Two-Step Confidence Sets
and the Trouble with the First Stage F Statistic

Isaiah Andrews\textsuperscript{1}

\textsuperscript{1}Harvard Society of Fellows

September 29, 2014
Two-step confidence sets

When weak identification is a concern, researchers typically proceed in two steps:

1. First calculate statistic intended to measure identification strength, for example first stage $F$
2. Then:
   - If model seems well-identified, proceed as usual and report non-robust confidence sets
   - If weak identification seems an issue, calculate robust confidence sets, seek new specification, or don’t report results

I focus on the case where we always report some confidence set

Other cases can lead to terrible properties for reported confidence sets
Introduction
Two-step confidence sets

- Represent the outcome of the first step by the identification statistic $\phi_{ICS} \in \{0, 1\}$
  - $\phi_{ICS} = 1$ indicates evidence of weak identification
- In the second step we’ll use
  - $CS_{NR}$ if identification seems strong
  - $CS_R$ if identification seems weak
- We can write two-step confidence sets as

$$CS_{2S} = \begin{cases} 
CS_{NR} & \text{if } \phi_{ICS} = 0 \\
CS_R & \text{if } \phi_{ICS} = 1 
\end{cases}$$
Introduction
Example: Linear IV

- We could let $CS_{NR}$ be the t-statistic confidence set based on 2SLS or LI ML.
- Let $CS_R$ to be an Anderson-Rubin confidence set.
- For $F$ the first-stage F-statistic, the commonly-used Staiger and Stock (1997) rule of thumb corresponds to

$$\phi_{ICS} = 1 \{ F \leq 10 \}$$

- For these choices, $CS_{2S}$
  - reports the usual t-statistic CS when $F > 10$
  - reports the Anderson-Rubin CS when $F \leq 10$
I’ll be interested in the coverage of the two-step confidence set

\[ Pr_{\beta_0} \{ \beta_0 \in CS_{2S} \} \]

Throughout I’ll assume that

- \( CS_{NR} \) has coverage at least \( 1 - \alpha \) under strong identification
- \( CS_R \) has coverage at least \( 1 - \alpha \) under both strong and weak identification

I’ll be interested in the maximal coverage distortion for \( CS_{2S} \), defined as the smallest \( \gamma \) such that

\[ Pr_{\beta_0} \{ \beta_0 \in CS_{2S} \} \geq 1 - \alpha - \gamma \]
Unsurprisingly, $CS_{2S}$ can exhibit large distortions if we choose $\phi_{ICS}$ poorly.

Perhaps more surprising- Stock and Yogo (2005) show that pretests based on the first-stage F-statistic can bound coverage distortions in linear IV with homoskedastic errors.

- For example, Stock and Yogo give a test for the null that the nominal 5% t-test has true size exceeding 10%.
- For 2SLS, critical values larger than the “rule of thumb” value $c = 10$.

Problem: we frequently think economic data are heteroskedastic, serially correlated, clustered, etc.
I demonstrate numerically that the first-stage F statistic is not a reliable measure of identification strength in IV with heteroskedastic errors.

I propose a general approach to constructing robust two-step confidence sets that:

1. is applicable to general GMM models
2. controls coverage up to an arbitrarily small, user-selected level of distortion $\gamma$ when identification is weak
3. indicates strong identification with probability tending to one under standard asymptotics
Introduction

Main Idea

- Under strong identification, many different confidence sets are asymptotically equivalent.
- In particular, we can construct robust confidence sets with coverage $1 - \alpha - \gamma$ which are contained in $CS_{NR}$ with probability tending to one under strong identification.
- Thus, to assess the quality to the usual asymptotic approximations, we can compare robust and non-robust confidence sets.
- By choosing $\phi_{ICS}$ in this way, we can ensure that $CS_{2S}$ has coverage at least $1 - \alpha - \gamma$. 
Introduction
Why study two-step procedures?

