The Stata Journal (2020)
20, Number 4, pp. 785–804 DOI: 10.1177/1536867X20976312

Fast leave-one-out methods for inference, model selection, and diagnostic checking

Federico Belotti
Department of Economics and Finance
University of Rome Tor Vergata
Rome, Italy
federico.belotti@uniroma2.it

Franco Peracchi
Department of Economics and Finance
University of Rome Tor Vergata
Rome, Italy
franco.peracchi@uniroma2.it

Abstract. In this article, we describe `jackknife2`, a new prefix command for jackknifing linear estimators. It takes full advantage of the available leave-one-out formula, thereby allowing for substantial reduction in computing time. Of special note is that `jackknife2` allows the user to compute cross-validation and diagnostic measures that are currently not available after `ivregress 2sls`, `xtreg`, and `xtivregress`.

Keywords: st0617, jackknife2, jackknife, heteroskedasticity-consistent standard errors, cross-validation, diagnostic checking, predictive residuals

1 Introduction

The jackknife (Quenouille 1956; Tukey 1958; Miller 1974; Efron 1982) is a method for assessing the accuracy of an estimator from data that are independently and identically distributed (i.i.d.) but not necessarily conditionally homoskedastic. Its basic idea is to exploit the information contained in the empirical distribution of the estimates computed from the n subsamples of size $n-1$ that can be obtained from a sample of size n by leaving out one data point at a time. The jackknife is known to work very well for linear estimators, such as ordinary least-squares (OLS) and instrumental-variables (IV) estimators, which are the workhorses of empirical research in a variety of fields. For these estimators, the jackknife may be implemented using simple formula for the effect of leaving out either one data point or one block of data points at a time. These leave-one-out (L1O) formula also represent the basis for other methods, including cross-validation (CV) procedures for model selection (Stone 1974, 1977) and diagnostic procedures for detecting heteroskedasticity, influential observations, and high-leverage points (Cook and Weisberg 1982).

The current Stata implementation of the jackknife is very general because it applies to both linear and nonlinear estimators. However, this generality comes at a cost in terms of computational speed when linear estimators are considered. For example, if one types `regress` *yvar xvar*`, vce(jackknife)`, Stata computes the jackknife estimate of the sampling variance of the OLS estimator by literally leaving out one observation at a time and then recomputing the OLS estimates for each of the n subsamples of $n-1$ observations. The same is true when using the `vce(jackknife)` option for the IV command `ivregress` or the panel versions of the OLS and IV commands, `xtreg` and

`xtivregress`, or when using the `jackknife` prefix command for statistics that are linear in the data. With "big data" (either a large sample size or many regressors), this way of implementing the jackknife causes unnecessarily long computing times and therefore restricts the applicability of the method to samples with at most a few thousand observations.

In this article, we introduce a new procedure for jackknifing linear estimators. Our procedure takes full advantage of the available L1O formula, thereby achieving substantial reductions in computing time. Because postestimation commands that implement CV and diagnostic procedures are currently available only after `regress`, we also extend these commands to `ivregress 2sls`, `xtreg`, and `xtivregress`. We hope that this will help promote a wider application of the jackknife and related methods in empirical research.

2 The basic L1O formula

This section presents the basic L1O formula for OLS and IV estimators, both for cross-sectional and panel data. The following sections then show how these formulas may be used for inference (section 3), model selection (section 4), and diagnostic checking (section 5).

2.1 Cross-sectional data

Let the random variable Y and the random vector $\mathbf{X}$ represent, respectively, the outcome of interest and a set of k regressors (including the constant term). We denote by $\mathbf{Y}$ the n-vector containing the observations on Y and by $\mathbf{X}$ the $n \times k$ matrix containing the observations on $\mathbf{X}$. We assume that $\mathbf{X}$ has full column rank $k < n$. We also denote by Y_i the ith element of $\mathbf{Y}$ and by $\mathbf{X}_i^\top$ the ith row of $\mathbf{X}$. Our parameter of interest is the unknown k-vector $\boldsymbol{\beta}$ in the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$, where $\mathbf{U}$ is an n-vector of unobservable regression errors.

