LEVERAGE AND COVARIANCE MATRIX ESTIMATION
IN FINITE-SAMPLE IV REGRESSIONS

Andreas Steinhauer, University of Zurich*   Tobias Wuergler, University of Zurich††

September 24, 2010

Abstract

This paper develops basic algebraic concepts for instrumental variables (IV) regressions which are used to derive the leverage and influence of observations on the 2SLS estimate and compute alternative heteroskedasticity-consistent (HC1-HC3) estimators for the 2SLS covariance matrix in a finite-sample context. In the second part, Monte Carlo simulations and an application to growth regressions are used to evaluate the performance of these estimators. The results suggest guidelines for applied IV projects, supporting the use of HC3 instead of standard White’s robust (HC0) errors in smaller, less balanced data sets with influential observations, in line with earlier results on alternative heteroskedasticity-consistent estimators for OLS. The leverage and influence of observations can be examined with the various measures derived in the paper.

Keywords: Two stage least squares, leverage, influence, heteroskedasticity-consistent covariance matrix estimation

---

*University of Zurich, Institute for Empirical Research in Economics, Muehlebachstrasse 86, CH-8008 Zürich, e-mail: steinhauer@iew.uzh.ch
†University of Zurich, Institute for Empirical Research in Economics, Muehlebachstrasse 86, CH-8008 Zürich, e-mail: wuergler@iew.uzh.ch
‡We thank Joshua Angrist, A. Colin Cameron, Andreas Kuhn, Jörn-Steffen Pischke, Kevin Staub, Michael Wolf and Jean-Philippe Wuellrich for valuable input.
1 Introduction

The implication of heteroskedasticity for inference based on OLS in the linear regression model has been extensively studied. If the form of heteroskedasticity is known, generalized least squares techniques can be performed, restoring the desirable finite-sample properties of the classical model. In practice the exact form is rarely known and the famous robust estimator presented by White (1980) is widely used to generate consistent estimates of the covariance matrix (see also Eicker, 1963). However, the sample size is required to be sufficiently large for this estimator to make valid inference. In a finite-sample context using Monte Carlo experiments, MacKinnon and White (1985) demonstrated the limits of the robust estimator and studied several alternative heteroskedasticity-consistent estimators with improved properties (known as $HC1$, $HC2$ and $HC3$ errors as opposed to the standard "$HC0" error due to White and Eicker). While $HC1$ incorporates a simple degrees of freedom correction, $HC2$ (due to Hinkley, 1977) and $HC3$ (the jackknife estimator, see Efron, 1982) aim to control explicitly for the influence of high leverage observations.

In the theoretical instrumental variables (IV) and generalized method of moments (GMM) literature heteroskedasticity has been addressed (e.g. White, 1982, Hansen, 1982). It is common practice to use White’s robust ($HC0$) standard errors which are consistent and valid in large samples. However, while biasedness of IV estimators in finite samples has received wide attention (e.g. Nelson and Startz, 1990a,b, Buse, 1992, Bekker, 1994, Bound, Jaeger, and Baker, 1995), the small-sample properties of heteroskedasticity-consistent covariance estimators in IV regressions have not been studied explicitly so far to the best of our knowledge. As it is possible to extend the alternative forms of heteroskedasticity-consistent covariance matrix estimators from an OLS to an IV environment, the present paper develops such estimators and evaluates them with Monte Carlo experiments and an application to growth regressions.

We base our main results on the two-stage least squares (2SLS) approach to IV/GMM estimation of single equation linear models. 2SLS is equivalent to other IV/GMM methods in the exactly identified case, and serves as a natural benchmark in the overidentified case as it is the efficient GMM estimator under conditional homoskedasticity (analogously to $HC1$-$2$ errors for OLS being based on properties derived under homoskedasticity). Moreover, it has been shown that the efficient GMM coefficient estimator has poor small-sample properties as it requires the estimation of fourth moments, and that approaches like 2SLS or using the identity matrix as weighting matrix are superior in smaller samples (see July 1996 issue of the Journal of Business and Economic Statistics). Finally and most importantly, the 2SLS approach is widely used in empirical studies.

In the first part of the paper, the robust covariance matrix estimators are derived. We ana-
lyze the hat and residual maker matrix in 2SLS regressions and extend the Frisch-Waugh-Lovell theorem for OLS (Frisch and Waugh, 1933, Lovell, 1963) to 2SLS. We then use these algebraic properties to derive explicit expressions for the leverage and influence of single observations on the 2SLS estimate. These concepts are valuable tools on their own as they can be used to perform influential diagnostics in IV regressions. Finally and most importantly, we demonstrate that, analogous to the case of OLS, the leverage of single observations is intrinsically linked to the problem of heteroskedasticity-consistent covariance estimation in finite-sample IV regressions, and compute the alternative forms of robust estimators for 2SLS.

In the second part, the performance of the various covariance matrix estimators is evaluated in Monte Carlo simulations as well as in existing growth regressions involving instrumental variables. We begin with the simplest case of one (endogenous) regressor and one instrument, parametrically generating data, simulating and computing the conventional non-robust as well as the robust HC0-3 standard errors. We compare size distortions and other measures in various parameterizations of the model with different degrees of heteroskedasticity across different sample sizes. We then analyze further specifications, changing the number of instruments as well as the data generating process. Finally, we re-examine two well-known growth regression studies of Persson and Tabellini (1994) and Acemoglu, Johnson and Robinson (2001) which use countries as units of observations, so they are naturally subject to smaller sample issues and consequently well suited to illustrate the derived concepts.

