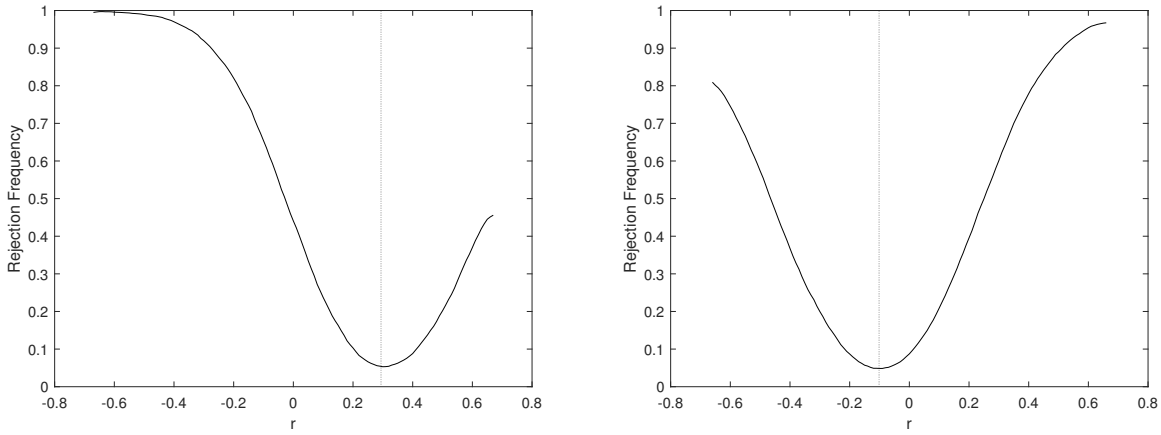
Because σ^{11} and σ^{12} are functions of ρ_{xz}, we can rewrite Equation 2.6 in terms of ρ_{xz} to see the effect of irrelevant instruments, i.e. $\rho_{xz} = 0$. We can easily calculate

Figure 2.1: P-Values from 2 Draws of the KLS Exclusion Restriction Test DGP for $\rho_{xu} = 0.3$, $\rho_{xz} = 0.5$, and $\rho_{zu} = 0.2$



Note: These are two examples of the results from the KLS exclusion test for a particular data generating process. The instruments in this case are invalid ($\rho_{zu} = 0.2$) so we would expect the test to suggest an incorrect value for the endogeneity. The left graph illustrates the expected response, with a p-value spike close to $r = -0.2$, when in fact the true value is $\rho_{xu} = 0.3$. The right graph illustrates a draw for which the test indicates an endogeneity level close to the true value ($r = 0.2, \rho_{xu} = 0.3$), which would fail to cast appropriate doubt on the validity of the instrument and is therefore a type II error.

Figure 2.2: Rejection Frequencies for 10,000 Runs of the KLS Exclusion Restriction Test for $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.5$ with $\rho_{zu} = 0$ (left) and $\rho_{zu} = 0.2$ (right)



Note: The left graph shows a minimum rejection rate of about 5% at about the true endogeneity level of 0.3, which is appropriate for the case of valid instruments. Under invalid instruments, on the right, the minimum value falls at a negative value despite the true positive endogeneity.

the inverse of the 2×2 matrix $\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xz} \\ \sigma_{xz} & \sigma_z^2 \end{bmatrix}$ to solve for σ^{11} and σ^{12}:

$$\Sigma^{-1} = \begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{12} & \sigma^{22} \end{pmatrix} = \frac{1}{\sigma_x^2\sigma_z^2 - \sigma_{xz}^2} \begin{pmatrix} \sigma_z^2 & -\sigma_{xz} \\ -\sigma_{xz} & \sigma_x^2 \end{pmatrix}$$

The definition of ρ_{xz} is

$$\rho_{xz} = \frac{\sigma_{xz}}{\sigma_x\sigma_z}$$

and so

$$\sigma_x^2\sigma_z^2 - \sigma_{xz}^2 = \sigma_x^2\sigma_z^2\left(1 - \rho_{xz}^2\right)$$

and

$$\Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_z^2\left(1 - \rho_{xz}^2\right)} \begin{pmatrix} \sigma_z^2 & -\sigma_{xz} \\ -\sigma_{xz} & \sigma_x^2 \end{pmatrix}.$$

Thus we have

$$\begin{aligned} \sigma^{11} &= \frac{1}{\sigma_x^2\left(1 - \rho_{xz}^2\right)} \\ \sigma^{12} &= -\frac{\sigma_{xz}}{\sigma_x^2\sigma_z^2\left(1 - \rho_{xz}^2\right)} \end{aligned}$$

and can rewrite the bias term in Equation 2.6 as

$$\begin{aligned} &\sigma_x\sigma^{12}\sigma_u\rho_{xu}\left(1 - \left(\frac{1 - \rho_{xu}^2\sigma_x^2\sigma^{11}}{1 - r_{xu}^2\sigma_x^2\sigma^{11}}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}}\right) \\ &= -\frac{\sigma_{xz}}{\sigma_x\sigma_z^2\left(1 - \rho_{xz}^2\right)}\sigma_u\rho_{xu}\left(1 - \left(\frac{1 - \rho_{xu}^2/\left(1 - \rho_{xz}^2\right)}{1 - r_{xu}^2/\left(1 - \rho_{xz}^2\right)}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}}\right) \\ &= \rho_{xz}\left[-\frac{1}{\sigma_z\left(1 - \rho_{xz}^2\right)}\sigma_u\rho_{xu}\left(1 - \left(\frac{1 - \rho_{xu}^2/\left(1 - \rho_{xz}^2\right)}{1 - r_{xu}^2/\left(1 - \rho_{xz}^2\right)}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}}\right)\right]. \end{aligned} \tag{2.8}$$

Given some amount of endogeneity ($\rho_{xu} \neq 0$), the above calculation demonstrates that the bias will be zero for irrelevant instruments. That is, when $\rho_{xz} = 0$,

$$\tilde{\gamma}\left(r_{xu}\right) \xrightarrow{p} \gamma$$

for any r_{xu}. This leads to the unusual result of having correct size and power when the instruments have no relationship with the endogenous variables they are meant to instrument for.

2.3.5 Weak Instruments

Recall that we can set γ equal to the negative bias to find the point at which the KLS estimator is zero. Combining Equation 2.7 and Equation 2.8 yields the following:

$$\left(\frac{1-\rho_{xu}^2/\left(1-\rho_{xz}^2\right)}{1-r_{xu}^2/\left(1-\rho_{xz}^2\right)}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}} = 1 - \frac{\gamma}{\frac{\rho_{xz}}{\sigma_z(1-\rho_{xz}^2)}\sigma_u\rho_{xu}}.$$

Because

$$\sigma^{11} = \frac{1}{\sigma_x^2\left(1-\rho_{xz}^2\right)},$$

we have

$$1/(1-\rho_{xz}^2) = \sigma_x^2\sigma^{11}$$

and so

$$\left(\frac{1-\rho_{xu}^2\sigma_x^2\sigma^{11}}{1-r_{xu}^2\sigma_x^2\sigma^{11}}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}} = 1 - \frac{\gamma}{\frac{\rho_{xz}}{\sigma_z(1-\rho_{xz}^2)}\sigma_u\rho_{xu}}$$

or equivalently

$$1-\left(\frac{1-\rho_{xu}^2\sigma_x^2\sigma^{11}}{1-r_{xu}^2\sigma_x^2\sigma^{11}}\right)^{1/2}\frac{r_{xu}}{\rho_{xu}} = \frac{\gamma}{\frac{\rho_{xz}}{\sigma_z(1-\rho_{xz}^2)}\sigma_u\rho_{xu}}.$$

Now consider the KLS estimate for β:

$$\tilde{\beta} = \hat{\beta} - \hat{\sigma}_x \hat{\sigma}^{11} \tilde{\sigma}_u \left(\rho_{xu}\right) \rho_{xu}$$

By the same arguments as before,

$$\hat{\beta} - \hat{\sigma}_x \hat{\sigma}^{11} \tilde{\sigma}_u \left(r_{xu}\right) r_{xu} \xrightarrow{p} \beta + \sigma_x \sigma^{11} \sigma_u \rho_{xu} \left(1 - \left(\frac{1 - \rho_{xu}^2 \sigma_x^2 \sigma^{11}}{1 - r_{xu}^2 \sigma_x^2 \sigma^{11}}\right)^{1/2} \frac{r_{xu}}{\rho_{xu}}\right)$$

If we choose the value r_{xu}^* that sets the KLS estimate of γ to zero as in

$$1 - \left(\frac{1 - \rho_{xu}^2 \sigma_x^2 \sigma^{11}}{1 - r_{xu}^{*2} \sigma_x^2 \sigma^{11}}\right)^{1/2} \frac{r_{xu}^*}{\rho_{xu}} = \frac{\gamma}{\frac{\rho_{xz}}{\sigma_z (1 - \rho_{xz}^2)} \sigma_u \rho_{xu}},$$

then

$$\hat{\beta} - \hat{\sigma}_x \hat{\sigma}^{11} \tilde{\sigma}_u \left(r_{xu}^*\right) r_{xu}^* \xrightarrow{p} \beta + \frac{\gamma \sigma_x \sigma^{11} \sigma_z \left(1 - \rho_{xz}^2\right)}{\rho_{xz}}.$$