OLS estimation

If there are no endogeneity problems, that is, the regressors are uncorrelated with the regression errors, an OLS regression of Y on $\mathbf{X}$ provides the standard way of estimating $\boldsymbol{\beta}$. The OLS estimate of $\boldsymbol{\beta}$ computed from the full sample is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, while the estimate computed by excluding the ith data point $(\mathbf{X}_i^\top, Y_i)$ is

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \frac{\mathbf{X}_i\left(Y_i - \mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}\right)}{1 - h_i} \qquad i = 1, \ldots, n \tag{1}$$

where h_i is the ith diagonal element of the "hat" matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ (see, for example, Peracchi [2001]). Because $\mathbf{H}$ is a projection matrix (that is, symmetric and idempotent), $0 \leq h_i \leq 1$. The k-vector $n(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})$, viewed as a function of $i = 1, \ldots, n$,

is called the sensitivity curve or empirical influence function (EIF) of OLS. The ith data point is said to be influential if the difference $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}$ is large in some norm. Notice that the influence of the ith data point on the OLS coefficient depends on both $\widehat{U}_i$ and h_i. If h_i is near one, then the ith data point is said to exert a high leverage.

IV estimation

If there are endogeneity problems, that is, the regressors are correlated with the regression errors, the available data on Y and $\mathbf{X}$ are generally insufficient to estimate $\boldsymbol{\beta}$ consistently. In this case, the IV method offers a solution provided one can find a set of $r \geq k$ valid instruments, namely, variables that are both exogenous (that is, uncorrelated with the regression errors) and relevant (that is, correlated with the regressors). We denote by $\mathbf{W}$ the $n \times r$ matrix containing the n observations on the r instruments and by $\mathbf{W}_i^\top$ the ith row of $\mathbf{W}$. We also assume that the matrix $\mathbf{W}^\top\mathbf{X}$ has full column rank $k \leq r$.

With $r = k$ instruments (the "exactly identified" case), the IV estimator of $\boldsymbol{\beta}$ is unique and is called a simple IV estimator. The simple IV estimate computed from the full sample is $\widetilde{\boldsymbol{\beta}} = (\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}^\top\mathbf{Y}$, while the estimate computed by excluding the ith data point $(\mathbf{X}_i^\top, \mathbf{W}_i^\top, Y_i)$ is

$$\widetilde{\boldsymbol{\beta}}_{(i)} = \widetilde{\boldsymbol{\beta}} - \left(\mathbf{W}^\top\mathbf{X}\right)^{-1} \frac{\mathbf{W}_i\left(Y_i - \mathbf{X}_i^\top\widetilde{\boldsymbol{\beta}}\right)}{1 - d_i} \qquad i = 1, \ldots, n \tag{2}$$

where $d_i = \mathbf{X}_i^\top(\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}_i$ is the ith diagonal element of the matrix $\mathbf{X}(\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}^\top$.

With $r > k$ instruments (the "overidentified" case), the number of IV estimators is infinite. By far the most popular among them is the two-stage least-squares (2SLS) estimator. The estimate computed from the full sample is $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{C}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{C}\mathbf{Y}$, where $\mathbf{C} = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top$ is an $n \times n$ matrix. Phillips (1977) showed that the estimate computed by excluding the ith data point is

$$\widetilde{\boldsymbol{\beta}}_{(i)} = \widetilde{\boldsymbol{\beta}} - \mathbf{P}^{-1}\mathbf{R}_i \qquad i = 1, \ldots, n \tag{3}$$

with

$$\begin{aligned}
\mathbf{P} &= \mathbf{X}^\top \mathbf{C} \mathbf{X} \\
\mathbf{R}_i &= \mathbf{J}_i \left\{ \left(Y_i - \widehat{\mathbf{X}_i}^\top \widetilde{\boldsymbol{\beta}} \right) - \left(Y_i - \mathbf{W}_i^\top \widehat{\boldsymbol{\pi}} \right) \right\} - \left(\mathbf{J}_i + \mathbf{K}_i \right) \left(Y_i - \mathbf{X}_i^\top \widetilde{\boldsymbol{\beta}} \right) \\
\mathbf{J}_i &= \frac{m_i}{e_i f_i} \widehat{\mathbf{V}}_i + \frac{\widehat{\mathbf{V}}_i^\top \mathbf{P}^{-1} \mathbf{X}_i}{e_i f_i} \mathbf{X}_i \\
\mathbf{K}_i &= \frac{\widehat{\mathbf{V}}_i^\top \mathbf{P}^{-1} \mathbf{X}_i}{e_i f_i} \widehat{\mathbf{V}}_i - \frac{1}{e_i} \mathbf{X}_i \\
e_i &= m_i + \frac{\left(\widehat{\mathbf{V}}_i^\top \mathbf{P}^{-1} \mathbf{X}_i \right)^2}{f_i} \\
f_i &= 1 - c_i + \widehat{\mathbf{V}}_i^\top \mathbf{P}^{-1} \widehat{\mathbf{V}}_i
\end{aligned}$$