The Monte Carlo simulations show that similar to OLS the size distortions in tests based on White’s robust (HC0) errors may be substantial. Empirical rejection rates may exceed nominal rates by a great margin depending on the design. HC1-3 estimators mitigate the problem with HC3 performing the best, bringing down the distortion substantially. HC3 errors have the further advantage of working relatively well both in a homoskedastic and in a heteroskedastic environment as opposed to conventional non-robust and White’s robust (HC0) errors. These results highlight the importance of computing and analyzing leverage measures as well as alternative covariance matrix estimators especially when performing IV regressions with smaller and less balanced data sets.

The application of the HC1-3 covariance matrix estimators to the two growth regression studies mentioned above demonstrates that results without adjustments to robust errors may indicate too high a precision of the estimation if the sample size is small and the design unbalanced. In the presence of a few influential observations as in one specification of the study of Persson and Tabellini (1994), the level of significance may be substantially lowered if the HC3 error estimator is used. On the other hand, in a relatively balanced design as the one of Acemoglu, Johnson and Robinson (2001) where no single observation is of particular influence, the use of HC3 does barely affect the significance level of the results.
The paper is organized as follows: Section 2 provides an overview of the issue of covariance matrix estimation in the presence of heteroskedasticity in 2SLS regressions. Section 3 derives the basic algebraic concepts for 2SLS regressions which are used to compute the leverage and influence of observations on the 2SLS estimate. Building on these concepts, Section 4 computes the various forms of heteroskedasticity-consistent covariance matrix estimators. In Section 5, we describe and present the results of parametric Monte Carlo experiments to examine and compare the performance of the alternative estimators. Section 6 applies the developed diagnostics and error estimators to two growth regression studies. Section 7 concludes.

2 The Model and Covariance Matrix Estimation

Consider estimation of the model

\[ y = X\beta + \varepsilon, \]

where \( y \) is a \((n \times 1)\) vector of observations of the dependent variable, \( X \) is a \((n \times L)\) matrix of observations of regressors, and \( \varepsilon \) is a \((n \times 1)\) vector of the unobservable error terms. Suppose that some of the regressors are endogenous, but there is a \((n \times K)\) matrix of instruments \( Z \) (including exogenous regressors) which are predetermined in the sense of \( E(Z'\varepsilon) = 0 \), and the \( K \times L \) matrix \( E(Z'X) \) is of full column rank. Furthermore, suppose that \( \{y_i, x_i, z_i\} \) is jointly ergodic and stationary, \( \{z_i, \varepsilon_i\} \) is a martingale difference sequence, and \( E[(z_{ik}x_{il})^2] \) exists and is finite.

Under these assumptions the efficient GMM estimator

\[
\hat{\beta}_{GMM} = \left(X'Z(Z'\hat{\Omega}Z)^{-1}Z'X\right)^{-1}X'Z(Z'\hat{\Omega}Z)^{-1}Z'y
\]

is consistent, asymptotically normal and efficient among linear GMM estimators, where \( \hat{\Omega} = \text{diag} (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_n^2) \) is an estimate of the covariance matrix of the error terms \( \Omega \) (see for example Hayashi, 2000). As mentioned in the Introduction, however, it has been shown that GMM estimators that do not require the estimation of fourth moments \((Z'\hat{\Omega}Z)\) tend to be superior in terms of bias and variance in smaller samples. One such estimator that is widely used in applied research is the 2SLS estimator,

\[
\hat{\beta} \equiv \hat{\beta}_{2SLS} = (X'PX)^{-1}X'Py,
\]

where \( P = Z(Z'Z)^{-1}Z' \) is the projection matrix associated with the instruments \( Z \). The 2SLS estimator is the efficient GMM estimator under conditional homoskedasticity \((E(\varepsilon_i^2 | z_i) = \sigma^2)\). Even if the assumption of conditional homoskedasticity cannot be made, 2SLS still is consistent although not necessarily efficient since it is a GMM estimator with \((Z'Z)^{-1}\) as weighting matrix.
If one cannot assume conditional homoskedasticity, the asymptotically valid estimator of the covariance matrix is given by

$$\widehat{\text{Var}}_{HC0}(\hat{\beta}) = (X'PX)^{-1}X'P\hat{\Omega}PX(X'PX)^{-1},$$

which is White’s robust estimator also known as $HC0$ (heteroskedasticity-consistent). Unlike the 2SLS coefficient estimator, the corresponding $HC0$ covariance matrix estimator requires fourth moment estimation which lies at the heart of the issues studied here. Although being valid in large samples, the $HC0$ estimator does not account for the fact that residuals tend to be too small and distorted in finite samples.

Robust estimators such as $HC0$ require the estimation of each diagonal element $\omega_i = \text{Var}(\varepsilon_i)$. $HC0$ plugs in the squared residuals from the 2SLS regression, $\hat{\omega}_i = \hat{\varepsilon}_i^2$ with $\hat{\varepsilon}_i \equiv y_i - x_i'\hat{\beta}$. It is well known from OLS that these estimates tend to underestimate the true variance of $\varepsilon_i$ in finite samples since least square procedures choose the residuals such that the sum of squared residuals is minimized. It is most apparent for influential observations which "pull" the line toward itself and thereby make their residuals "too small". Since the influence of any single observation vanishes (under the assumptions stated above), $HC0$ is asymptotically valid. But in small samples, using the simple $HC0$ estimator tends to lead to (additional) bias in covariance matrix estimation. For OLS regressions, a set of alternative covariance estimators with improved finite sample properties is available (MacKinnon and White, 1985) with $HC1$ using a simple degrees of freedom correction and $HC2$ as well as $HC3$ aiming to control explicitly for the influence of observations.

In the case of 2SLS regressions, the issue is similar but involves two stages. An observation affects the regression line in the first stage and in the reduced form. The effect on the residual of the observation is ambiguous. In contrast to OLS, an observation might not pull the 2SLS regression line toward itself but push it away through the combined effect of the two stages. The next section derives and studies leverage and influence of observations before moving to the computation and interpretation of the various $HC$ covariance matrix estimators in 2SLS regressions.