As

$$\sigma^{11} \left(1 - \rho_{xz}^2\right) = 1/\sigma_x^2,$$

we have

$$\begin{aligned} \beta + \frac{\gamma \sigma_x \sigma^{11} \sigma_z \left(1 - \rho_{xz}^2\right)}{\rho_{xz}} &= \beta + \frac{\gamma \sigma_x \sigma_z}{\rho_{xz} \sigma_x^2} \\ &= \beta + \frac{\gamma \sigma_z}{\rho_{xz} \sigma_x} \\ &= \beta + \frac{\gamma}{\pi} \end{aligned}$$

where π is the coefficient on z in the first stage equation $x = z\pi + v$, which means $\sigma_{xz} = \sigma_z^2 \pi$ and so

$$\pi = \frac{\sigma_{xz}}{\sigma_z^2} = \frac{\rho_{xz} \sigma_x \sigma_z}{\sigma_z^2} = \frac{\rho_{xz} \sigma_x}{\sigma_z}.$$

Hence,

$$\hat{\beta} - \hat{\sigma}_x \hat{\sigma}^{11} \tilde{\sigma}_u \left(r^*_{xu} \right) r^*_{xu} \xrightarrow{p} \beta + \frac{\gamma}{\pi},$$

which is the IV estimand of β in Equation 2.4 because

$$\hat{\beta}_{IV} = \frac{z'y}{z'x} = \beta + \frac{z'(z\gamma + u)}{z'x}$$

and

$$\frac{z'(z\gamma + u)}{z'x} \xrightarrow{p} \frac{\sigma_z^2 \gamma}{\sigma_{xz}} = \frac{\sigma_z^2 \gamma}{\sigma_z^2 \pi} = \frac{\gamma}{\pi}.$$

A potential problem with the method of Kiviet (2022) is then that when the instruments are weak, $\pi \to 0$ and so $\beta + \frac{\gamma}{\pi} \to \infty$, which means the KLS estimator of β becomes very large as the instruments become very weak, which is not a desirable outcome.

2.3.6 Finite Sample Failure of the Exclusion Restriction Test

Kiviet (2022) exposes the key problem of the exclusion restriction test by showing that the endogeneity correlation implied by the IV estimator produces a KLS estimate of γ that tends to zero. Before we were aware of this update, we independently derived our own proof. Additionally, Kiviet (2022) only proves the result asymptotically, while we show the result to hold exactly in finite samples, which is an original contribution to the literature.

Proposition 1 Consider $\tilde{\gamma}\left(r_{xu}\right) = \hat{\gamma} - \hat{\sigma}_x \hat{\sigma}^{12} \tilde{\sigma}_u \left(r_{xu}\right) r_{xu}$, the KLS estimator for γ in the equation $y = x\beta + z\gamma + u$. Define $r_{x\hat{u}_{IV}} = \frac{\hat{\sigma}_{x\hat{u}_{IV}}}{\hat{\sigma}_x \hat{\sigma}_{\hat{u}_{IV}}}$, where $\hat{\sigma}_{AB} = \frac{A'B}{n}$, $\hat{\sigma}_A = \sqrt{\frac{A'A}{n}}$, and $\hat{u}_{IV} = y - X\hat{\beta}_{IV}$. Then $\tilde{\gamma}\left(r_{x\hat{u}_{IV}}\right) = 0$.

Proof. See Appendix 2.8.1.

In Appendix 2.8.1, we provide our proof that the KLS estimator evaluated at the endogeneity correlation implied by the IV estimator, $\tilde{\gamma}\left(r_{x\hat{u}_{IV}}\right)$, does not merely tend to zero but is exactly zero. Thus we cannot use the KLS-based test as a true test of the exclusion restriction because we have proved there will always be a point where the test fails to reject the validity of the instruments, regardless of their actual validity or invalidity.

In order for this finite sample result to be precise, we disregard degrees of freedom corrections that would otherwise introduce a small amount of mismatch (e.g. n versus $n-1$). This has no practical relevance because the KLS exclusion restriction test will also fail to reject the null that $\gamma = 0$ if $\tilde{\gamma}$ is almost exactly rather than exactly zero.

2.3.7 KLS as Sensitivity Analysis

Although we have shown that the KLS test of the exclusion restriction is misleading, KLS is still a valid method of estimation. Recall that there is a distinction between estimating β without instruments according to the model $y = x\beta + u$ and attempting to apply the logic of this estimation method to test the value of γ in the model $y = x\beta + z\gamma + u$. The former is perfectly acceptable, while the latter is more objectionable. But because the KLS method of estimation is based on unverifiable assumptions about the population parameter ρ_{xu}, it is best used as a sensitivity analysis rather than a method yielding a single point estimate.

Examining the KLS estimates of β across assumed correlation values of $(-1, 1)$, however, may produce an unhelpfully large range in β estimates. Applied researchers sometimes provide the KLS estimates $\tilde{\beta}$ in this way, without imposing assumptions on ρ_{xu}, and as a result their estimates are robust but imprecise (e.g. Mariella, 2021; Kiviet and Kripfganz, 2020). Although the correlation between x

and u is not known, some applied researchers assume its sign or a range of plausible values in order to limit the range of the resulting KLS estimates, which are less robust but more precise than if they were to forgo assumptions (e.g. Wang and Cheng, 2022; Chen, 2022).

Another use for the KLS estimation technique is to calculate the value of ρ_{xu} for which $\tilde{\beta}$ loses significance. Briel et al. (2020), for example, use a confidence index to predict starting salaries and calculate the level of endogeneity for which the confidence index loses significance in both a male and female model (see Figure 2.3). They calculate that $\tilde{\beta}$ loses significance at $\rho_{xu} = 0.72$ in the male model and $\rho_{xu} = 0.93$ in the female model, which are high enough levels of correlation to suggest that confidence is indeed a significant contributor to starting salaries, especially for women.

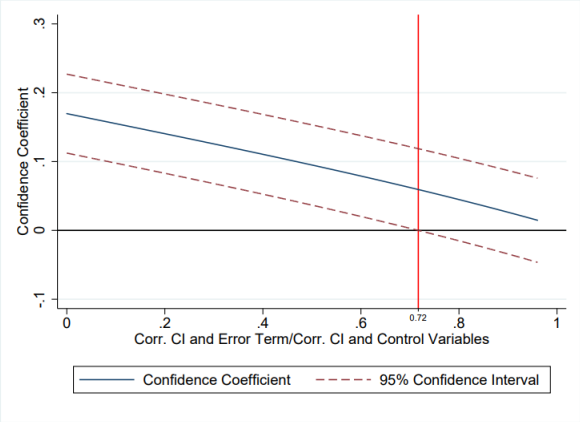
It is important to note that KLS sensitivity analysis graphs will not always look like those in Figure 2.3. It is possible to have a positively sloping line, for example, or to have a situation in which the line approximately passes through the origin, in which case we could calculate the correlation values for which the treatment variable gains rather than loses significance.

2.4 Tests of ρ_{xu}

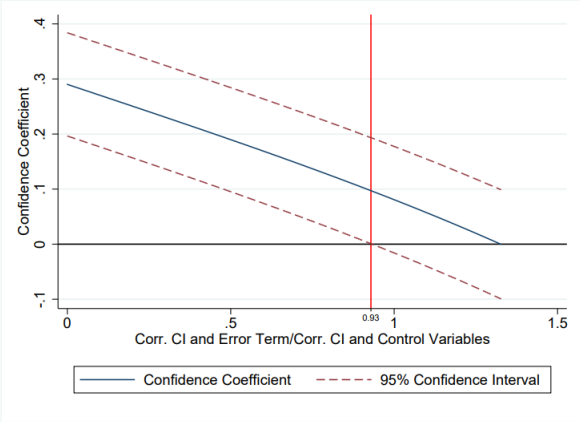
While the purported KLS test of the exclusion restriction does not properly function as a true exclusion restriction test, it does provide some information and can partially illuminate the link between instrument validity and treatment variable endogeneity. The framing as a test of the exclusion restriction is misleading, however, and framing it as a test of the treatment variable endogeneity correlation, ρ_{xu}, may be more straightforward, although neither concept can be tested

Figure 2.3: KLS Loss of Significance in Briel et al. (2020): "Sensitivity of Confidence to Correlation with the Error Term"

(a) Male Model



(b) Female Model



Note: KLS estimation can be used as a sensitivity analysis. Here we see the minimum value of the endogeneity needed for the coefficient to lose significance. As the values are quite high (0.72 in the male model and 0.93 in the female model), this provides assurance that the coefficient is truly significant.

individually - they are necessarily intertwined.

If we reformat the test of the exclusion restriction as a test of the endogeneity correlation of the treatment variable, the null is $H_0 : \rho_{xu} = r_{xu}$ rather than $H_0 : \gamma = 0$. This test will identify the true level of endogeneity when the exclusion restriction is satisfied. So if we have valid instruments, the r_{xu} value corresponding to the highest point on a p-value graph as in Figure 2.1 is a consistent estimate of ρ_{xu}. It is important to recognise that the two tests - of $H_0 : \rho_{xu} = r_{xu}$ and of $H_0 : \gamma = 0$ - are identical. However, the interpretation is somewhat altered. Whereas a test of $H_0 : \gamma = 0$ is misleading because it always indicates a degree of treatment variable endogeneity for which the instruments appear valid, even if they are not, the spike in p-values is more easily understood in the case of testing $H_0 : \rho_{xu} = r_{xu}$, where it represents an estimate for the endogeneity correlation, which is valid if the instruments are valid.