where $\widehat{\boldsymbol{\pi}} = (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{Y}$ is the r-vector of coefficients from the "reduced-form" OLS regression of $\mathbf{Y}$ on the instruments in $\mathbf{W}$, $\widehat{\mathbf{X}_i} = \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}_i$ and $\widehat{\mathbf{V}}_i = \mathbf{X}_i - \widehat{\mathbf{X}_i}$ are the k-vectors of fitted values and residuals for the ith unit from the "first-stage" OLS regressions of the k variables in $\mathbf{X}$ on the r instruments in $\mathbf{W}$, and $m_i = 1 - \mathbf{X}_i^\top \mathbf{P}^{-1} \mathbf{X}_i$ and $c_i = \mathbf{W}_i (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}_i^\top$ are the ith diagonal elements of the matrices $\mathbf{M} = \mathbf{I}_n - \mathbf{X} \mathbf{P}^{-1} \mathbf{X}^\top$ and $\mathbf{C}$.

2.2 Panel data

To simplify the notation and with little loss of generality, let us consider a balanced panel dataset in which n units are all observed at the same T time points. Our parameter of interest is the unknown k-vector $\boldsymbol{\beta}$ in the linear panel-data model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$, where now $\mathbf{Y}$ denotes the nT-vector containing the observations on Y, $\mathbf{X}$ denotes the $nT \times k$ matrix containing the observations on $\mathbf{X}$, and $\mathbf{U}$ denotes the nT-vector of regression errors. We denote by Y_{it} the generic element of $\mathbf{Y}$ and by $\mathbf{X}_{it}^\top$ the generic row of $\mathbf{X}$. A popular specification of the vector of regression errors is $\mathbf{U} = \alpha \otimes \iota_T + \epsilon$, where $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ is an n-vector of unknown unit-specific effects, $\otimes$ is Kronecker's product, ι_T is a T-vector with elements all equal to 1, and ϵ is an nT-vector of unobservable random errors. Endogeneity problems arise if either α or ϵ is correlated with the regressors.

Fixed-effects estimation

If only α is correlated with the regressors, the standard estimator of $\boldsymbol{\beta}$ in a linear panel-data model is the so-called fixed-effects (FE) estimator, which treats the unit-specific effects as additional parameters to estimate. The FE estimate computed from the full sample is $\widehat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*\top} \mathbf{X}^*)^{-1} \mathbf{X}^{*\top} \mathbf{Y}^*$, where $\mathbf{Y}^*$ is the nT-vector with generic element $Y_{it}^* = Y_{it} - \overline{Y}_i$, $\mathbf{X}^*$ is the $nT \times k$ matrix with generic row $\mathbf{X}_{it}^{*\top} = (\mathbf{X}_{it} - \overline{\mathbf{X}}_i)^\top$, $\overline{Y}_i = T^{-1} \sum_{t=1}^T Y_{it}$, and $\overline{\mathbf{X}}_i = T^{-1} \sum_{t=1}^T \mathbf{X}_{it}$. Banerjee and Frees (1997) showed that

the estimate computed by excluding the block of T observations $[\mathbf{X}_i, \mathbf{Y}_i]$ on the ith unit is

$$\widehat{\boldsymbol{\beta}}^*_{(i)} = \widehat{\boldsymbol{\beta}}^* - \left(\mathbf{X}^{*\top}\mathbf{X}^*\right)^{-1} \mathbf{X}_i^{*\top} \left(\mathbf{I}_T - \mathbf{H}_i^*\right)^{-1} \left(\mathbf{Y}_i^* - \mathbf{X}_i^* \widehat{\boldsymbol{\beta}}^*\right) \qquad i = 1, \ldots, n \tag{4}$$

where $\mathbf{H}_i^* = \mathbf{X}_i^* (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1} \mathbf{X}_i^{*\top}$ and $[\mathbf{X}_i^*, \mathbf{Y}_i^*]$, respectively, are the $T \times T$ diagonal block of the matrix $\mathbf{X}^* (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1} \mathbf{X}^{*\top}$ and the $T \times (k+1)$ submatrix of $[\mathbf{X}^*, \mathbf{Y}^*]$ corresponding to the ith unit.

Fixed-effects IV estimation

If α and ϵ are both correlated with the regressors but one can find a set of $r \geq k$ valid instruments, a consistent estimator of $\boldsymbol{\beta}$ is the so-called fixed-effects instrumental-variables (FE-IV) estimator, which is the IV estimator for the transformed model where the unit-specific effects are eliminated by taking deviations of all variables from their unit-specific means over the T periods.