### 3 2SLS Leverage and Influence

We first compute the 2SLS hat and residual maker matrix whose diagonal elements play a central role for the leverage of an observation. Then, the Frisch-Waugh-Lovell theorem is adapted to 2SLS regressions and used to derive the leverage and influence of observations.
3.1 2SLS Hat and Residual Maker Matrix

The fitted values of a 2SLS regression are given by

$$\hat{y} = X\hat{\beta} = X(X'PX)^{-1}X'Py \equiv Qy,$$

where $Q$ is defined as the 2SLS hat matrix. The 2SLS hat matrix involves both the regressors $X$ and the conventional projection (hat) matrix $P = Z(Z'Z)^{-1}Z'$ associated with the instruments $Z$. The residuals are given by

$$\hat{\varepsilon} = y - X\hat{\beta} = (I - Q)y \equiv Ny,$$

where $N$ is defined as the 2SLS residual maker matrix.

It is easy to verify that the 2SLS hat and residual maker matrix are idempotent but not symmetric like their OLS counterparts:

$$Q = X(X'PX)^{-1}X'P, \quad N = I - Q,$$

$$Q' = PX(X'PX)^{-1}X', \quad N' = I - Q',$$

$$QQ' = X(X'PX)^{-1}X', \quad NN' = I - PX(X'PX)^{-1}X'P = I - \tilde{Q},$$

where we have defined $\tilde{Q} = PX(X'PX)^{-1}X'P$. While the 2SLS hat matrix self-projects and the residual maker matrix annihilates $X$, the transposed 2SLS hat matrix self-projects and the transposed residual maker matrix matrix annihilates $PX$. The products of the matrices in the last line are used to compute covariance matrices (for fixed regressors and instruments) of the fitted values and residuals under conditional homoskedasticity:

$$Var(\hat{y}) = Var(Qy) = QVar(y)Q' = \sigma^2 \cdot QQ',$$

$$Var(\hat{\varepsilon}) = Var(N\varepsilon) = NVar(\varepsilon)N' = \sigma^2 \cdot NN' = \sigma^2 \cdot (I - \tilde{Q}). \quad (1)$$

Note that these matrices simplify in the exactly identified case of $K = L$ since $Z'X$ is a square matrix, $Q = X(Z'X)^{-1}Z'$ and $\tilde{Q} = P$, where $P$ is the conventional projection matrix of the instruments.

3.2 A Frisch-Waugh-Lovell Theorem for 2SLS

Split the instruments into two groups, $Z = (Z_1; Z_2)$, where $Z_2$ are predetermined regressors, and rewrite the regression equation as

$$y = X_1\beta_1 + Z_2\beta_2 + \varepsilon. \quad (2)$$

The normal equations are split accordingly,

$$Z_1'X_1\hat{\beta}_1 + Z_1'Z_2\hat{\beta}_2 = Z_1'y, \quad \text{and} \quad Z_2'X_1\hat{\beta}_1 + Z_2'Z_2\hat{\beta}_2 = Z_2'y.$$
Premultiply the first group of equations by \( Z_1(Z_1'Z_1)^{-1} \),
\[
P_1X_1\hat{\beta}_1 + P_1Z_2\hat{\beta}_2 = P_1y,
\]
where \( P_1 = Z_1(Z_1'Z_1)^{-1}Z_1' \) is the conventional (OLS) projection matrix associated with the first group of instruments \( Z_1 \). Next, use the 2SLS hat matrix associated with \( X_1 \) and \( Z_1 \), 
\[
Q_1 = X_1(X_1'P_1X_1)^{-1}X_1'P_1,
\]
and premultiply again,
\[
X_1\hat{\beta}_1 + Q_1Z_2\hat{\beta}_2 = Q_1y,
\]
since \( P_1 \) is idempotent, and plug \( X_1\hat{\beta}_1 \) into the second group of normal equations,
\[
Z_2'( -Q_1Z_2\hat{\beta}_2 + Q_1y) + Z_2'Z_2\hat{\beta}_2 = Z_2y,
\]
which yields
\[
\hat{\beta}_2 = (Z_2'N_1Z_2)^{-1}Z_2'N_1y,
\]
where \( N_1 \) is the 2SLS residual maker matrix associated with \( X_1 \) and \( Z_1 \). As a result, we have

**Theorem 1** The 2SLS estimate of \( \beta_2 \) from regression (2) with \( Z \) as instruments is numerically identical to the 2SLS estimate of \( \beta_2 \) from regression
\[
N_1y = N_1Z_2\beta_2 + \text{errors}
\]
with \( Z_2 \) as instruments.

**Proof.** Since the normal equation
\[
Z_2'N_1Z_2\hat{\beta}_2 = Z_2'N_1y
\]
is exactly identified and \( Z_2'N_1Z_2 \) is a square matrix, respectively, we can solve for \( \hat{\beta}_2 \) by premultiplying with the inverse, proving that the two estimates for \( \beta_2 \) are indeed numerically identical.

With this Frisch-Waugh-Lovell (FWL) theorem for 2SLS at hand, we are now set to compute the influence of single observations.