2.4.1 Bootstrap

In order to conduct a test of ρ_{xu} one may follow the procedure already described for testing $H_0 : \gamma = 0$. Namely, use the sample data to calculate the KLS-based test statistic directly at each assumed endogeneity r_{xu} within the range. However, it is also possible to bootstrap the sample and calculate a different estimate of ρ on each iteration. Recall that a bootstrap repeatedly resamples with replacement from the original sample data, reshuffling the data in different combinations. For each of the $B = 599$ iterations of the bootstrap, we create a set $\{x_b^*, y_b^*, z_b^*\}$ by jointly resampling from $\{x, y, z\}$ a total of n times. Thus we have the same number of observations on x, y, and z in the bootstrap as in the original sample, with some observations repeated and some left out. We then subtract the mean from each variable and calculate the following:

$$\begin{aligned}
\hat{\beta}^*_{IV_b} &= \frac{z_b^{*\prime} y_b^*}{z_b^{*\prime} x_b^*} \\
\hat{u}^*_{IV_b} &= y_b^* - x_b^* \hat{\beta}^*_{IVb} \\
\hat{\sigma}^{*2}_{u_b} &= \hat{u}^*_{IV_b}{}' \hat{u}^*_{IV_b}/(n-2) \\
\hat{\sigma}^{*2}_{x_b} &= x_b^{*\prime} x_b^*/(n-1) \\
\hat{\rho}^*_{xu_b} &= \frac{x_b^{*\prime} \hat{u}^*_{IV_b}}{\sqrt{\hat{\sigma}^{*2}_{x_b} \hat{\sigma}^{*2}_{u_b}}}
\end{aligned}$$

This procedure is repeated for every iteration, across the grid of r_{xu} values. Bootstrapping therefore produces a distribution of $\hat{\rho}_{xu}$ estimates and we propose two different methods for utilising this distribution.

Notably, jointly resampling from $\{x, y, z\}$ makes the bootstrap robust to heteroskedasticity. This is important because the method of Kiviet (2022) is only valid under homoskedasticity, so using the bootstrap allows us to relax a key

assumption.

2.4.1.1 Percentile Method

First, one could calculate the percentiles of the distribution of $\hat{\rho}_{xu}$ estimates corresponding to $\alpha/2$ and $1-\alpha/2$ (the 2.5th and 97.5th percentiles in the case of $\alpha = 0.05$). These would then serve as the bounds of the confidence interval. Thus, this method directly estimates the confidence interval of $\hat{\rho}_{xu}$ instead of calculating a p-value at each r_{xu} value, which is why it appears simply as two vertical lines demarcating the lower and upper bounds in the example p-value graph in Figure 2.4.

2.4.1.2 Wald Statistic with Bootstrapped Variance

Another possible method for using the bootstrapped $\hat{\rho}_{xu}$ distribution is to calculate its variance and then plug that value into the Wald statistic:

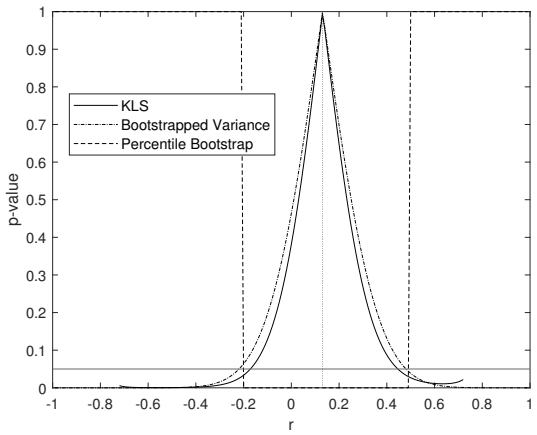
$$W = \frac{(\hat{\rho}_{xu} - r_{xu})^2}{Var(\hat{\rho}^*_{xu_b})},$$

where $Var(\hat{\rho}^*_{xu_b})$ is the variance of the bootstrapped $\hat{\rho}_{xu}$ values. The resulting W statistic is then compared to the critical value from the χ^2_1 or $F_{1,n-3}$ distribution as usual.

2.4.1.3 Performance of the Two Bootstrap Methods

As shown in Figure 2.5, the standard KLS test and the bootstrap methods perform quite similarly under moderately strong instruments and homoskedasticity. Under heteroskedasticity, all methods are at least somewhat oversized, the standard KLS test most severely and the percentile bootstrap least severely. When

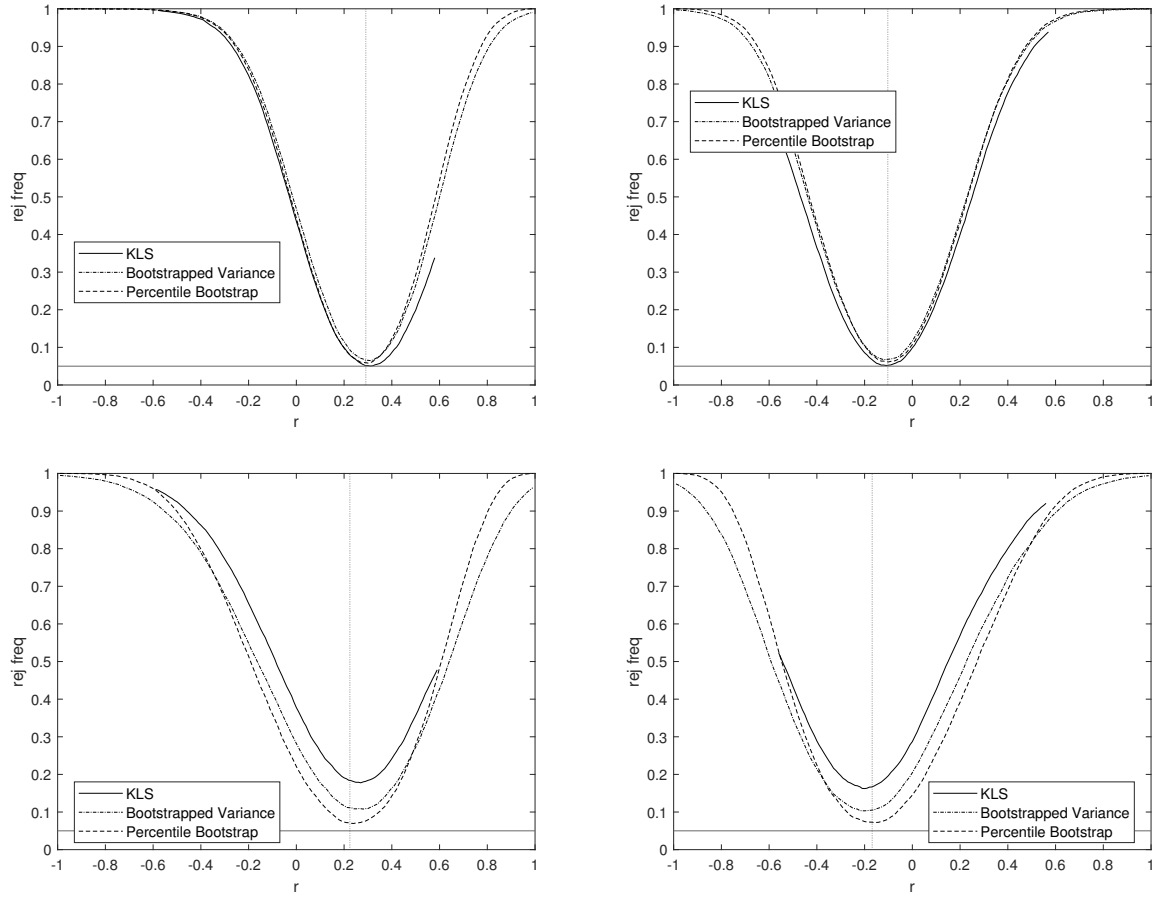
Figure 2.4: P-Values from a Single Draw of KLS Exogeneity Test DGP Compared to Two Bootstrap Methods for $\rho_{xu} = 0.3$, $\rho_{xz} = 0.5$, and $\rho_{zu} = 0$



Note: In a particular run, the standard KLS test and the KLS test with bootstrapped variance will appear as spikes, while the percentile bootstrap will appear as bounds.