When $r = k$, the FE-IV estimator of $\boldsymbol{\beta}$ is unique and is called a simple FE-IV estimator. The simple FE-IV estimate computed from the full sample is $\widetilde{\boldsymbol{\beta}}^* = (\mathbf{W}^{*\top}\mathbf{X}^*)^{-1}\mathbf{W}^{*\top}\mathbf{Y}^*$, where $\mathbf{W}^*$ is the $nT \times r$ matrix with generic row $\mathbf{W}_{it}^{*\top} = (\mathbf{W}_{it} - \overline{\mathbf{W}}_i)^\top$ and $\overline{\mathbf{W}}_i = T^{-1} \sum_{t=1}^T \mathbf{W}_{it}$, while the estimate computed by excluding the block of T observations $[\mathbf{X}_i, \mathbf{W}_i, \mathbf{Y}_i]$ is

$$\widetilde{\boldsymbol{\beta}}^*_{(i)} = \widetilde{\boldsymbol{\beta}}^* - \left(\mathbf{W}^{*\top}\mathbf{X}^*\right)^{-1} \mathbf{W}_i^{*\top} \left(\mathbf{I}_T - \mathbf{D}_i^*\right)^{-1} \left(\mathbf{Y}_i^* - \mathbf{X}_i^* \widetilde{\boldsymbol{\beta}}^*\right) \qquad i = 1, \ldots, n \tag{5}$$

where $\mathbf{D}_i^* = \mathbf{X}_i^* (\mathbf{W}^{*\top}\mathbf{X}^*)^{-1} \mathbf{W}_i^{*\top}$ and $[\mathbf{X}_i^*, \mathbf{W}_i^*, \mathbf{Y}_i^*]$, respectively, are the $T \times T$ diagonal block of the matrix $\mathbf{X}^* (\mathbf{W}^{*\top}\mathbf{X}^*)^{-1} \mathbf{W}^\top$ and the $T \times (k+r+1)$ submatrix of $[\mathbf{X}^*, \mathbf{W}^*, \mathbf{Y}^*]$ corresponding to the ith unit.

When $r > k$, a popular FE-IV estimator is FE-2SLS. Assuming that the $nT \times r$ instrument matrix $\mathbf{W}$ has full column rank, the FE-2SLS estimate of $\boldsymbol{\beta}$ computed from the full sample is $\widetilde{\boldsymbol{\beta}}^* = (\mathbf{X}^{*\top}\mathbf{C}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{C}^*\mathbf{Y}^*$, where $\mathbf{C}^* = \mathbf{W}^*(\mathbf{W}^{*\top}\mathbf{W}^*)^{-1}\mathbf{W}^{*\top}$ is an $nT \times nT$ matrix, while the estimate computed by excluding the block $[\mathbf{X}_i, \mathbf{W}_i, \mathbf{Y}_i]$ of T observations on the ith unit is

$$\widetilde{\boldsymbol{\beta}}^*_{(i)} = \widetilde{\boldsymbol{\beta}}^* - (\mathbf{P}^*)^{-1} \mathbf{R}_i^* \qquad i = 1, \ldots, n \tag{6}$$

with

$$\begin{aligned}
\mathbf{P}^* &= \mathbf{X}^{*\top}\mathbf{C}^*\mathbf{X}^* \\
\mathbf{R}_i^* &= \mathbf{J}_i^{*\top} \left\{ \left(\mathbf{Y}_i^* - \widehat{\mathbf{X}}_i^* \widetilde{\boldsymbol{\beta}}^*\right) - \left(\mathbf{Y}_i^* - \mathbf{W}_i^* \widehat{\boldsymbol{\pi}}^*\right) \right\} - \left(\mathbf{J}_i^* + \mathbf{K}_i^*\right)^\top \left(\mathbf{Y}_i^* - \mathbf{X}_i^* \widetilde{\boldsymbol{\beta}}^*\right) \\
\mathbf{J}_i^* &= (\mathbf{E}_i^*)^{-1} \mathbf{M}_i^* (\mathbf{F}_i^*)^{-1} \widehat{\mathbf{V}}_i^* + (\mathbf{E}_i^*)^{-1} \widehat{\mathbf{V}}_i^* (\mathbf{P}^*)^{-1} \mathbf{X}_i^{*\top} (\mathbf{F}_i^*)^{-1} \mathbf{X}_i^* \\
\mathbf{K}_i^* &= (\mathbf{E}_i^*)^{-1} \widehat{\mathbf{V}}_i^* (\mathbf{P}^*)^{-1} \mathbf{X}_i^{*\top} (\mathbf{F}_i^*)^{-1} \widehat{\mathbf{V}}_i^* - (\mathbf{E}_i^*)^{-1} \mathbf{X}_i^* \\
\mathbf{E}_i^* &= \mathbf{M}_i^* + \widehat{\mathbf{V}}_i^* (\mathbf{P}^*)^{-1} \mathbf{X}_i^{*\top} (\mathbf{F}_i^*)^{-1} \mathbf{X}_i^* (\mathbf{P}^*)^{-1} \widehat{\mathbf{V}}_i^{*\top} \\
\mathbf{F}_i^* &= \mathbf{I}_T - \mathbf{C}_i^* + \widehat{\mathbf{V}}_i^* (\mathbf{P}^*)^{-1} \widehat{\mathbf{V}}_i^{*\top}
\end{aligned}$$