- Two-step confidence sets are often unappealing from a theoretical perspective
  - Nonetheless, pretests for identification are extremely common in empirical work
  - Commonly used pretests are unreliable, so alternatives are needed

- Robust confidence sets that are efficient under strong ID give a natural, reliable way to assess ID strength
  - Further argument for computing and reporting such confidence sets
Outline

1. Introduction
2. Problems with the First-Stage F-statistic
3. A Simple Two-Step Confidence Set
4. Simulation Performance
5. Reporting Results
6. Conclusion
Problems with the First-Stage F-statistic

Linear IV model

Consider the linear IV model

\[ Y_t = X_t \beta_0 + V_{1,t} \]

\[ X_t = Z_t' \pi + V_{2,t} \]

with \( Z_t \) a \( k \times 1 \) vector of instruments,
\[ E [Z_t V_{1,t}] = E [Z_t V_{2,t}] = 0 \]

Simulations set \( \beta_0 = 0 \)

Calibrations with moderate and high endogeneity
Problems with the First-Stage F-statistic
Simulation design

- I report results for two-step confidence sets with
  - $CS_{NR}$ the nominal 95% Wald confidence set based on:
    - 2SLS
    - LIML
  - $CS_R$ the Anderson-Rubin confidence set
  - $\phi_{ICS} = 1 \{F \leq c\}$ for:
    - the rule-of-thumb cutoff $c = 10$ (RoT)
    - the Stock and Yogo (2005) cutoffs (SY)

- SY cutoffs guarantee coverage at least 85% in homoskedastic case
- For heteroskedastic simulations, I use a heteroskedastic-robust form of the first stage F
# Problems with the First-Stage F-statistic

## Minimal coverage: homoskedastic case

<table>
<thead>
<tr>
<th>Minimal Coverage</th>
<th>Moderate Endogeneity</th>
<th>High Endogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 5$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td>RoT LIML</td>
<td>92.9%</td>
<td>93.1%</td>
</tr>
<tr>
<td>RoT 2SLS</td>
<td>90.4%</td>
<td>89.1%</td>
</tr>
<tr>
<td>SY LIML</td>
<td>91.6%</td>
<td>90.6%</td>
</tr>
<tr>
<td>SY 2SLS</td>
<td>92.6%</td>
<td>92.4%</td>
</tr>
</tbody>
</table>

**Table**: Minimal coverage for confidence sets in homoskedastic IV simulations with 10,000 observations.
Problems with the First-Stage F-statistic

Minimal coverage: heteroskedastic case

<table>
<thead>
<tr>
<th>Minimal Coverage</th>
<th>Moderate Endogeneity</th>
<th>High Endogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 5$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td>RoT LIML</td>
<td>63.2%</td>
<td>44.8%</td>
</tr>
<tr>
<td>RoT 2SLS</td>
<td>55.1%</td>
<td>30.8%</td>
</tr>
<tr>
<td>RoT CUGMM</td>
<td>45.4%</td>
<td>18.4%</td>
</tr>
<tr>
<td>RoT 2SGMM</td>
<td>35.4%</td>
<td>13%</td>
</tr>
<tr>
<td>SY LIML</td>
<td>61.2%</td>
<td>43%</td>
</tr>
<tr>
<td>SY 2SLS</td>
<td>63.6%</td>
<td>40.2%</td>
</tr>
<tr>
<td>SY CUGMM</td>
<td>39.5%</td>
<td>15.3%</td>
</tr>
<tr>
<td>SY 2SGMM</td>
<td>46.7%</td>
<td>19.8%</td>
</tr>
</tbody>
</table>

Table: Minimal coverage for confidence sets in heteroskedastic IV simulations with 10,000 observations.
Problems with the First-Stage F-statistic
Minimal coverage: heteroskedastic case

- As we can see, all of the two-step procedures studied can exhibit large coverage distortions.
- The problem: when the data are heteroskedastic, the first stage F-statistic doesn’t give a reliable guide to identification strength.
- In the $k = 10$ moderate endogeneity calibration, substantial coverage distortions persist even for $E[F] = 500$.
  - High endogeneity calibration much more extreme: significant distortions even for $E[F] = 100,000$. 
Outline

1. Introduction
2. Problems with the First-Stage F-statistic
3. A Simple Two-Step Confidence Set
4. Simulation Performance
5. Reporting Results
6. Conclusion
The idea: compare robust and non-robust confidence sets.