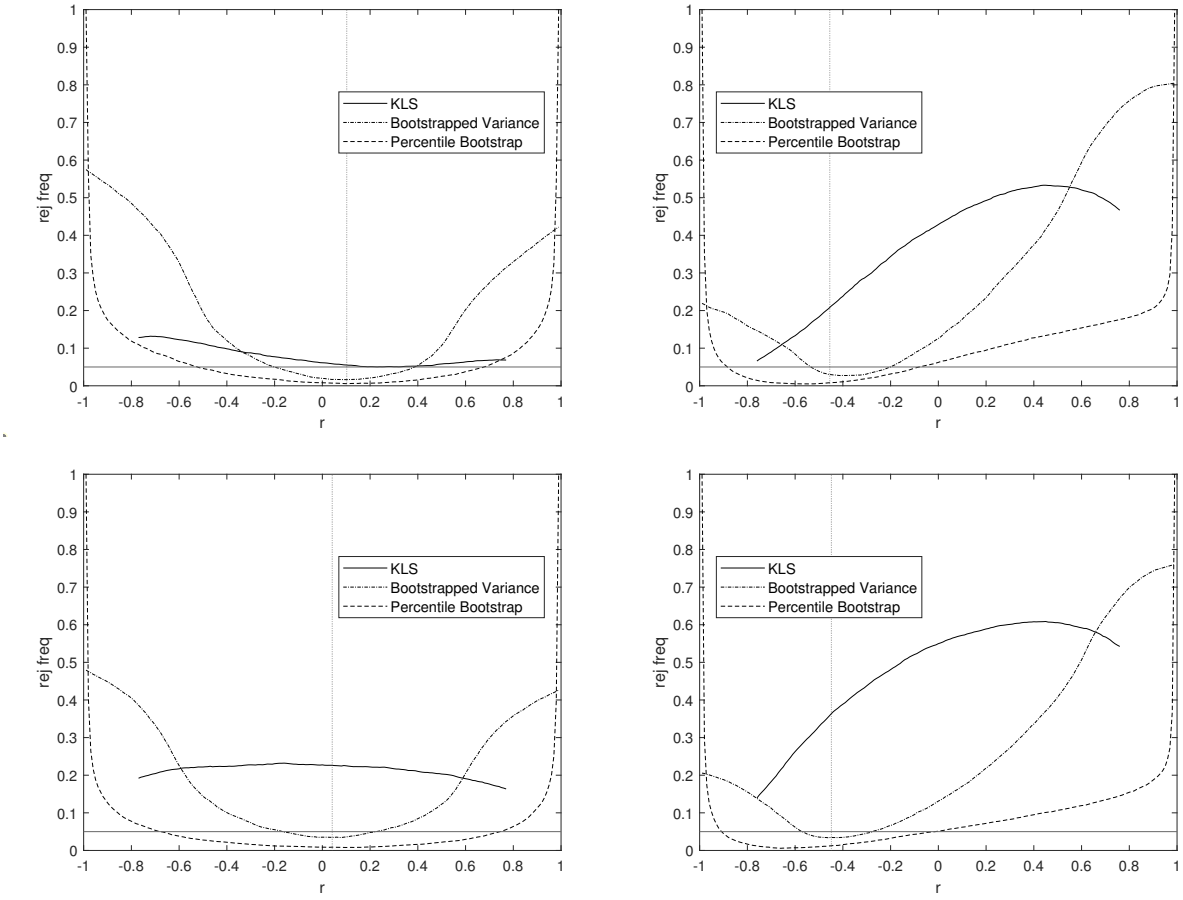
the instruments are weak, the bootstrap methods maintain the expected shape to a greater extent than the KLS test, which has downwards rather than upwards concavity in three of the four scenarios illustrated in Figure 2.6. For valid instruments, the standard KLS test has far less power than the bootstrapped variance method under homoskedasticity and is greatly oversized under heteroskedasticity. For invalid instruments, although the KLS test has a smaller range of rejection rates across r_{xu} than the bootstrapped variance method, its rejection rate at the true value $\rho_{xu} = 0.3$ is higher under both homoskedasticity and heteroskedasticity, which means that it is more likely that a researcher will correctly detect the invalidity of the instruments if testing $H_0 : \rho_{xu} = 0.3$. In all scenarios represented in Figure 2.6, the percentile bootstrap tends to have low rejection rates across most of the range of r_{xu} and very high rejection rates at the extremities (i.e. near -1 and 1), which is reasonable in the weak instrument context.

Figure 2.5: Rejection Frequencies for 10,000 Runs of the Bootstrap and KLS Endogeneity Tests for $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.5$ with $\rho_{zu} = 0$ (left) and $\rho_{zu} = 0.2$ (right), under homoskedasticity (top) and heteroskedasticity (bottom)



Note: The three test perform quite similarly on average under homoskedasticity (top), but the bootstrapped versions perform better than the standard KLS test under heteroskedasticity (bottom).

Figure 2.6: Rejection Frequencies for 10,000 Runs of the Bootstrap and KLS Endogeneity Tests for $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.1$ with $\rho_{zu} = 0$ (left) and $\rho_{zu} = 0.2$ (right), under homoskedasticity (top) and heteroskedasticity (bottom)



Note: When the instruments are quite weak, all tests perform poorly, which is to be expected. The standard KLS test displays reversed concavity in three of the four illustrated scenarios, whereas the bootstrapped versions better hold the expected shape.

2.4.2 Anderson-Rubin Test of ρ_{xu}

So far, we have recommended treating the KLS exclusion test as an endogeneity test and proposed two different bootstrap methods for such a test. Now we devise an alternative test of the endogeneity correlation based on the Anderson-Rubin (AR) statistic and compare its performance to the KLS-based test of endogeneity.

In the last chapter, we proposed a heteroskedasticity-robust form of the AR statistic,

$$ARS_r(\beta_0) = u_0' Z (Z' H_{u_0} Z)^{-1} Z' u_0,$$

which is useful here because the KLS method is not robust to heteroskedasticity. The idea of the AR test is to calculate an estimate for ρ_{xu} based on the IV estimator for β in the equation

$$y = x\beta + u$$

with instrument z and then find the upper and lower bounds of a confidence interval for ρ_{xu}.

Given an IV estimator $\hat{\beta}_{IV} = (z'x)^{-1} z'y$, we can estimate ρ_{xu} using the correlation formula:

$$r_{x\hat{u}_{IV}} = \frac{\hat{\sigma}_{x\hat{u}_{IV}}}{\hat{\sigma}_x \hat{\sigma}_{\hat{u}_{IV}}},$$

where $\hat{u}_{IV} = y - x\hat{\beta}_{IV}$. To find the upper and lower bounds, we need to find β_0 such that the AR statistic passes the threshold of the critical value in either direction. This can be done with a stepping method that starts at the IV estimator and then increases β until the point at which it crosses the critical value, which is the value $\hat{\beta}_{up}$ associated with the upper bound, and, restarting at the IV estimator, decreases β until again surpassing the critical value, resulting in the value $\hat{\beta}_{low}$

associated with the lower bound. The values $\hat{\beta}_{up}$ and $\hat{\beta}_{low}$ are then converted to a bound on the confidence interval of $\hat{\rho}_{xu}$ by the following process:

$$\hat{u}_{up} = y - x\hat{\beta}_{up}$$
$$r_{up} = \frac{\hat{\sigma}_{x\hat{u}_{up}}}{\hat{\sigma}_x \hat{\sigma}_{u_{up}}}$$
$$\hat{u}_{low} = y - x\hat{\beta}_{low}$$
$$r_{low} = \frac{\hat{\sigma}_{x\hat{u}_{low}}}{\hat{\sigma}_x \hat{\sigma}_{u_{low}}}$$

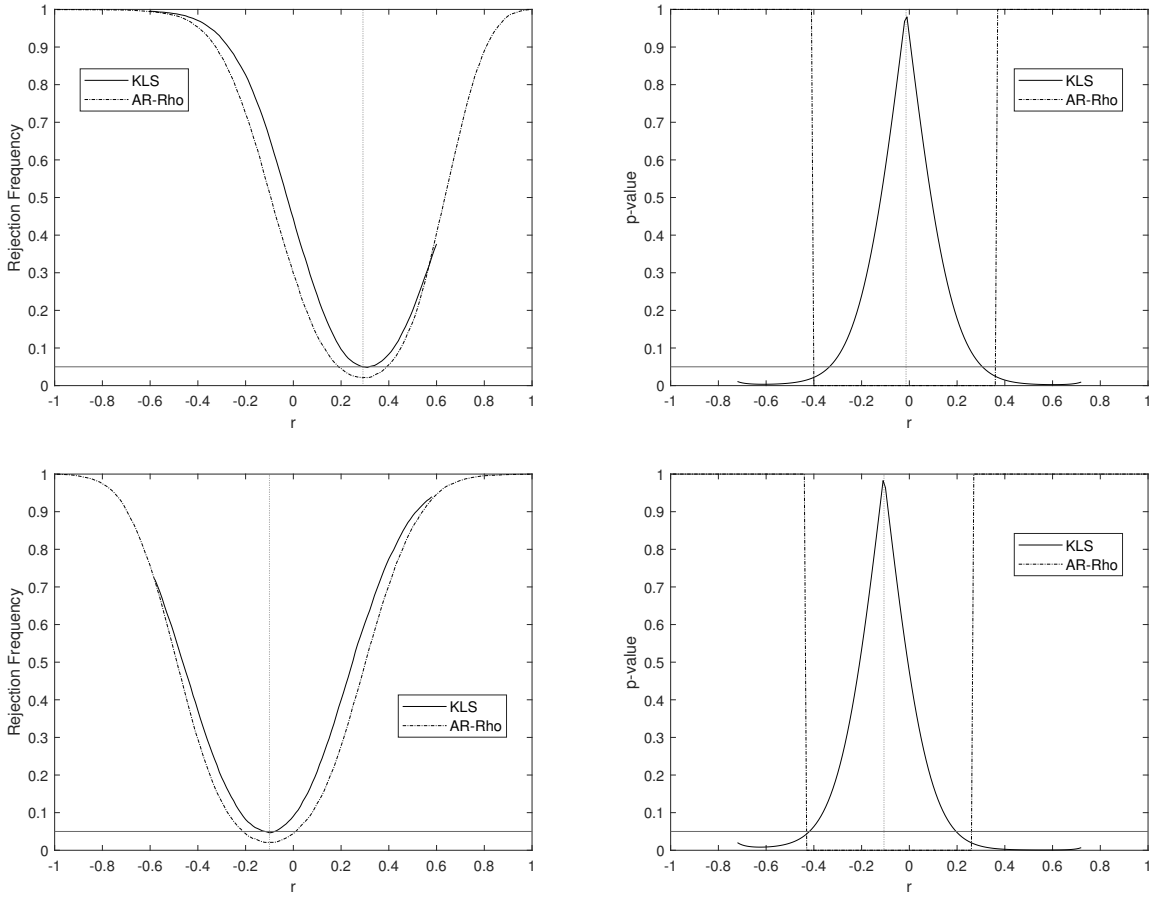
If the AR statistic associated with the IV estimator exceeds the critical value, then the confidence interval is empty. Alternatively, when the instruments are especially weak, there may be a situation in which very large values of β_0 do not produce AR statistics that exceed the critical value, in which case the confidence interval for ρ_{xu} is $(-1, 1)$ because there is simply not enough information to estimate it more precisely. Figure 2.7 shows that the results of the AR and KLS endogeneity tests are fairly similar over 10,000 runs with $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.5$. They both correctly identify the endogeneity when the instruments are valid and they both go equally wrong when the instruments are invalid. Thus, if a researcher expected a small positive value for ρ_{xu}, (s)he would come to the same conclusions about the validity of the instruments using either method. The AR test is, however, slightly undersized compared to the correctly sized KLS test and also has slightly reduced power. And while the left-hand column in Figure 2.7 demonstrates that both methods work on average, the right-hand column, which depicts the results of a single p-value draw and therefore represents what applied researchers would observe in practice, reminds us that any one particular draw can be difficult to interpret. In the case of valid instruments (top right), the