where $\widehat{\boldsymbol{\pi}}^* = (\mathbf{W}^{*\top}\mathbf{W}^*)^{-1}\mathbf{W}^{*\top}\mathbf{Y}^*$ is the r-vector of coefficients from the "reduced-form" OLS regression of the demeaned $\mathbf{Y}^*$ on the demeaned instruments in $\mathbf{W}^*$, $\widehat{\mathbf{X}}_i^* = \mathbf{X}^{*\top}\mathbf{W}^*(\mathbf{W}^{*\top}\mathbf{W}^*)^{-1}\mathbf{W}_i^{*\top}$ and $\widehat{\mathbf{V}}_i^* = \mathbf{X}_i^* - \widehat{\mathbf{X}}_i^{*\top}$ are the $T \times k$ matrices of fitted values and residuals for the ith unit from the "first-stage" OLS regressions of the k-demeaned variables in $\mathbf{X}^*$ on the r-demeaned instruments in $\mathbf{W}^*$, and $\mathbf{M}_i^* = \mathbf{I}_T - \mathbf{X}_i^*(\mathbf{P}^*)^{-1}\mathbf{X}_i^{*\top}$ and $\mathbf{C}_i^* = \mathbf{W}_i^*(\mathbf{W}^{*\top}\mathbf{W}^*)^{-1}\mathbf{W}_i^{*\top}$ are the $T \times T$ diagonal blocks of the matrices $\mathbf{M}^* = \mathbf{I}_{nT} - \mathbf{X}^*(\mathbf{P}^*)^{-1}\mathbf{X}^{*\top}$ and $\mathbf{C}^*$ corresponding to the ith unit.

3 Inference

Monte Carlo experiments (MacKinnon and White 1985) and theoretical calculations (Chesher and Jewitt 1987) show that conventional heteroskedasticity-consistent (HC) estimates of the OLS variance matrix can be severely downward biased in finite samples, particularly in the presence of high-leverage points, leading to overrejection of statistical hypotheses of interest (Chesher 1989). Young (2020) documents similar problems for inference based on conventional HC estimates of variance in the IV case. For both OLS and IV, the available evidence shows that inference based on the jackknife estimate of variance is more accurate. In addition, IV estimators are known to be biased in finite samples. Here, again, the jackknife can help by reducing the order of magnitude of the bias.

3.1 Estimating sampling variability

The jackknife estimate of the sampling variance of a k-dimensional estimator $\widehat{\boldsymbol{\theta}}$ is defined as

$$\widehat{\mathbb{V}}_{\mathrm{J}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}_{(\cdot)}\right)\left(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}_{(\cdot)}\right)^{\top} \tag{7}$$

where $\widehat{\boldsymbol{\theta}}_{(i)}$ is the ith L1O estimate and $\widehat{\boldsymbol{\theta}}_{(\cdot)} = n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{\theta}}_{(i)}$ is the average of the ith L1O estimate.

It follows from (1) that the jackknife estimate of the sampling variance of the OLS estimator is

$$\frac{n-1}{n}\,\mathbf{P}^{-1}\left\{\sum_{i=1}^{n}\left(\mathbf{R}_i - \overline{\mathbf{R}}\right)\left(\mathbf{R}_i - \overline{\mathbf{R}}\right)^{\top}\right\}\left(\mathbf{P}^{\top}\right)^{-1} \tag{8}$$

where $\mathbf{P} = \mathbf{X}^{\top}\mathbf{X}$, $\mathbf{R}_i = \mathbf{X}_i(Y_i - \mathbf{X}_i^{\top}\widehat{\boldsymbol{\beta}})/(1-h_i)$, and $\overline{\mathbf{R}} = n^{-1}\sum_{i=1}^{n}\mathbf{R}_i$. Ignoring $\overline{\mathbf{R}}$ gives the estimate proposed by Horn, Horn, and Duncan (1975) and Hinkley (1977), while ignoring the denominator $1-h_i$ in $\mathbf{R}_i$ gives the conventional HC estimate, implemented in Stata with the option `robust` after the command `regress`.