In general GMM models, for all the commonly-used non-robust confidence sets $CS_{NR}$ and any $\gamma > 0$, we can construct a preliminary robust confidence set $CS_{R,P}$ with

1. $Pr_{\beta_0} \{ \beta_0 \in CS_{R,P} \} \geq 1 - \alpha - \gamma$ regardless of identification strength
2. $Pr \{ CS_{R,P} \subseteq CS_{NR} \} \to 1$ under strong identification
3. $CS_{R,P} \subseteq CS_R$
If we take

\[ \phi_{ICS} = 1 \{ CS_R,P \not\subseteq CS_{NR} \} \]

then \( \phi_{ICS} \rightarrow_p 0 \) under strong identification. Moreover \( CS_R,P \subseteq CS_{2S} \) so

\[ Pr_{\beta_0} \{ \beta_0 \in CS_{2S} \} \geq 1 - \alpha - \gamma \]

regardless of identification strength
GMM Model

- Consider a potentially weakly identified GMM model with identifying assumption $E_{\beta_0} [g_T (\beta_0)] = 0$

- Linear IV: $g_T (\beta) = \frac{1}{T} \sum Z_t (Y_t - X_t \beta)$
- For convenience, continue to assume $\dim (\beta) = 1$

- Let

$$\lim_{T \to \infty} \text{Var} \begin{pmatrix} \sqrt{T} g_T (\beta) \\ \sqrt{T} \frac{\partial}{\partial \beta} g_T (\beta) \end{pmatrix} = \begin{pmatrix} \Sigma (\beta) & \Sigma_\beta (\beta) \\ \Sigma_\beta (\beta) & \Sigma_{\beta \beta} (\beta) \end{pmatrix}$$
A Simple Two-Step Confidence Set

Robust test statistics

- Define the $S$ statistic of Stock and Wright (2000) as

$$S(\beta) = T \cdot g_T(\beta)' \hat{\Sigma}(\beta)^{-1} g_T(\beta).$$

- $S(\beta_0) \rightarrow_d \chi_k^2$ regardless of identification strength

- Following Kleibergen (2005), let

$$D_T(\beta) = \frac{\partial}{\partial \beta} g_T(\beta) - \hat{\Sigma}_\beta(\beta) \hat{\Sigma}(\beta)^{-1} g_T(\beta)$$

and define the $K$ statistic

$$K(\beta) = T \cdot g_T(\beta)' \hat{\Sigma}(\beta)^{-\frac{1}{2}} P_{\hat{\Sigma}(\beta)^{-\frac{1}{2}} D_T(\beta)} \hat{\Sigma}(\beta)^{-\frac{1}{2}} g_T(\beta)$$

- $K(\beta_0) \rightarrow_d \chi_1^2$ regardless of identification strength
A Simple Two-Step Confidence Set

Local asymptotic equivalence

- Under standard assumptions, for $W(\beta)$ a Wald statistic based on an efficient estimator $\hat{\beta}$

$$\sup_{\sqrt{T}\|\beta - \beta_0\| \leq C} \| W(\beta) - K(\beta) \| = o_p(1)$$

under strong identification

- Thus, the K-statistic confidence set

$$\{ \beta : K(\beta) \leq \chi^2_{1,1-\alpha} \}$$

is asymptotically equivalent to the usual Wald confidence set on $\sqrt{T}$-neighborhoods of $\beta_0$