### 3.3 2SLS Leverage and Influential Analysis

The influence of an observation \( i \) on the 2SLS estimate \( \hat{\beta} \) can be measured as the difference of the estimate omitting the \( i \)-th observation \( \hat{\beta}^{(i)} \) and the original estimate. Instead of running a separate 2SLS regression on a sample with the observation dropped, one can derive a closed-form expression for the difference by including a dummy variable for the \( i \)-th observation \( d_i \) and applying the FWL theorem,\(^1\)

\[
y = X\beta + ad_i + \varepsilon
\]

\(^1\)See Davidson and MacKinnon (2004) for the case of OLS.
with instruments \((Z_i' d_i)\), and \(E[(Z_i' d_i)' \varepsilon] = 0\). Use the 2SLS hat and residual maker matrix associated with \((X_i' d_i)\) and \((Z_i' d_i)\), \(Q_D\) and \(N_D\), to express the regression as

\[
y = Q_D y + N_D y = X_{\hat{\beta}}^{(i)} + \hat{\alpha} d_i + N_D y.
\]

Premultiply with the regular 2SLS hat matrix \(Q\),

\[
Q y = X_{\hat{\beta}}^{(i)} + Q \hat{\alpha} d_i,
\]
as the transposed 2SLS residual maker annihilates \(P X, N_D' Q' = 0\), cancelling out the residual. Since \(Q y = X_{\hat{\beta}}\), we can rewrite the equation as

\[
X (\hat{\beta}^{(i)} - \hat{\beta}) = -Q \hat{\alpha} d_i.
\]

Applying the FWL theorem for 2SLS, the estimate of \(\alpha\) from the full 2SLS regression is numerically equivalent to running the 2SLS regression

\[
N y = N d_i \alpha + \text{errors}
\]
with \(d_i\) as instrument, which yields

\[
\hat{\alpha} = \frac{d_i' N y}{d_i' N d_i} = \frac{\hat{\varepsilon}_i}{1 - q_i},
\]

where \(q_i\) is defined as the \(i\)th diagonal element of the original 2SLS hat matrix \(Q\). Plug this expression into the equation,

\[
X (\hat{\beta}^{(i)} - \hat{\beta}) = -Q \frac{\hat{\varepsilon}_i}{1 - q_i} d_i,
\]

and premultiply with \((X' P X)^{-1} X' P\) to obtain the change of the coefficient due to observation \(i\),

\[
\hat{\beta}^{(i)} - \hat{\beta} = -\frac{\hat{\varepsilon}_i}{1 - q_i} (X' P X)^{-1} X' P d_i = -\frac{\hat{\varepsilon}_i}{1 - q_i} (X' P X)^{-1} X' Z (Z' Z)^{-1} z_i. \tag{3}
\]

This measure looks similar to the one of OLS, but the leverage of observation \(i\),

\[
q_i = x'_i (X' P X)^{-1} X' Z (Z' Z)^{-1} z_i,
\]
is the diagonal element of the 2SLS hat matrix \(Q\). The influence of an observation depends on both the residual \(\hat{\varepsilon}_i\) and the leverage \(q_i\) of the observation. The higher they are, the more likely an observation is influential. Note that since \(Q\) is idempotent, its trace is equal to the rank \((L \leq K)\) which implies \(\sum_i q_i = L\).

\(\text{Trace}(Z(Z' P Z)^{-1} Z' P) = \text{Trace}((Z' P Z)^{-1} Z' P Z) = \text{Trace}(I_L) = L\)
Consider the leverage in the most simple case of a constant and one nonconstant regressor and instrument,
\[ q_i = x'_i (Z'X)^{-1} z = \frac{1}{n} + \frac{(x_i - \bar{x})(z_i - \bar{z})}{\sum_j (x_j - \bar{x})(z_j - \bar{z})}. \]

In contrast to the OLS leverage, \( q_i \) is not necessarily larger than \( n^{-1} \), as the individual leverage depends on the sign of \( \text{Cov}(x_i, z_i) \) and the sign and magnitude of \( (x_i - \bar{x})(z_i - \bar{z}) \) (which might be of opposite sign). Hence, \( q_i \) can potentially be smaller than \( n^{-1} \) (and even negative) for some observations, pushing away the regression line as in the example above.

Leverage and influence are useful tools to investigate the impact of single observations in 2SLS regressions (cf. Hoaglin and Welsch, 1978, and others for OLS). The derived expressions can be used to compute further diagnostic measures such as the studentized residual \( t_i \)
\[ t_i = \frac{\hat{E}_i}{s \sqrt{1 - q_i}} \]
where \( s^2 = (\hat{\varepsilon}'\hat{\varepsilon}) / (n - L) \) and \( (1 - \hat{q}_i) \) is the \( i \)-th diagonal element of the matrix \( NN' = I - \hat{Q} \) as well as Cook’s distance (Cook, 1977) based on the \( F \)-test,
\[ D_i = \frac{(\hat{\beta} - \hat{\beta}^{(i)})(X'PX)(\hat{\beta} - \hat{\beta}^{(i)})}{Ls^2} = \frac{\hat{E}_i^2}{(1 - \hat{q}_i)^2 Ls^2}. \]

This measure tells us that the removal of data point \( i \) moves the 2SLS estimate to the edge of the \( z\% \)-confidence region for \( \beta \) based on \( \hat{\beta} \) where \( z \) is implicitly given by \( D_i = F(L, n - L, z) \). Leverage and influence play a crucial role in the computation and interpretation of the alternative covariance matrix estimators which we now turn to.

4 Some Heteroskedasticity-Consistent Covariance Matrix Estimators for IV Regressions

The asymptotically valid covariance matrix of 2SLS is given by
\[ \text{Var} \left( \hat{\beta} \right) = (X'PX)^{-1} X'P \Omega PX (X'PX)^{-1}, \]
where \( \Omega \) is a diagonal matrix with diagonal elements \( \omega_i = \text{Var}(\varepsilon_i) \). In the case of conditional homoskedasticity \( \Omega = \sigma^2 \cdot I \), the covariance matrix can be consistently estimated by
\[ \text{Var}_C(\hat{\beta}) = \hat{\sigma}^2 \cdot (X'PX)^{-1}, \]
which only requires a consistent estimate of the variance of the error term, \( \hat{\sigma}^2 \). In the case of conditional heteroskedasticity, the analyst faces the task of estimating every single diagonal element \( \omega_i = \text{Var}(\varepsilon_i^2) \) of \( \Omega \).