Estimators based on the IV method only have moments up to order $r-k$, the number of overidentifying restrictions (see, for example, Davidson and MacKinnon [2007]). In

particular, a simple IV estimator has no moments.[1] When second moments do not exist, jackknife estimates of variance need to be properly interpreted as estimating the asymptotic variance divided by n (see, for example, Shao and Wu [1989]). Of course, the same note of caution applies to conventional HC estimates of variance.

From (2), the jackknife estimate of variance for a simple IV estimator with $k = r$ has the same form as (8) with $\mathbf{P} = \mathbf{W}^\top \mathbf{X}$ and $\mathbf{R}_i = \mathbf{W}_i(Y_i - \mathbf{X}_i^\top \widetilde{\boldsymbol{\beta}})/(1 - d_i)$. Ignoring the term $1 - d_i$ in $\mathbf{R}_i$ gives the conventional HC estimate, implemented in Stata with the option `robust` after the command `ivregress 2sls`. In the case of overidentified 2SLS estimators, the jackknife estimate of variance for a 2SLS estimator has the same form as (8) with $\mathbf{P} = \mathbf{P}^*$ and $\mathbf{R}_i = \mathbf{R}_i^*$, where $\mathbf{P}^*$ and $\mathbf{R}_i^*$ are defined after (3).

From (4), the jackknife estimate of variance for an FE estimator has the same form as (8) with $\mathbf{P} = \mathbf{X}^{*\top}\mathbf{X}^*$ and $\mathbf{R}_i = \mathbf{X}_i^{*\top}(\mathbf{I}_T - \mathbf{H}_i^*)(\mathbf{Y}_i^* - \mathbf{X}_i^* \widehat{\boldsymbol{\beta}}^*)$. Ignoring the matrix $\mathbf{I}_T - \mathbf{H}_i^*$ in $\mathbf{R}_i$ gives the so-called clustered standard errors (Stock and Watson 2008; Cameron and Miller 2015), implemented in Stata with the option `vce(cluster)` after the command `xtreg, fe`. A Monte Carlo comparison of inference based on jackknife and clustered standard errors is presented in section 7.2.

From (5), the jackknife estimate of variance for a simple FE-IV estimator has the same form as (8) with $\mathbf{P} = \mathbf{W}^{*\top}\mathbf{X}^*$ and $\mathbf{R}_i = \mathbf{W}_i^{*\top}(\mathbf{I}_T - \mathbf{D}_i^*)(\mathbf{Y}_i^* - \mathbf{X}_i^* \widetilde{\boldsymbol{\beta}}^*)$. Finally, from (6), the jackknife estimate of variance for an FE-2SLS estimator has the same form as (8) with $\mathbf{P} = \mathbf{P}^*$ and $\mathbf{R}_i = \mathbf{R}_i^*$, where $\mathbf{P}^*$ and $\mathbf{R}_i^*$ are defined after (6).

3.2 Correcting for bias

If $\widehat{\boldsymbol{\theta}}$ is a biased estimator of a population parameter $\boldsymbol{\theta}$, in the sense that $\mathbb{E}(\widehat{\boldsymbol{\theta}}) \neq \boldsymbol{\theta}$, the jackknife estimate of its (mean) bias is defined as

$$\widehat{\text{Bias}}_{\text{J}} = (n-1)\left(\widehat{\boldsymbol{\theta}}_{(\cdot)} - \widehat{\boldsymbol{\theta}}\right)$$

Suppose that $\widehat{\boldsymbol{\theta}}$ has a finite bias of order $1/n$; that is,

$$\mathbb{E}\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta} = \frac{\mathbf{b}_1}{n} + \frac{\mathbf{b}_2}{n^2} + \frac{\mathbf{b}_3}{n^3} + \cdots$$

with $\mathbf{b}_1 \neq \mathbf{0}$. Then, the bias of the jackknife bias-corrected estimator $\widetilde{\theta} = \widehat{\boldsymbol{\theta}} - \widehat{\text{Bias}}_{\text{J}}$ is

$$\mathbb{E}\left(\widetilde{\theta}\right) - \boldsymbol{\theta} = -\frac{\mathbf{b}_2}{n(n-1)} - \frac{(2n-1)\,\mathbf{b}_3}{n^2(n-1)^2} + \cdots$$

which is of the smaller order $1/n^2$. This argument relies on the existence of first moments, so it works only for overidentified 2SLS and FE-2SLS estimators with $r \geq k + 1$. Jackknife bias-corrected 2SLS estimators were first proposed by Owen and Phillips (1975). Related estimators have been proposed by Angrist, Imbens, and Krueger (1999),

1. We thank the anonymous referee for stressing this point.

Blomquist and Dahlberg (1999), Hahn, Hausman, and Kuersteiner (2004), and Ackerberg and Devereux (2009). The actual performance of these estimators in finite samples typically depends on the strength of the instruments.