- More general statement given in paper
A Simple Two-Step Confidence Set

Non-local non-equivalence

- Unfortunately, the equivalence of the K and Wald confidence sets holds only locally, not globally.
- K confidence sets are often inconsistent, in the sense that even in strongly identified models they fail to shrink towards the true parameter value as the sample grows.
A Simple Two-Step Confidence Set

A consistent robust confidence set

- To obtain a consistent confidence set, for $a > 0$ consider

$$CS_{R,P} = \{ \beta : K(\beta) + a \cdot S(\beta) \leq \chi^2_{1,1-\alpha} \} .$$

$K(\beta) + a \cdot S(\beta)$ is a linear combination statistic, as in Andrews (2013)

- This confidence set has coverage

$$1 - \alpha - \gamma(a) = Pr \{(1 + a) \cdot \chi^2_1 + a \cdot \chi^2_{k-1} \leq \chi^2_{1,1-\alpha}\}$$

regardless of identification strength

- $\gamma \rightarrow 0$ as $a \rightarrow 0$, and we can choose $a$ to obtain any desired level of $\gamma$
A Simple Two-Step Confidence Set
Detecting weak identification

- For $CS_{NR}$ the Wald confidence set, under strong identification

$$Pr \{ CS_{R,P} \subseteq CS_{NR} \} \to 1$$

- To assess whether the usual strong-identification approximations are reasonable, can check whether $CS_{R,P} \subseteq CS_{NR}$
  - Motivates choice $\phi_{ICS} = 1 \{ CS_{R,P} \not\subseteq CS_{NR} \}$
  - Gives $Pr \{ \beta_0 \in CS_{2S} \} \geq 1 - \alpha - \gamma$
Outline

1. Introduction
2. Problems with the First-Stage F-statistic
3. A Simple Two-Step Confidence Set
4. Simulation Performance
5. Reporting Results
6. Conclusion
For comparability with my earlier simulation results, I set \( \gamma = 10\% \), so \( CS_{R,P} \) has coverage 85%.

Thus, \( CS_{2S} \) has coverage at least 85% as well.

I again start by considering the homoskedastic case.
## Simulation Performance

### Homoskedastic Linear IV

<table>
<thead>
<tr>
<th>Minimal Coverage</th>
<th>Moderate Endogeneity</th>
<th>High Endogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CS_{2S}$ LIML</td>
<td>$92.6%$ $92.4%$ $90.4%$</td>
<td>$86%$ $87%$ $85%$</td>
</tr>
<tr>
<td>$CS_{2S}$ 2SLS</td>
<td>$92.8%$ $92.8%$ $92.9%$</td>
<td>$86%$ $87%$ $85%$</td>
</tr>
</tbody>
</table>

*Table*: Minimal coverage for different confidence sets in homoskedastic IV simulations with 10,000 observations.
We see that $CS_{2S}$ controls coverage distortions in the homoskedastic case.

Since procedures based on the first-stage $F$ statistic also work here, we can compare their performance.

In particular, we might worry that my $CS_{2S}$ controls coverage distortions by setting $\phi_{ICS} = 1$ more often than procedures based on the first-stage $F$.

This isn’t the case.
Figure: $E[\phi_{ICS}] = Pr \{\phi_{ICS} = 1\}$ plotted against the mean of the first stage F-statistic in moderate endogeneity homoskedastic linear IV calibration with $k = 10$. 
Simulation Performance
Heteroskedastic Linear IV

<table>
<thead>
<tr>
<th>Minimal Coverage</th>
<th>Moderate Endogeneity</th>
<th>High Endogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 5$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td>CS$_2$S LIML</td>
<td>94.1%</td>
<td>92.4%</td>
</tr>
<tr>
<td>CS$_2$S 2SLS</td>
<td>93.7%</td>
<td>94%</td>
</tr>
<tr>
<td>CS$_2$S CUGMM</td>
<td>95%</td>
<td>94.1%</td>
</tr>
<tr>
<td>CS$_2$S 2SGMM</td>
<td>94.5%</td>
<td>94%</td>
</tr>
</tbody>
</table>