If conditional homoskedasticity cannot be assumed, the standard approach to estimate the covariance matrix consistently is to plug the individual squared residuals of the 2SLS regression

\[ p_i = x'_i (X'X)^{-1} x_i = 1/n + (x_i - \bar{x})^2 / \sum_j (x_j - \bar{x})^2. \]

\[ \text{For OLS the expression is } p_i = x'_i (X'X)^{-1} x_i = 1/n + (x_i - \bar{x})^2 / \sum_j (x_j - \bar{x})^2. \]
into the diagonal elements, \( \hat{\omega}_i = \hat{\varepsilon}_i^2 \),

\[
\sqrt{\text{Var}_{HC0}(\hat{\beta})} = (X'PX)^{-1}X'P\hat{\Omega}PX(X'PX)^{-1},
\]
with \( \hat{\Omega} = \text{diag}(\hat{\varepsilon}_1^2, \hat{\varepsilon}_2^2, \ldots, \hat{\varepsilon}_n^2) \), which produces the robust standard errors due to White (1980) and Eicker (1963) also called HC0.

As already mentioned above, this estimator does not account for the fact that residuals tend to be too small in finite samples. In order to see this for 2SLS, consider again the case of conditional homoskedasticity where the variance of the residuals (for fixed regressors and instruments; see equation 1) is

\[
\text{Var}(\hat{\varepsilon}) = \sigma^2 \cdot NN' = \sigma^2 \cdot (I - PX(X'PX)^{-1}X'P) = \sigma^2 \cdot (I - \hat{Q}),
\]
Recall that the 2SLS residual maker is not symmetric idempotent, \( NN' = (I - \hat{Q}) \neq N \), and we have defined \( \hat{Q} \equiv PX(X'PX)^{-1}X'P \). The diagonal elements \( \hat{q}_i \) of the matrix \( \hat{Q} \) sum to \( L \) given that \( \hat{Q} \) is idempotent and of rank \( L \). We can compute the expectation of the average squared residuals,

\[
E(\hat{\varepsilon}^2) = \frac{1}{n} \sum_{i=1}^{n} E(\hat{\varepsilon}_i^2) = \frac{1}{n} \sum_{i=1}^{n} \sigma^2 (1 - \hat{q}_i) = \frac{n - L}{n} \sigma^2,
\]
which shows that the average tends to underestimate the true variance of the error term. Analogously to OLS, an improved estimator for the standard error in finite samples is \( s^2 \equiv \hat{\varepsilon}'\hat{\varepsilon}/(n - L) \) in the case of conditional homoskedasticity. In the case of conditional heteroskedasticity, using the same degree of freedom correction to the single diagonal elements, \( \hat{\varepsilon}_i^2 (n/(n - L)) \), yields an improved robust estimator for the covariance matrix,

\[
\sqrt{\text{Var}_{HC1}(\hat{\beta})} = \frac{n}{n - L} (X'PX)^{-1}X'P\hat{\Omega}PX(X'PX)^{-1},
\]
which is referred to as HC1.

Not all residuals are necessarily biased equally. A residual of an observation with a high leverage is biased downward more given its influence on the regression line. Hence, one should inflate the residuals of high leverage observations more than of low leverage observations. In the case of no heteroskedasticity, we have seen that the expected squared residuals can be expressed as

\[
E(\hat{\varepsilon}_i^2) = \sigma^2 (1 - \hat{q}_i),
\]
with the diagonal elements of \( \hat{Q} \),

\[
\hat{q}_i = z_i (Z'Z)^{-1}Z'X(X'PX)^{-1}X'Z(Z'Z)^{-1}z_i.
\]
Hence, one could inflate the \( \hat{\varepsilon}_i^2 \) with \( (1 - \hat{q}_i)^{-1} \), which is referred to as an "almost unbiased" estimator by Horn, Horn and Duncan (1975) who suggested this adjustment for OLS. Using
these adjustments generates the HC2 estimator,

\[
\hat{\text{Var}}_{HC2}(\hat{\beta}) = (X'PX)^{-1}X'P\hat{\Omega}PX(X'PX)^{-1},
\]

where

\[
\hat{\Omega} = \text{diag}(\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_n) \quad \text{with} \quad \hat{\omega}_i \equiv \hat{\varepsilon}_i^2(1 - \hat{q}_i)^{-1},
\]

which consistently estimates the covariance matrix (as the leverage of a single observation vanishes asymptotically, \( \hat{q}_i \to 0 \) as \( n \to \infty \), under the assumptions stated at the beginning), yet adjusts residuals of high leverage observations in finite samples.

In the presence of conditional heteroskedasticity, observations with a large error variance tend to influence the estimation "very much", which suggests using an even stronger adjustment for high leverage observations. Such an estimator can be obtained by applying a jackknife method. The jackknife estimator (Efron, 1982) of the covariance matrix of \( \hat{\beta} \) is

\[
\hat{\text{Var}}_{JK}(\hat{\beta}) = ((n - 1)/n) \sum_{i=1}^{n} \left[ \hat{\beta}^{(i)} - (1/n) \sum_{j=1}^{n} \hat{\beta}^{(j)} \right] \left[ \hat{\beta}^{(i)} - (1/n) \sum_{j=1}^{n} \hat{\beta}^{(j)} \right]',
\]

Plugging the expression for the influence above (equation 3) into the covariance matrix estimator, after some considerable manipulations (analogous to MacKinnon and White, 1985)\(^6\) one obtains

\[
\hat{\text{Var}}_{JK}(\hat{\beta}) = ((n - 1)/n) (X'PX)^{-1} \left[ X'P\Omega^*PX - (1/n) X'P\omega^*\omega^*PX \right] (X'PX)^{-1}
\]

where

\[
\Omega^* \equiv \text{diag}(\omega_1^*, \omega_2^*, \ldots, \omega_n^*) \quad \text{with} \quad \omega_i^* \equiv \hat{\varepsilon}_i/(1 - q_i),
\]

and \( \omega^* \equiv (\omega_1^*, \omega_2^*, \ldots, \omega_n^*)' \). This expression usually is simplified by dropping the \((n - 1)/n\)-and the \(1/n\)-terms which vanish asymptotically and whose omission is conservative since the covariance matrix becomes larger in a matrix sense, yielding the HC3 estimator

\[
\hat{\text{Var}}_{HC3}(\hat{\beta}) = (X'PX)^{-1}X'P\Omega^*PX(X'PX)^{-1},
\]

which again is consistent, yet tends to adjust residuals of high leverage observations more than HC2.