treatment variable appears to be exogenous on this draw and its true correlation, $\rho_{xu} = 0.3$, appears on the edge of plausibility. Of course, with any correctly sized test of size 0.05, a true null is rejected 5% of the time. The p-value draw for the invalid test illustrates how the test is supposed to function. If the researcher had correctly guessed the true endogeneity and tested $H_0 : \rho_{xu} = 0.3$, both the AR and KLS tests would reject and the researcher would correctly conclude that the instruments were invalid. Crucially, this depends on a good estimate for ρ_{xu}. If the researcher had tested $H_0 : \rho_{xu} = 0.1$, also a small positive correlation, (s)he would have failed to reject.

While the AR and KLS tests perform similarly under moderately strong instruments and homoskedasticity, significant differences emerge under weak instruments and heteroskedasticity. Under heteroskedasticity with moderately strong instruments, the KLS test becomes severely oversized, with minimum rejection rates over 16%. Meanwhile the AR test, which is robust to heteroskedasticity, remains undersized, as it was under homoskedasticity (Figure 2.8).

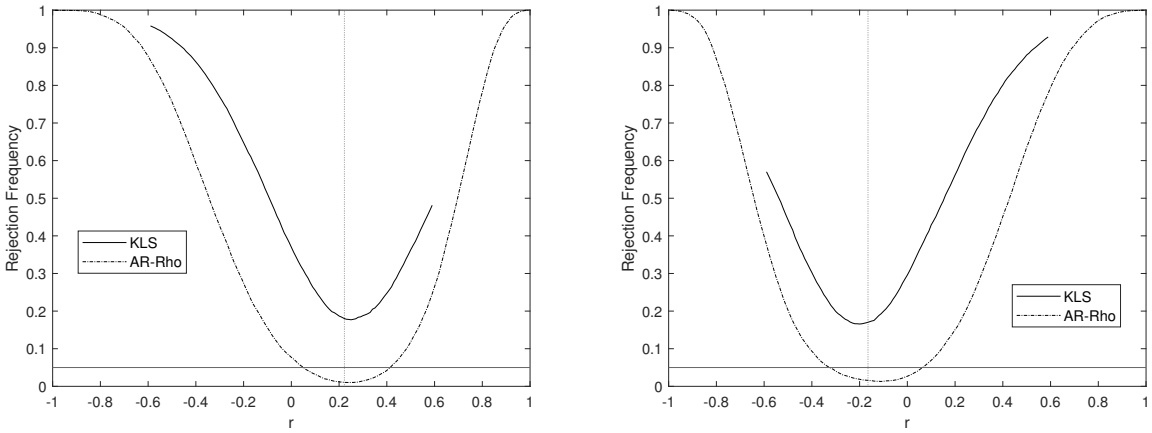
Figure 2.9 shows the rejection frequencies of the two tests under weak instruments. In the case of homoskedasticity and valid instruments (top left), the KLS test performs reasonably well, with just over a 5% rejection rate at the true value and some amount of power against negative endogeneities. Under weak instruments, a lack of power is to be expected. The AR test is more conservative, with a minimum rejection rate under 1%. Under homoskedasticity and invalid instruments (top right), both tests have lower rejection rates for large negative values of r_{xu}, which would lead a researcher who believed the endogeneity to be positive to correctly reject the validity of the instruments. However, the KLS test has much more power to reject the true value ($\rho_{xu} = 0.3$) in order to indicate the invalidity of the instruments: the rejection rate at $\rho_{xu} = 0.3$ is 52% for the KLS test and

Figure 2.7: Rejection Frequencies (left) for 10,000 Runs of the AR and KLS Endogeneity Tests and P-Values from One DGP Draw (right) for $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.5$ with $\rho_{zu} = 0$ (top) and $\rho_{zu} = 0.2$ (bottom)



Note: Both the KLS test and the AR endogeneity test function well under moderately strong instruments and homoskedasticity, although the latter is undersized on average. In a particular draw (right), the KLS test generates a spike in p-values, while the AR test simply estimates a confidence interval.

Figure 2.8: Rejection Frequencies for 10,000 Runs of the AR and KLS Endogeneity Tests for $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.5$ with $\rho_{zu} = 0$ (left) and $\rho_{zu} = 0.2$ (right), under heteroskedasticity



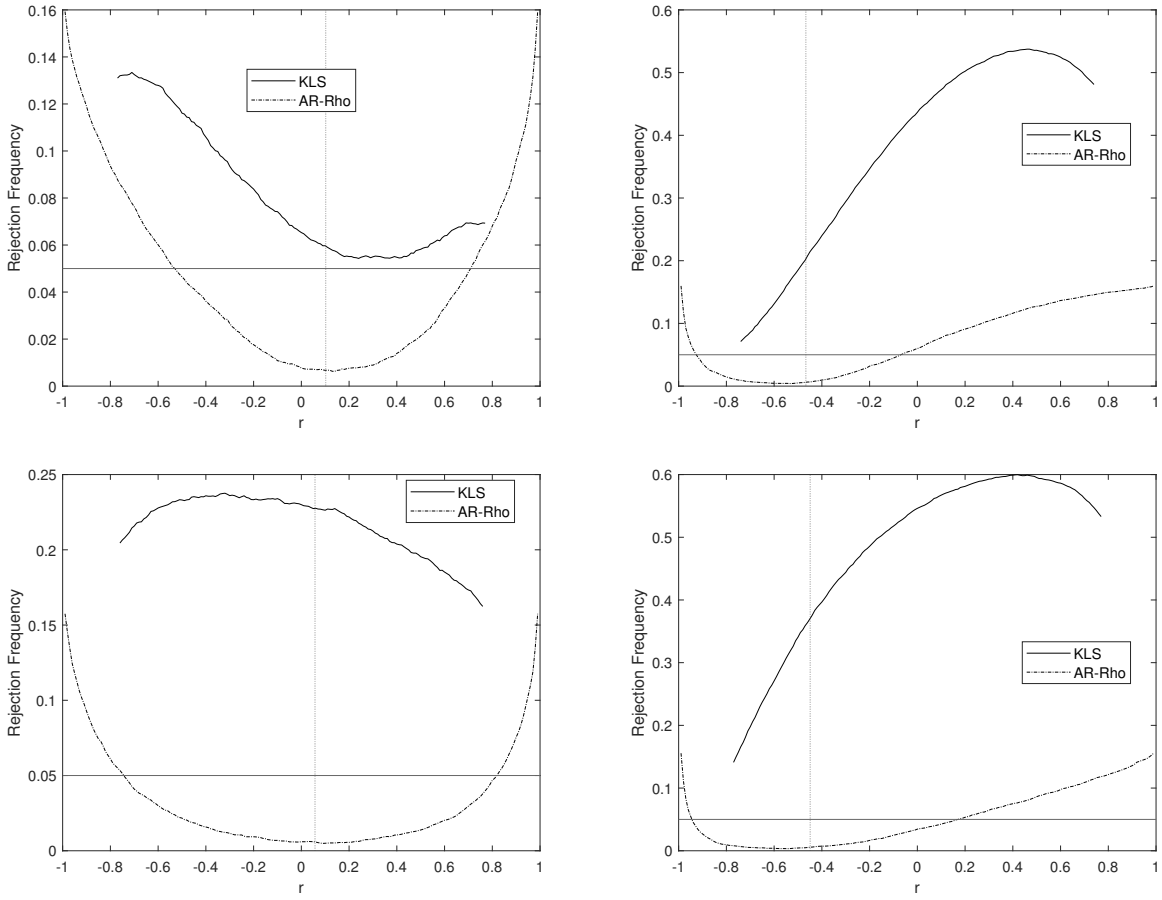
Note: Under heteroskedasticity, the KLS test is greatly oversized, while the AR test remains undersized.

only 10% for the AR test. Thus, under homoskedasticity the KLS test is preferred due to its superior power properties.

The KLS test, however, does not perform as well under heteroskedasticity as it does under homoskedasticity because the AR test uses a heteroskedasticity robust statistic while the KLS test does not. Hence under heteroskedasticity with valid instruments (bottom left), the KLS test is far oversized, with a rejection rate at the true value of 21%, rather than the 5% rejection rate a correctly sized test would have. Interestingly, the KLS test maintains high rejection rates at the true value under heteroskedasticity and invalid instruments (bottom right), as desired. It fails, however, to produce any rejection rates below 14%, whereas the AR test produces much lower rejection rates far away at large negative values.