Table: Minimal coverage for different confidence sets in heteroskedastic IV simulations with 10,000 observations.
Outline

1. Introduction
2. Problems with the First-Stage F-statistic
3. A Simple Two-Step Confidence Set
4. Simulation Performance
5. Reporting Results
6. Conclusion
Reporting Results

Picking $\gamma$

1. My discussion so far assumes a choice of the maximal distortion $\gamma$
   - What if you don’t like my $\gamma$?

2. When used at all in empirical work robust confidence sets typically supplement, rather than replace, non-robust ones.
Pick some \( \gamma_{\text{min}} \geq 0 \). For \( \gamma \geq \gamma_{\text{min}} \), consider the family of robust confidence sets

\[
CS_{R,P}(\gamma) = \{ \beta : K(\beta) + a(\gamma) \cdot S(\beta) \leq \chi^2_{1,1-\alpha} \}
\]

where \( CS_{R,P}(\gamma) \) has coverage \( 1 - \alpha - \gamma \)
Note that decreasing in $\gamma$: $\gamma \leq \gamma' \Rightarrow CS_{R,P}(\gamma') \subseteq CS_{R,P}(\gamma)$

Consider $CS_R$ such that $CS_{R,P}(\gamma_{\text{min}}) \subseteq CS_R$

Define the *maximal distortion cutoff*

$$\widehat{\gamma} = \min \{ \gamma \geq \gamma_{\text{min}} : CS_{R,P}(\gamma) \subseteq CS_{NR} \}$$
Reporting Results

- We can report $CS_{NR}$, $CS_R$, and $\hat{\gamma}$
- If I adopt the rule that I will use $CS_{NR}$ if $\hat{\gamma} \leq \gamma^*$, and will use $CS_R$ otherwise, “my” confidence set has coverage at least $1 - \alpha - \gamma^*$
Acemoglu, Johnson, Robinson, and Yared (2008) study the relationship between per-capita income and measures of political democracy. They argue that once one controls for country fixed effects, there is no significant relationship.

Consider a number of specifications, including one which instruments income with the trade-weighted income of a country’s trading partners.

Cervellati, Jung, Sunde and Vischer (2014) argue that this zero effect finding masks heterogeneity across countries, and in particular that the relationship between income and democracy is negative in former European colonies.

I re-examine the identifying power of the trade-weighted income instrument.
## Reporting Results

### Income and Democracy

<table>
<thead>
<tr>
<th>Specification</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>2SLS</td>
<td>-0.120 (0.11)</td>
<td>-1.37 (23.0)</td>
<td>-0.16 (0.07)</td>
</tr>
<tr>
<td>First Stage F</td>
<td>26.5</td>
<td>0.0</td>
<td>57.6</td>
</tr>
<tr>
<td>( \hat{\gamma} )</td>
<td>20.2%</td>
<td>0%</td>
<td>57.9%</td>
</tr>
<tr>
<td>SCS</td>
<td>[-1,1]</td>
<td>[-1,1]</td>
<td>[-1,1]</td>
</tr>
<tr>
<td>Sample Restriction</td>
<td>None</td>
<td>Non-colonies</td>
<td>Former Colonies</td>
</tr>
<tr>
<td>Observations</td>
<td>895</td>
<td>218</td>
<td>718</td>
</tr>
</tbody>
</table>

*Table*: Results based on Acemoglu et. al. (2008) data
Conclusion

- I show that two-step confidence sets based on the first-stage F-statistic can be highly unreliable in non-homoskedastic data.
- I propose an alternative approach to detecting weak identification and constructing two-step confidence sets which:
  - Controls coverage distortions under weak identification and is applicable to general GMM models.
  - Indicates strong identification with probability tending to one under standard asymptotics.
- While I’ve framed my discussion using linear combination confidence sets, many other options also work.
- Close connection between efficient robust confidence sets and reliable assessments of identification strength.
The End