Another related issue in covariance matrix estimation is the bias of White’s robust (HC0) errors if there is no or only little heteroskedasticity. Chesher and Jewitt (1987) showed that HC0 errors in OLS are biased downward if heteroskedasticity is only moderate. Consider an adaption of a simple OLS example by Angrist and Pischke (2009) to the IV context. Take our simplest IV regression but without a constant \((L = 1)\). Let \( s_x^2 = \sum_{i=1}^{n} z_i^2/n \) and \( s_{xz} = \sum_{i=1}^{n} x_i z_i/n \),

\(^6\)Given that for OLS \( \hat{\beta}^{(i)} - \hat{\beta} - [\hat{\varepsilon}_i/(1 - q_i)](X'X)^{-1} x_i \), replace \( \hat{\beta} \) with \( \hat{\delta} \), [\( \hat{\varepsilon}_i/(1 - q_i) \)] with [\( \hat{\varepsilon}_i/(1 - p_i) \)], \((X'X)^{-1} \) with \((Z'PZ)^{-1}\) and \( x_i \) with \((Pz_i)'\).
and note that $\tilde{Q} = P = z'(zz')^{-1}z'$ since the equation is exactly identified. The expectation of the conventional non-robust covariance matrix estimator (for fixed regressors and instruments) in the case of conditional homoskedasticity is

$$E\left[\text{Var}_C(\hat{\beta})\right] = E\left[\sigma^2 \cdot (xz'(zz')^{-1}xx')^{-1}\right] = \frac{\sigma^2 s_{zz}}{s_{zz}s_{xx}} \left[\frac{1}{n} \sum_{i=1}^{n} (1 - \hat{q}_i)\right] = \frac{\sigma^2 s_{zz}}{n s_{zz}s_{xx}} \left[1 - \frac{1}{n}\right],$$

as the 2SLS hat matrix $\tilde{Q} = P$ has trace 1. The bias due to the missing correction of degrees of freedom is small. The expectation of the robust White covariance estimator ($HC0$) in the case of conditional homoskedasticity is

$$E\left[\text{Var}_{HC0}(\hat{\beta})\right] = E\left[(xz')^{-1}z\hat{\Omega}(xz')^{-1}\right] = \frac{\sigma^2}{s_{xx}s_{zz}} \left[\frac{1}{n} \sum_{i=1}^{n} \hat{q}_i^2 (1 - \hat{q}_i)\right] = \frac{\sigma^2 s_{zz}}{ns_{xx}s_{zz}} \left[1 - \sum_{i=1}^{n} \hat{q}_i^2\right],$$

where the downward bias is larger if $\sum_{i=1}^{n} \hat{q}_i^2 > n^{-1}$ which is the case if $\exists i$ such that $\hat{q}_i > n^{-1}$ since $\hat{q}_i = p_i \geq n^{-1} \forall i$, that is if there is some variation in $z$. Hence, the downward bias from using traditional $HC0$ may be worse than from using the conventional non-robust covariance estimator in 2SLS regressions if there is conditional homoskedasticity, in line with the results for OLS. Using $HC1-3$ mitigates the problem. However, one should optimally compute and compare conventional non-robust as well as $HC0-3$ errors and investigate the role of influential observations using the 2SLS regression diagnostics derived in the last section. If there are influential observations and substantial differences between the various covariance estimates, one should err on the side of caution by using the most conservative estimate. In the next section we will compare the finite sample performance of the various covariance matrix estimators.

5 Parametric Monte Carlo Simulations

For OLS regressions, MacKinnon and White (1985) examined the performance of $HC0-3$ estimators and found that $HC3$ performs better than $HC1-2$ which in turn outperform the traditional White’s robust ($HC0$) estimator in all of their Monte Carlo experiments. Later OLS simulations by Long and Ervin (2000) confirmed these results. Angrist and Pischke (2009, Chapter 8) provided further support with a very simple and illustrative example of an OLS regression with one binary regressor. We first perform simulations for a simple IV regression model with one as well more than one continuous instruments, and then redo the simulations with binary instruments.
5.1 Continuous Instruments

Consider the simplest case of a linear model with one endogenous regressor and one instrument,

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \]

completing the system of equations with the first stage

\[ x_i = \pi_0 + \pi_1 z_i + v_i, \]

where \( z_i \in \mathbb{R} \) is a continuous instrument and \( v_i \) the error term of the first stage. The model is parameterized such that \( \beta_0 = 0, \beta_1 = 0, \pi_0 = 1, \) and \( \pi_1 = 5. \) While the instrument is valid, the true effect of the regressor on the dependent variable is zero. The data generating process is such that in each iteration observations are drawn from a standard normal distribution, \( z_i \sim N(0, 1) \). The error terms are then drawn from a joint normal distribution with the structural disturbance \( \varepsilon \) potentially being conditionally heteroskedastic,

\[
\begin{pmatrix}
\varepsilon_i \\
v_i
\end{pmatrix}
\sim N
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\alpha(z_i)^2 & \rho \alpha(z_i) \eta \\
\rho \alpha(z_i) \eta & \eta^2
\end{pmatrix},
\text{with } \alpha(z_i) = \sqrt{\alpha^2 + (1 - \alpha^2) z_i^2}.
\]

While we set \( \eta = 1 \) and \( \rho = 0.8 \), we run simulations with substantial, moderate and no heteroskedasticity, \( \alpha \in \{0.5, 0.85, 1\} \), respectively. The true standard deviation of the error is \( \alpha \) at \( z_i = 0 \), 1 at \( z_i = -1, 1 \), and increases in \( |z_i| \). We vary the sample size, analyzing a very small sample of \( n = 30 \), a moderate sample of \( n = 100 \), and larger samples of \( n = 200 \) and \( n = 500 \) observations, performing 25,000 replications for each sample size and heteroskedasticity regime.