Under weak instruments, then, the AR test is almost always less likely to reject than the KLS test. In the context of a researcher testing a hypothesised value for ρ_{xu}, a failure to reject would lead a researcher to conclude that it is plausible that the instruments are valid, which is a good conclusion if they truly are valid and a

Figure 2.9: Rejection Frequencies for 10,000 Runs of the AR and KLS Endogeneity Tests for $\rho_{xu} = 0.3$ and $\rho_{xz} = 0.1$ with $\rho_{zu} = 0$ (left) and $\rho_{zu} = 0.2$ (right), under homoskedasticity (top) and heteroskedasticity (bottom)



Note: Under weak instruments, the KLS test sometimes displays reversed concavity, while the AR test is conservative but maintains the expected shape.

bad conclusion if they are not.

2.5 Using the Zero First Stage Plausibly Exogenous Method to Estimate Endogeneity

If the researcher has a subgroup of the sample for which Π_X in Equation 2.2 is zero by construction, (s)he can use the zero first stage plausibly exogenous

method (van Kippersluis and Rietveld, 2018) to first obtain an estimate of γ (or β) and next obtain an estimate of ρ. Setting the r_{xu}-dependent KLS estimate equal to the plausibly exogenous estimate $\hat{\gamma}_{PE}$ provides a way to solve for r_{xu} and the solution is then an estimate of ρ_{xu}. In particular, the difference between the OLS and plausibly exogenous estimates of γ is equal to the bias term in the KLS formula:

$$\tilde{\gamma}\left(r_{xu}\right) = \hat{\gamma} - \tilde{\sigma}_u\left(r_{xu}\right)\hat{\sigma}_x r_{xu}\hat{\sigma}_{12} = \hat{\gamma}_{PE}$$

$$\begin{aligned}\hat{\gamma} - \hat{\gamma}_{PE} &= \tilde{\sigma}_u\left(r_{xu}\right)\hat{\sigma}_x r_{xu}\hat{\sigma}_{12} \\ &= \frac{\hat{\sigma}_u^2\hat{\sigma}_x r_{xu}\hat{\sigma}_{12}}{1 - r_{xu}^2\hat{\sigma}_x^2\hat{\sigma}_{11}}\end{aligned}$$

Alternatively, the researcher could equate the KLS and PE estimates of β rather than γ:

$$\tilde{\beta}\left(r_{xu}\right) = \hat{\beta} - \tilde{\sigma}_u\left(r_{xu}\right) r_{xu}/\hat{\sigma}_x = \hat{\beta}_{PE}$$

$$\begin{aligned}\hat{\beta} - \hat{\beta}_{PE} &= \tilde{\sigma}_u\left(r_{xu}\right) r_{xu}/\hat{\sigma}_x \\ &= \frac{\hat{\sigma}_u^2 r_{xu}}{(1 - r_{xu}^2)\hat{\sigma}_x}\end{aligned}$$

The r_{xu} value that solves these equations is then an estimate of the endogeneity of the treatment variable. This method thus provides a way to consistently estimate the endogeneity correlation even with invalid instruments. Previous literature provides guidance on how to test whether a treatment variable is endogenous (e.g. Hausman, 1978) but we go beyond testing for the existence of endogeneity

to estimate the extent of the endogeneity with invalid instruments.

2.6 Application to Catholic Schooling Data

Altonji et al. (2005) examine the effect of attending Catholic high school on educational outcomes and use the zero first stage method to evaluate the exclusion restriction of the instrument, Catholicism. The idea is that attending Catholic schools may affect educational outcomes like test scores, but the type of families who send their children to Catholic schools may be systematically different from the ones who do not and thus Catholic schooling is a potentially endogenous treatment variable. Altonji et al. (2005) attempt to evaluate the suitability of Catholicism as an instrument. It is clearly relevant - being Catholic is highly correlated with attending Catholic school, but whether or not it is exogenous is less clear. If Catholic students tend to have different educational outcomes regardless of the type of school they attend, then the instrument violates the exclusion restriction. To see if this is the case, Altonji et al. (2005) study the educational outcomes in a subsample of students, some Catholic and some not, who attended public school the year before high school. Only a very small number of the students in this subsample go on to Catholic high school. In this situation, the first stage coefficient on the instrument is zero by construction because Catholicism is not related to the type of school the students attend - they nearly all attend non-religious schools. Hence, van Kippersluis and Rietveld (2018) are able to apply their zero first stage plausibly exogenous method to estimate the coefficient on Catholic schools in the structural equation.

Here we consider 12th grade math scores as the educational outcome, with Catholic school attendance in 10th grade as the potentially endogenous treat-

ment variable ($n = 6839$). As summarised in Table 2.1, the OLS estimate of the effect of Catholic schooling on 12th grade math scores is $\hat{\beta} = 0.882$ ($SE = 0.250$, $p = 0.000$) and the 2SLS estimate is $\hat{\beta}_{IV} = 3.745$ ($SE = 0.922$, $p = 0.000$). With uncertainty introduced through the Imben's Rule method introduced in Section 2.2.3.2, the plausibly exogenous estimate is $\hat{\beta}_{PE} = 0.225$ ($SE = 0.967$, $p = 0.816$). Hence, the estimated effect of attending Catholic high school on 12th grade math scores appears highly significant in the OLS and 2SLS regressions, but loses significance when the invalidity of the instrument is accounted for, through the plausibly exogenous zero first stage method. Simply accounting for the endogeneity through 2SLS regression analysis with the Catholicism instrument greatly increases the magnitude of the coefficient estimate as compared to the estimate from OLS regression, from $\hat{\beta} = 0.882$ to $\hat{\beta}_{IV} = 3.745$, but this large and significant coefficient is misleading because the instrument is invalid. 2SLS regression assumes the instrument is exogenous, but in the subsample for which the first stage coefficient on the instrument is zero by construction, we estimate a direct effect of Catholicism on 12th grade math scores of $\hat{\gamma} = 0.554$ ($SE = 0.168$, $p = 0.001$). This suggests Catholic students receive higher math scores regardless of the type of school they attend and so the significant positive effect of attending Catholic school identified by OLS ($\hat{\beta} = 0.882$, $p = 0.000$) is actually due to the concentration of higher scoring students. Thus we do not have evidence that Catholic schools perform any better than public schools, as it seems they serve a population of students predisposed to higher scores. This is why the plausibly exogenous estimate, after accounting for the invalidity of the Catholicism instrument, is not significantly different from zero ($\hat{\beta} = 0.225$, $p = 0.816$).

We can also apply the KLS method to view the adjusted coefficient estimates across a range of assumed values for the endogeneity coefficient (Figure 2.10, left).

Table 2.1: Regression Results for the Effect of Attending Catholic School on 12th Grade Reading Scores

	OLS	2SLS	PE
$\hat{\beta}$ (SE)	0.882* (0.250)	3.745* (0.922)	0.225 (0.967)
Cragg-Donald F Statistic		584.240	

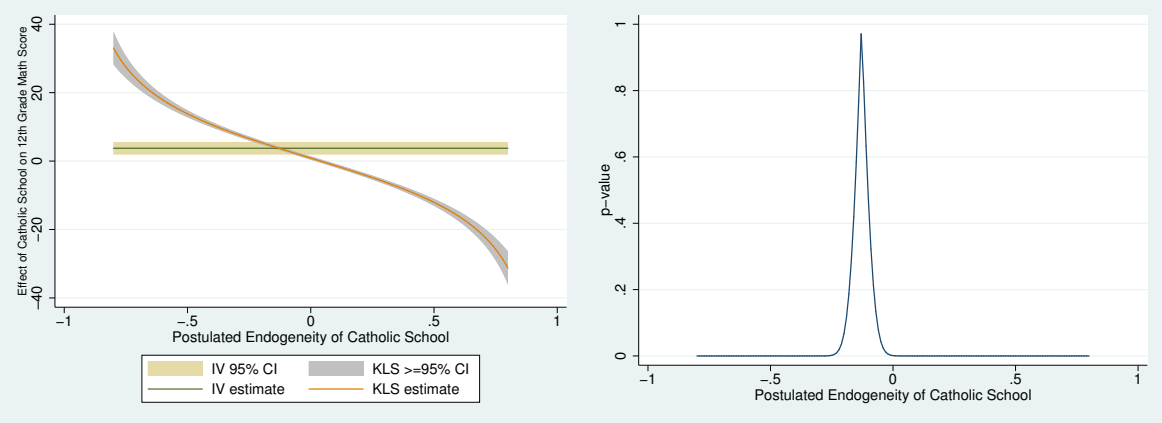
Table 2.2: Estimated Endogeneity of Attending Catholic School in 12th Grade Reading Scores Regression

	KLS	PE
$\hat{\rho}_{xu}$	-0.129	0.029

If we have a hypothesized value of ρ_{xu}, we can calculate the KLS estimate for that value (e.g. $r_{xu} = 0.2$ implies $\hat{\beta} = -3.630$). In Figure 2.10 (right), we provide the output from the KLS exclusion test for the Catholicism instrument, which shows a spike in p-values at $r_{xu} = -0.129$. A naive analysis would conclude that Catholicism is a valid instrument for attending Catholic school if $r_{xu} = -0.129$ is a plausible value for the endogeneity. However, as discussed in Section 2.3.3, the supposed test of the exclusion restriction is misleading and we can never conclude that the instrument is valid based on its results, but the negative endogeneity correlation may raise doubt about the validity of the instrument if we expected a positive correlation.