4 Model selection

Model selection is about choosing, from a given set of models, one that is best in terms of out-of-sample prediction. It differs from hypothesis testing, which is instead about deciding whether the available data support a particular model against some alternatives. The distinguishing features of model selection are the emphasis on predictive accuracy and the concern for overfitting.

A variety of model-selection criteria are available, including the adjusted R^2, Mallow's C_p (Mallows 1973), and information criteria such as the Akaike information criterion (Akaike 1973) and the Bayesian information criterion (Schwarz 1978). All of these criteria may be regarded as analytical approximations to measures of out-of-sample predictive risk.

An alternative approach, purely data driven, is CV. Its simplest version is sample splitting, which randomly divides the data in two halves, one used to fit a model (the "training set") and the other to assess predictive accuracy (the "validation set"). The mean squared error for the validation set provides an estimate of the mean squared prediction error (MSPE). Though easy to implement, sample splitting uses the data asymmetrically and inefficiently and tends to produce results that are highly variable.

An alternative method, K-fold CV, randomly divides the data into $K \leq n$ groups or folds of about equal size n/K. Then, it iteratively holds out one of the folds, fitting the data in the other $K-1$ folds and using the results to predict the outcomes in the held-out fold. Finally, it estimates the MSPE by averaging the prediction error over the K folds.

When $K = n$, this method is equivalent to holding out one observation at a time and then using the results to predict the held-out case. Because of this, n-fold CV is also known as leave-one-out cross-validation (L1OCV). The L1OCV criterion is defined as

$$\text{CV} = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_{(i)} \right)^2$$

where $\widehat{Y}_{(i)}$ is a predictor of Y_i that does not make use of Y_i. The L1OCV procedure selects the model with the smallest CV.

The L1OCV criterion may be used to choose an appropriate value for "tuning parameters" such as the number of regressors in a linear model fit by OLS or the number of instruments in an IV procedure. As argued by Varian (2014), "even if there is no tuning parameter, it is prudent to use CV to report goodness-of-fit measures because it measures out-of-sample performance, which is generally more meaningful than in-sample performance."

4.1 OLS and IV

Because $\widehat{Y}_{(i)} = \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_{(i)}$ for a linear model fit by OLS, the L1OCV criterion becomes

$$\text{CV} = \sum_{i=1}^{n} \left(\frac{\widehat{U}_i}{1 - h_i} \right)^2$$

Under the classical homoskedastic linear model $\mathbb{E}(\widehat{U}_i^2) = (1 - h_i)\sigma^2$, where σ^2 is the variance of a regression error,

$$\mathbb{E}\,(\text{CV}) = \sigma^2 \sum_{i=1}^{n} \frac{1}{1 - h_i}$$

If n is large enough and there are no high-leverage points, a first-order Taylor series expansion of $(1 - h_i)^{-1}$ about $h_i = 0$ gives

$$\mathbb{E}\,(\text{CV}) \approx \sigma^2 \sum_{i=1}^{n} (1 + h_i) = (n + k)\sigma^2$$

Thus, in this case, CV is an approximately unbiased estimator of the MSPE.

Similar criteria are easily constructed for simple IV, 2SLS, FE, FE-IV, or FE-2SLS estimates using (2)–(6).

5 Diagnostic checking

We focus on predictive residuals and measures of influence and leverage.

5.1 Predictive residuals

Predictive OLS residuals are defined as

$$\widehat{U}_{(i)} = Y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_{(i)} = \frac{\widehat{U}_i}{1 - h_i} \qquad i = 1, \ldots, n$$

The main advantage of predictive residuals is that they tend to give more emphasis to high-leverage points, because $\widehat{U}_{(i)} \geq \widehat{U}_i$ because $0 \leq h_i \leq 1$. Notice that predictive residuals are in fact ubiquitous, because they are a part of (1), the formula for the jackknife estimate of the sampling variance of OLS, and the L1OCV criterion for OLS. Also notice that the predictive residuals are related to the internally Studentized residuals $\widehat{U}_i^S = \widehat{U}_i / \sqrt{s^2(1 - h_i)}$, with $s^2 = (n - k)^{-1} \sum_{i=1}^{n} \widehat{U}_i^2$, which have approximately unit variance under the assumptions of the classical linear model. The externally Studentized residuals instead replace s^2 by $s_{(i)}^2 = (n - k - 1)^{-1} \sum_{j \neq i} \widehat{U}_{(i)}^2$.