The results of the simulations are presented in Table 1. The second column reports mean and standard deviation of \( \hat{\beta}_1 \) and the third and fourth columns mean and standard deviation of the error estimates, respectively. The last two columns show the empirical rejection rates for a nominal 5% and 1% two-sided \( t \)-test for the (true) hypothesis, \( \beta_1 = 0. \) As expected, tests based on the conventional non-robust error estimator lead to massive size distortions in heteroskedastic environments. In the case of substantial heteroskedasticity for the sample size of \( n = 30 \), the true null is rejected in \( \sim 20\% \) of the iterations instead of 5% (and \( \sim 9\% \) instead of 1%). Although White’s robust (\( HC0 \)) estimator mitigates the problem, the size distortion still is substantial with an empirical rejection rate of \( \sim 11\% \) (\( \sim 4\% \)). The degree of freedom correction of \( HC1 \) and leverage adjustments of \( HC2 \) and \( HC3 \) successively lower the distortion. Tests based on \( HC3 \) errors come closest to the nominal rate by a clear margin compared to the other estimators, especially \( HC0 \), yet inference is still somewhat too liberal in

\[ \text{The slope coefficient is chosen to be sufficiently high, } \pi_1 = 5, \text{ such that in small samples the probability of drawing a sample generating an estimate close to zero which leads to exploding estimates in the second stage is avoided.} \]
this highly heteroskedastic environment. While the average standard errors produced by more robust estimators are higher, the variability increases as well. Using a rule of thumb of taking the more conservative of the conventional and HC3 error reduces the distortion only slightly compared to taking HC3 alone.

**TABLE 1**

For a moderate level of heteroskedasticity ($\alpha = 0.85$), conventional standard errors perform similarly to White’s robust (HC0) errors, both leading to inference substantially too liberal. HC3 errors, on the other hand, lead to much smaller size distortions. Taking the higher of the conventional and HC3 error removes the distortion almost completely in this case. Finally, the case of no heteroskedasticity ($\alpha = 1$) confirms that using HC0 errors in smaller samples may lead to large size distortions in contrast to conventional non-robust errors if one has conditional homoskedasticity. HC3 removes the distortion by adjusting for the impact of high leverage observations on error estimation.

For $n = 100$, the distortions of tests based on HC0 errors are lower but still nonnegligible, while using HC3 leads to very small distortions. Table 2 reports the same results for larger sample sizes. In the case of $n = 200$, size distortions become small with the difference between HC0 and HC3 being less than 1%. With the larger sample size of $n = 500$ the leverage of single observations is washed out such that HC0-3 perform similarly well.

**TABLE 2**

Figure 2 illustrates the size distortions for the smallest sample size of $n = 30$, plotting the empirical size against various nominal size levels (1, 2, ..., 20%). The absolute size distortions increase with the nominal size for all estimators except HC3 where the distortions are small and relatively constant across nominal levels (as for conventional errors in the case of homoskedasticity). The graphs demonstrate that HC3 may perform well both in heteroskedastic and homoskedastic environments as opposed to HC0 in smaller samples and non-robust estimates in heteroskedastic environments.

**FIGURE 2**

Next, consider an extension of the basic model with three instruments, 

$$x_i = \pi_0 + \pi_1 z_{i,1} + \pi_2 z_{i,2} + \pi_3 z_{i,3} + u_i,$$

drawn from independent standard normal distributions, and all being excluded from the regres-
The true standard deviation of the structural disturbance may depend on the instruments,

\[ \alpha(z_i) = \sqrt{\alpha^2 + (1 - \alpha^2)} \parallel z_i \parallel^2, \]

where \( \parallel z_i \parallel = \sqrt{z_{i,1}^2 + z_{i,2}^2 + z_{i,3}^2} \) is the Euclidian norm of the instruments. All other parameters are the same as for the base case studied above. The results are reported in Table 3 and are similar to the ones with one continuous, normally distributed instrument. Some observations of the instruments tend to be far away of the center in an Euclidian sense, potentially leading to liberal HC0 estimates. We also computed the efficient GMM estimates (given overidentification). The size distortions are even larger than for tests in 2SLS regressions based on HC0 errors. As efficient GMM requires estimation of fourth moments to compute coefficient estimates, influential observations interfere both with coefficient vector and covariance matrix estimation in smaller samples, worsening size distortions.

### TABLE 3

In the simulations with normally distributed, continuous instruments the HC3 errors tend to perform very well and substantially better than White’s robust (HC0) errors in smaller sample IV regressions. The normal distribution of the instrument leads to a design with some high leverage observations interfering with robust error estimation.

#### 5.2 Binary Instruments

Let us redo the simulations with one binary instrument, \( z_i \in \{0, 1\} \), which is nonstochastic. We consider a highly unbalanced design where 90\% of the observations are "untreated", \( z_i = 0 \), and 10\% "treated", \( z_i = 1 \), as in Angrist and Pischke (2009), which allows us to explicitly compare our results to theirs for OLS. As above, the errors are drawn from a joint normal distribution with the structural disturbance \( \varepsilon \) now being conditionally heteroskedastic of the form

\[ \alpha(z_i) = \begin{cases} 
1 & \text{if } z_i = 1, \\
\alpha & \text{if } z_i = 0.
\end{cases} \]

As above, we run simulations with substantial, moderate and no heteroskedasticity, \( \alpha \in \{0.5, 0.85, 1\} \), respectively, and leave all other parameters the same.