Following the proposal in Section 2.5, we can use the zero first stage plausibly exogenous estimate of β to estimate the endogeneity. We set $\tilde{\beta}(r_{xu}) = \hat{\beta}_{PE}$ to solve for r_{xu}. We know $\hat{\beta}_{PE} = 0.225$ and we calculate $\tilde{\beta}(r_{xu})$ across the range $[-0.8, 0.8]$ at intervals of 0.01, finding that the plausibly exogenous estimate falls between $\tilde{\beta}(0.02) = 0.441$ and $\tilde{\beta}(0.03) = 0.220$. Linear extrapolation then suggests that the r_{xu} value for which $\tilde{\beta}(r_{xu}) = 0.225$ is 0.029, our estimate of the endogeneity correlation of attending Catholic school.

Figure 2.10: KLS and IV Estimates of the Effect of Attending Catholic School on 12th Grade Reading Scores (left), KLS Exclusion Restriction Test for Catholicism Instrument (right)



Note: A standard IV estimate does not depend on the postulated endogeneity and so appears as a flat line in the graph on the left. The KLS estimate, in contrast, changes with the assumed endogeneity level. The KLS endogeneity test, right, shows a spike in p-values at $r_{xu} = -0.129$, which provides evidence against the exclusion restriction if the researcher's prior was that the endogeneity was positive.

2.7 Conclusion

While Kiviet (2020) claimed to be able to "test the impossible" by devising the first hypothesis test of the exclusion restriction purported to work even in the case of a single instrument, we have proved this test to be deficient not just asymptotically but in finite samples. The crucial problem is that the endogeneity correlation implied by the IV estimator always produces an exactly zero KLS estimate of the coefficient on the instruments, $\tilde{\gamma}$, provided we disregard small degrees of freedom corrections. If the test is marketed as a test of the exclusion restriction, the instruments will always appear valid at this point, which is highly misleading. Therefore, we recommend considering the test as a test of endogeneity that is valid under valid instruments. We cannot escape the fundamental problem of the failure to reject at the identified point, but we can adjust our interpretation of the test to better reflect what it can and cannot do. While it cannot properly test

the exclusion restriction, it can flag possible violations of the exclusion restriction if a test of a plausible value of the endogeneity correlation is rejected. However, a failure to reject a plausible value of the endogeneity correlation should not be taken as proof that the instruments are valid because a true test of the exclusion restriction is indeed impossible.

We offer several extensions to our reformulation of the KLS test of the exclusion restriction as a test of the endogeneity correlation. We first propose a percentile bootstrap of the endogeneity correlation, which has superior performance to the KLS test at least when instruments are moderately strong. We also suggest the possibility of using a bootstrapped variance in the standard Wald statistic, a method that performs fairly well under weak instruments. Finally, we construct an entirely new test of the endogeneity correlation that estimates confidence intervals based on an inversion of the robust AR statistic. This test has better size properties than the KLS test under heteroskedasticity and tends to be conservative while KLS tends to overreject. Unlike the KLS test, all three of these alternative tests of ρ_{xu} - the percentile bootstrap, the bootstrapped variance, and the AR test - are robust to heteroskedasticity.

2.8 Appendix

2.8.1 Derivation of equivalence

We run the following regression

$$y = x\beta + z\gamma + u$$

The OLS estimators are given by

$$\hat{\beta} = \frac{1}{A}\left(\hat{\sigma}_z^2\hat{\sigma}_{xy} - \hat{\sigma}_{xz}\hat{\sigma}_{zy}\right)$$
$$\hat{\gamma} = \frac{1}{A}\left(\hat{\sigma}_x^2\hat{\sigma}_{zy} - \hat{\sigma}_{xz}\hat{\sigma}_{xy}\right)$$

where $A = \hat{\sigma}_x^2\hat{\sigma}_z^2 - \hat{\sigma}_{xz}^2 = \hat{\sigma}_x^2\hat{\sigma}_z^2\left(1 - r_{xz}^2\right)$ and e.g. $\hat{\sigma}_{xz} = \frac{1}{n}x'z$, $\hat{\sigma}_z^2 = \frac{1}{n}z'z$ and $r_{xz}^2 = \frac{\hat{\sigma}_{xz}^2}{\hat{\sigma}_x^2\hat{\sigma}_z^2}$.

The KLS estimator for γ as a function of specified correlation between x and u, r_{xu}, is given by

$$\tilde{\gamma}\left(r_{xu}\right) = \hat{\gamma} + \frac{1}{A}\left(r_{xu}\tilde{\sigma}\left(r_{xu}\right)\hat{\sigma}_x\hat{\sigma}_{xz}\right)$$

where

$$\begin{aligned}\tilde{\sigma}^2\left(r_{xu}\right) &= \frac{\hat{\sigma}_u^2}{1 - \frac{1}{A}r_{xu}^2\hat{\sigma}_x^2\hat{\sigma}_z^2} = \frac{A\hat{\sigma}_u^2}{A - r_{xu}^2\hat{\sigma}_x^2\hat{\sigma}_z^2} \\ &= \frac{\left(1 - r_{xz}^2\right)\hat{\sigma}_u^2}{1 - r_{xz}^2 - r_{xu}^2}\end{aligned}$$

and

$$\hat{\sigma}_u^2 = \frac{1}{n}\hat{u}'\hat{u}$$
$$\hat{u} = y - x\hat{\beta} - z\hat{\gamma}.$$

It follows that $\tilde{\gamma}\left(r_{xu}\right) = 0$ if

$$r_{xu}\tilde{\sigma}\left(r_{xu}\right) = \frac{\hat{\sigma}_{xz}\hat{\sigma}_{xy} - \hat{\sigma}_x^2\hat{\sigma}_{zy}}{\hat{\sigma}_x\hat{\sigma}_{xz}}.$$

Consider the IV estimator for β in the model

$$y = x\beta + u$$

using as instrument z. Then

$$\hat{\beta}_{IV} = \frac{z'y}{z'x} = \frac{\hat{\sigma}_{zy}}{\hat{\sigma}_{zx}}.$$

Let the IV residual $\hat{u}_{IV} = y - x\hat{\beta}_{IV}$, it follows that

$$\hat{\sigma}_{z\hat{u}_{iv}} = \hat{\sigma}_{zy} - \hat{\sigma}_{zx}\hat{\beta}_{IV} = 0$$
$$\hat{\sigma}_{x\hat{u}_{IV}} = \hat{\sigma}_{xy} - \hat{\sigma}_x^2\hat{\beta}_{IV}$$

or

$$\hat{\sigma}_{zy} = \hat{\sigma}_{xz}\hat{\beta}_{IV}$$
$$\hat{\sigma}_{xy} = \hat{\sigma}_{x\hat{u}_{IV}} + \hat{\sigma}_x^2\hat{\beta}_{IV}.$$

It follows that

$$\hat{\sigma}_{xz}\hat{\sigma}_{xy} - \hat{\sigma}_x^2\hat{\sigma}_{zy} = \hat{\sigma}_{xz}\hat{\sigma}_{x\hat{u}_{IV}} + \hat{\sigma}_x^2\hat{\sigma}_{xz}\hat{\beta}_{IV} - \hat{\sigma}_x^2\hat{\sigma}_{xz}\hat{\beta}_{IV}$$
$$= \hat{\sigma}_{xz}\hat{\sigma}_{x\hat{u}_{IV}}$$

and so $\tilde{\gamma}(r_{xu}) = 0$ if

$$r_{xu}\tilde{\sigma}(r_{xu}) = \frac{\hat{\sigma}_{x\hat{u}_{IV}}}{\hat{\sigma}_x}.$$

For $\hat{\beta}$ we have

$$\begin{aligned}\hat{\beta} &= \frac{1}{A}\left(\hat{\sigma}_z^2\hat{\sigma}_{xy} - \hat{\sigma}_{xz}\hat{\sigma}_{zy}\right) \\ &= \frac{1}{A}\left(\hat{\sigma}_z^2\hat{\sigma}_{x\hat{u}_{IV}} + \hat{\sigma}_z^2\hat{\sigma}_x^2\hat{\beta}_{IV} - \hat{\sigma}_{xz}^2\hat{\beta}_{IV}\right) \\ &= \hat{\beta}_{IV} + \frac{1}{A}\hat{\sigma}_z^2\hat{\sigma}_{x\hat{u}_{IV}}\end{aligned}$$

and for $\hat{\gamma}$,

$$\begin{aligned}\hat{\gamma} &= \frac{1}{A}\left(\hat{\sigma}_x^2\hat{\sigma}_{zy} - \hat{\sigma}_{xz}\hat{\sigma}_{xy}\right) \\ &= -\frac{1}{A}\hat{\sigma}_{xz}\hat{\sigma}_{x\hat{u}_{IV}}.\end{aligned}$$