Thank you!
Formal Results

Weak identification assumptions

Assumption

For \( J_T(\beta) = E_T \left[ \frac{\partial}{\partial \beta'} g_T(\beta) \right] \),

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \begin{array}{c}
 vec \left( \frac{\partial}{\partial \beta'} g_t(\beta_0) - J_T(\beta_0) \right)
\end{array} \right) \rightarrow_d N \left( 0, \left( \begin{array}{cc}
 \Sigma_g & \Sigma_{g\theta} \\
 \Sigma_{\theta g} & \Sigma_\theta
\end{array} \right) \right)
\]

where \( \Sigma_g \) is positive definite and

\[
\left( \begin{array}{cc}
 \Sigma_g & \Sigma_{g\beta} \\
 \Sigma_{\beta g} & \Sigma_{\beta\beta}
\end{array} \right) = \lim_{T \rightarrow \infty} \text{Var}_{T,\xi} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \begin{array}{c}
 vec \left( \frac{\partial}{\partial \beta'} g_t(\beta_0) \right)
\end{array} \right) \right)
\]

is consistently estimable.
Formal Results
Weak identification assumptions

Assumption

There exists sequence of full-rank $O(\sqrt{T})$ normalizing matrices $\Lambda_{1,T}$ such that $D_T (\beta_0) \Lambda_{1,T} \rightarrow_d D$ for a (possibly degenerate) Gaussian random matrix $D$ which is full rank almost surely.
Formal Results
Identification-robust tests

Theorem

*Under these assumptions,*

\[(K(\beta_0), S(\beta_0) - K(\beta_0)) \rightarrow_d (\chi^2_p, \chi^2_{k-p})\]

*and \(K(\beta_0)\) and \(S(\beta_0) - K(\beta_0)\) are asymptotically independent*
Assumption

1. $g_T(\beta) \xrightarrow{p} \lim_{T \to \infty} E_T [g_t(\beta)]$ uniformly, and $\lim_{T \to \infty} E_T [g_t(\beta)]$ is continuous in $\beta$.

2. $\hat{\Sigma}_g(\beta) \xrightarrow{p} \Sigma_g(\beta)$ uniformly for $\Sigma_g(\beta)$ continuous in $\beta$ and everywhere positive definite with a uniformly bounded maximal eigenvalue and a minimal eigenvalue bounded away from zero.

3. For all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$
\left( \lim_{T \to \infty} E_T [g_t(\beta)] \right)' \Omega(\beta) \left( \lim_{T \to \infty} E_T [g_t(\beta)] \right) < \delta
$$

only if $\|\theta - \theta_0\| < \varepsilon$. 
Formal Results
Strong identification assumptions

Assumption

1. \( \beta_0 \) belongs to the interior of its parameter space

2. \( g_T(\beta) \) and \( \hat{\Sigma}_g(\beta) \) are almost surely continuously differentiable on some open ball \( B(\beta_0) \) around \( \beta_0 \)

3. For

\[
J(\beta) = \lim_{T \to \infty} E_T \left[ \frac{\partial}{\partial \beta'} g_T(\beta) \right],
\]

\( J(\beta) \) is continuous at \( \beta_0 \), \( G_T(\beta) = \frac{\partial}{\partial \beta'} g_T(\beta) \to_p J(\beta) \) uniformly on \( B(\beta_0) \), and \( J(\beta_0) \) is full-rank

4. \( \sup_{\beta \in B(\beta_0)} \left\| \frac{\partial vec(\hat{\Sigma}_g(\beta))}{\partial \beta'} \right\| = O_p(1) \)
Formal Results
Identification-robust tests

Theorem

Under these assumptions we have that for all $C \geq 0$

$$\sup_{\sqrt{T} \| \beta - \beta_0 \| \leq C} \| W(\beta) - K(\beta) \| = o_p(1).$$