The results shown in Table 4 for the small sample of \( n = 30 \) and the 5\% level are quantitatively similar to the one reported by Angrist and Pischke (2009) for OLS. In the case of substantial heteroskedasticity (\( \alpha = 0.5 \)), the conventional non-robust error estimator leads to substantial size distortions which are barely cured by White’s robust (HC0) estimator in this highly unbalanced design. The degree of freedom correction of HC1 does not help much,
either. On the other hand, HC2 reduce the size distortion substantially by around 5% (4% for the 1% level), while HC3 performs the best, bringing the empirical rate down by 10% (6%). Although much closer to the truth, HC3 inference is still too liberal in such a design. The rule of thumb of taking the higher of the conventional and HC3 error mitigates the distortion further but does not entirely remove it. For moderate heteroskedasticity (\( \alpha = 0.85 \)) conventional non-robust errors perform much better than White’s robust (HC0) errors and even better than HC3. The variability of the robust standard error estimates counteracts the higher average levels. However, taking the higher of the conventional and HC3 error removes the size distortion completely in this case. Finally, the case of \( \alpha = 1 \) shows again that using White’s robust errors may lead to large size distortions in homoskedastic models with small samples in contrast to conventional non-robust errors. HC3 substantially counteracts but does not entirely eliminate the distortion.

TABLE 4

As the distortions are driven by high leverage points, working with a more balanced design should mitigate these issues. Simulations with the most balanced case of 50% of the observations being untreated and 50% treated have shown that the size distortions are substantially smaller for all error estimators and across the different regimes of heteroskedasticity. As the potential for influential observations is minimized by the design, the choice of the error estimator is of lower consequence for inference even in smaller samples. The simulations with binary instruments confirm that the performance of robust errors is highly dependent on the design in addition to the sample size. HC3 errors perform much better than White’s robust (HC0) errors in smaller and less balanced IV regressions in line with the results for OLS. As high leverage observations are interfering with error estimation, using HC0 errors leads to larger size distortions in heteroskedastic as well as homoskedastic environments. When working with such designs, one should be very cautious, compute conventional as well as HC1-3 error estimates, use the most conservative estimate and complement inference with influential analysis and other diagnostics.

6 Application to Growth Regressions

In this section, we apply the alternative covariance matrix estimators to growth regressions with instruments using the data of Persson and Tabellini (1994) and Acemoglu, Johnson and Robinson (2001). As growth regressions use countries as units of observations, they are naturally subject to smaller sample issues and thus well suited to test the performance of the alternative error estimators and diagnostics.
Persson and Tabellini (1994) estimated the effect of inequality on growth. According to their theoretical framework, there should be a negative relationship between the two in democracies. In one of their settings, they worked with a cross section of 46 countries, splitting the sample into democracies (29 observations) and nondemocracies (17), and used three instruments for inequality: percentage of labor force participation in the agricultural sector, male life expectancy, and secondary-school enrollments. Table 6 reports the coefficient of MIDDLE (a measure of equality), the original non-robust as well as our computations of the \( HC0-3 \) error estimates, and corresponding \( t \)- and \( p \)-values. While the original estimates for the whole sample differ slightly from ours (potentially due to data or computational issues), the ones for democracies match. Using conventional (non-robust) errors, one finds a positive coefficient for democracies as predicted by Persson and Tabellini’s theory with a \( p \)-value of 2\%. When we use \( HC3 \) errors instead, significance is substantially reduced as the \( p \)-value increases to 8\%. The shift in significance due to using \( HC3 \) errors hints at the presence of influential observations. Panel a) of Figure 3 plots the leverages (\( q_i \)’s) against the squared residuals. Observations like Colombia (COL), Venezuela (VEN) and India (IND) combine a relatively high leverage with a large residual, influencing the regression substantially. Without appropriate adjustments to robust errors, the results may suggest too high a precision of the estimation. Also note that the leverage for Jamaica (JAM) is negative which can happen in IV regressions as discussed above.

TABLE 5

Finally, one of the most famous growth regressions using instrumental variables was performed by Acemoglu, Johnson and Robinson (2001) who exploited differences in European mortality rates to estimate the effect of institutions on economic performance. They used a sample of 64 countries that were ex-colonies and for which settler mortality data was available, and instrument an index for expropriation risk (measuring institutional quality) with European settler mortality. Despite the relatively small sample, our computations of \( HC0-3 \) only deviate negligibly from their reported error estimates. Panel b) of Figure 3 plotting the leverage against the squared residuals of the base case shows that the design is well-balanced. Compared to Persson and Tabellini (1994), the leverage of observations is much more bounded from above especially for observations with larger residuals.

FIGURE 3
7 Conclusion

In this paper, we developed basic algebraic concepts for IV regressions which were used to derive the leverage and influence of observations on the 2SLS estimate and compute alternative heteroskedasticity-consistent ($HC1$-$HC3$) estimators for the 2SLS covariance matrix. The performance of these concepts was evaluated in Monte Carlo simulations showing that size distortions are substantial when using White’s robust ($HC0$) errors in smaller and less balanced IV designs. An application to growth regressions showed that the significance level of an estimator can be decisively reduced by using $HC3$ in the presence of influential observations.

The results suggest guidelines for applied IV projects, supporting the use of $HC3$ instead of conventional White’s robust ($HC0$) errors especially in smaller, unbalanced data sets with influential observations, in line with earlier results on alternative heteroskedasticity-consistent estimators for OLS. The results also demonstrate the importance of analyzing leverage and influence of observations in smaller samples which can be done conveniently with the measures derived in the paper.
References