It then follows that

$$\begin{aligned}\hat{u} &= y - x\hat{\beta} - z\hat{\gamma} \\ &= y - x\hat{\beta}_{IV} - \frac{1}{A}\left(x\hat{\sigma}_z^2\hat{\sigma}_{x\hat{u}_{IV}} - z\hat{\sigma}_{xz}\hat{\sigma}_{x\hat{u}_{IV}}\right) \\ &= \hat{u}_{IV} + \frac{\hat{\sigma}_{x\hat{u}_{IV}}}{A}\left(z\hat{\sigma}_{xz} - x\hat{\sigma}_z^2\right),\end{aligned}$$

and

$$\begin{aligned}\hat{\sigma}_u^2 &= \hat{\sigma}_{\hat{u}_{IV}}^2 + \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2}{A^2}\left(\hat{\sigma}_z^2\hat{\sigma}_{xz}^2 + \hat{\sigma}_x^2\hat{\sigma}_z^4 - 2\hat{\sigma}_{xz}^2\hat{\sigma}_z^2 - 2A\hat{\sigma}_z^2\right) \\ &= \hat{\sigma}_{\hat{u}_{IV}}^2 + \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2\hat{\sigma}_z^2}{A^2}\left(\hat{\sigma}_{xz}^2 + \hat{\sigma}_x^2\hat{\sigma}_z^2 - 2\hat{\sigma}_{xz}^2 - 2A\right) \\ &= \hat{\sigma}_{\hat{u}_{IV}}^2 + \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2\hat{\sigma}_z^2}{A^2}\left(\hat{\sigma}_x^2\hat{\sigma}_z^2 - \hat{\sigma}_{xz}^2 - 2A\right) \\ &= \hat{\sigma}_{\hat{u}_{IV}}^2 - \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2\hat{\sigma}_z^2}{A} \\ &= \hat{\sigma}_{\hat{u}_{IV}}^2 - \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2}{\hat{\sigma}_x^2\left(1 - r_{xz}^2\right)}\end{aligned}$$

$$\begin{aligned}\tilde{\sigma}^2\left(r_{xu}\right) &= \frac{\left(1-r_{xz}^2\right)\hat{\sigma}_u^2}{1-r_{xz}^2-r_{xu}^2} \\ &= \frac{\left(1-r_{xz}^2\right)\hat{\sigma}_{\hat{u}_{IV}}^2 - \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2}{\hat{\sigma}_x^2}}{1-r_{xz}^2-r_{xu}^2} \\ &= \frac{\left(1-r_{xz}^2\right)\hat{\sigma}_{\hat{u}_{IV}}^2 - \hat{\sigma}_{\hat{u}_{IV}}^2 r_{x\hat{u}_{IV}}^2}{1-r_{xz}^2-r_{xu}^2} \\ &= \hat{\sigma}_{\hat{u}_{IV}}^2 \frac{1-r_{xz}^2-r_{x\hat{u}_{IV}}^2}{1-r_{xz}^2-r_{xu}^2},\end{aligned}$$

where $r_{x\hat{u}_{IV}}^2 = \frac{\hat{\sigma}_{x\hat{u}_{IV}}^2}{\hat{\sigma}_x^2\hat{\sigma}_{\hat{u}_{IV}}^2}$.

It therefore follows that $\tilde{\sigma}^2\left(r_{x\hat{u}_{IV}}^2\right) = \hat{\sigma}_{\hat{u}_{IV}}^2$, and $\tilde{\sigma}\left(r_{x\hat{u}_{IV}}^2\right) = \hat{\sigma}_{\hat{u}_{IV}}$. It follows then that

$$r_{x\hat{u}_{IV}}^2\tilde{\sigma}\left(r_{x\hat{u}_{IV}}^2\right) = \frac{\hat{\sigma}_{x\hat{u}_{IV}}}{\hat{\sigma}_x\hat{\sigma}_{\hat{u}_{IV}}}\hat{\sigma}_{\hat{u}_{IV}} = \frac{\hat{\sigma}_{x\hat{u}_{IV}}}{\hat{\sigma}_x}$$

and so

$$\tilde{\gamma}\left(r_{x\hat{u}_{IV}}^2\right) = 0.$$

Thus,

$$\tilde{\gamma}\left(r_{x\hat{u}_{IV}}\right) = 0.$$

Bibliography

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *The Journal of Human Resources*, 40, 791–821.

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. and P. Guggenberger (2009): "Asymptotic Size and a Problem with Subsampling and the M out of N Bootstrap," *Econometric Theory*, 26, 426–468.

Andrews, D. W. K. (2017): "Identification-Robust Subvector Inference," *SSRN Electronic Journal*.

Andrews, D. W. K. and P. Guggenberger (2006): "The Limit of Finite-Sample Size and a Problem With Subsampling," *Yale Economics Department Research Papers*.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J., V. Lavy, and A. Schlosser (2010): "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *Journal of Labor Economics*, 28, 773–824.

Basmann, R. L. (1960): "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics," *Journal of the American Statistical Association*, 55, 650–659.

Briel, S., A. Osikominu, G. Pfeifer, M. Reutter, and S. Satlukal (2020): "Overconfidence and Gender Differences in Wage Expectations," *SSRN Electronic Journal*.

Chaudhuri, S. and E. Zivot (2011): "A new method of projection-based inference in GMM with weakly identified nuisance parameters," *Journal of Econometrics*.

Chen, G. (2022): "Parasocial While Meaningful: How Does Exposure to Foreign Cultures Affect One's Opinion of Foreign Countries?" Ph.D. thesis.

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

Davidson, R. and E. Flachaire (2008): "The Wild Bootstrap, Tamed at Last," *Journal of Econometrics*, 146, 162–169.

Davidson, R. and J. MacKinnon (2015): "Bootstrap Tests for Overidentification in Linear Regression Models," *Econometrics*, 3, 825–863.

Davidson, R. and J. G. MacKinnon (2010): "Wild Bootstrap Tests for IV Regression," *Journal of Business & Economic Statistics*, 28, 128–144.

Davies, N. M., S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer, and G. D. Smith (2014): "The Many Weak Instruments Problem and Mendelian Randomization," *Statistics in Medicine*, 34, 454–468.

Dufour, J.-M. and M. Taamouti (2005): "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," *Econometrica*, 73, 1351–1365.

Gospodinov, N., R. Kan, and C. Robotti (2017): "Too Good to Be True? Fallacies in Evaluating Risk Factor Models," Working Paper Series 2017-9.

Gray-Lobe, G., P. Pathak, and C. Walters (2021): "The Long-Term Effects of Universal Preschool in Boston," Tech. rep.

Grohmann, J. (2021): "Power of the subvector Anderson-Rubin and bootstrap tests in linear instrumental variables regression with weak instruments," mathesis, University of Oxford.

Grohmann, J. and F. Windmeijer (2021): "Power of the Subvector Anderson-Rubin and Bootstrap Tests in Linear Instrumental Variables Regression with Weak Instruments," Master's thesis, University of Oxford.

Guggenberger, P., F. Kleibergen, and S. Mavroeidis (2019): "A More Powerful Subvector Anderson Rubin Test in Linear Instrumental Variables Regression," *Quantitative Economics*, 10, 487–526.

——— (2020): "A Test for Kronecker Product Structure Covariance Matrix," *University of Oxford*.

——— (2021): "A Powerful Subvector Anderson Rubin Test in Linear Instrumental Variables Regression with Conditional Heteroskedasticity," *University of Oxford*.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression," *Econometrica*, 80, 2649–2666.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Kiviet, J. F. (2020): "Testing the impossible: Identifying exclusion restrictions," *Journal of Econometrics*, 218, 294–316.

——— (2022): "Instrument-free inference under confined regressor endogeneity and mild regularity," *Econometrics and Statistics*.

Kiviet, J. F. and S. Kripfganz (2020): "Reassessment of classic case studies in labor economics with new instrument-free methods," Tech. rep.

Kleibergen, F. and R. Paap (2006): "Generalized Reduced Rank Tests Using the Singular Value Decomposition," *Journal of Econometrics*, 133, 97–126.

Liu, R. Y. (1988): "Bootstrap Procedures under Some Non-I.I.D. Models," *The Annals of Statistics*, 16, 1696–1708.

MacKinnon, J. G. (2013): *Thirty years of heteroskedasticity-robust inference*, Springer.

Mammen, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21.

Mariella, V. (2021): "Landownership Concentration and Human Capital Accumulation in Post-Unification Italy (1871–1921)," *SSRN Electronic Journal*.

Masten, M. A. and A. Poirier (2021): "Salvaging Falsified Instrumental Variable Models," *Econometrica*, 89, 1449–1469.

Moreira, M. J., J. R. Porter, and G. A. Suarez (2009): "Bootstrap Validity for the Score Test when Instruments May Be Weak," *Journal of Econometrics*, 149, 52–64.

Sanderson, E. and F. Windmeijer (2016): "A Weak Instrument F-test in Linear IV Models with Multiple Endogenous Variables," *Journal of Econometrics*, 190, 212–221.

Sargan, J. D. (1958): "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, 26, 393–415.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Van de Sijpe, N. and F. Windmeijer (2022): "On the Power of the Conditional Likelihood Ratio and Related Tests for Weak-Instrument Robust Inference," *Journal of Econometrics*.

van Kippersluis, H. and C. A. Rietveld (2018): "Beyond plausibly exogenous," *The Econometrics Journal*, 21, 316–331.

Wang, H. and Z. Cheng (2022): "Kids eat free: School feeding and family spending on education," *Journal of Economic Behavior &amp Organization*, 193, 196–212.

Wang, W. and F. Doko Tchatoka (2018): "On Bootstrap Inconsistency and Bonferroni-Based Size-Correction for the Subset Anderson–Rubin Test Under Conditional Homoskedasticity," *Journal of Econometrics*, 207, 188–211.

Windmeijer, F. (2019): "Weak Instruments, First-Stage Heteroskedasticity and the Robust F-Test," Dept. of Economics Discussion Paper 19/708.

——— (2021): "Testing Underidentification in Linear Models, with Applications to Dynamic Panel and Asset Pricing Models," *Journal of Econometrics*.

Wu, C.-F. J. (1986): "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *The Annals of Statistics*, 14, 1261–1295.