Although Studentized residuals are defined only for OLS, predictive residuals are easily defined for all other estimators we consider. For IV and 2SLS, they are defined as

$$\widetilde{U}_{(i)} = Y_i - \mathbf{X}_i^\top \widetilde{\boldsymbol{\beta}}_{(i)} \qquad i = 1, \ldots, n$$

For FE, they are defined as

$$\widehat{\mathbf{U}}^*_{(i)} = \mathbf{Y}^*_i - \mathbf{X}^*_i \widehat{\boldsymbol{\beta}}^*_{(i)} \qquad i = 1, \ldots, n$$

while for FE-IV and FE-2SLS, they are defined as

$$\widetilde{\mathbf{U}}^*_{(i)} = \mathbf{Y}^*_i - \mathbf{X}^*_i \widetilde{\boldsymbol{\beta}}^*_{(i)} \qquad i = 1, \ldots, n$$

5.2 Measures of influence and leverage

To measure the overall influence of the ith observation on the OLS estimates, Cook (1977) proposed the index

$$D_i = \frac{\sum_{j=1}^n \left(\mathbf{X}_j^\top \widehat{\boldsymbol{\beta}}_{(i)} - \mathbf{X}_j^\top \widehat{\boldsymbol{\beta}}\right)^2}{k s^2} = \frac{h_i}{1-h_i} \frac{\left(\widetilde{U}_i^S\right)^2}{k} \qquad i = 1, \ldots, n$$

where $\widetilde{U}_i^S$ is the ith internally Studentized residual. The index D_i is proportional to the norm of the EIF of OLS in the metric of the matrix $\mathbf{X}^\top\mathbf{X}$. A large value of D_i indicates that the ith observation has a strong influence on the OLS estimate. Cook and Weisberg (1982) suggest choosing $D_i = 1$ as a cutoff. An extension of Cook's D-statistic to linear panel-data models was proposed by Banerjee and Frees (1997).

Notice that Cook's distance may be written as $D_i = (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^\top \widehat{\mathbb{V}}_{\text{OLS}}^{-1} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})/k$, where $\widehat{\mathbb{V}}_{\text{OLS}} = s^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is the classical estimate of the sampling variance of OLS, which assumes homoskedasticity. To avoid this assumption, we propose the following generalization,

$$D_i^J = \frac{1}{k} \left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right)^\top \widehat{\mathbb{V}}_{\text{J}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right) \qquad i = 1, \ldots, n \tag{9}$$

where $\widehat{\boldsymbol{\theta}}$ is any of our linear estimators and $\widehat{\mathbb{V}}_{\text{J}}$ is the jackknife estimate of their sampling variance.

5.3 Diagnostic plots

A leverage plot shows on the x axis the leverage measure $h_i/(1-h_i)$ and on the y axis the square of the internally Studentized residuals $\widetilde{U}_i^2$. These plots are very useful to detect the presence of outliers in the data and understand their nature but are not routinely produced by Stata.

6 The jackknife2 prefix command

`jackknife2` is a prefix command written using Mata. The basic `jackknife2` syntax, similar to the official `jackknife` prefix command, is as follows,

`jackknife2` [, `eif(`*filename*[, `replace`]) `hat(`*newvar*[, `replace`])
`fehat(`*filename*[, `replace`]) `presidual(`*newvar*[, `replace`])
`irstudent(`*newvar*[, `replace`]) `erstudent(`*newvar*[, `replace`])
`cooksd(`*newvar*[, `replace`]) `bpd(`*newvar*[, `replace`]) `dots(`#`) `nodots`] :
command

where *command* can be `regress`, `xtreg` with the `fe` option, `ivregress 2sls`, or `xtivreg` with the `fe` option. Only `pweight` and `iweight` are allowed, even if *command* supports other weight types.

`jackknife2` automatically computes the L1OCV criterion and the bias-corrected estimate. The latter, computed using the formula reported in section 3.2, is reported and stored in `e()`, while post diagnostics and measures of leverage are computed only when explicitly requested by the user through the corresponding options.

6.1 Options

`eif(`*filename*[, `replace`]) saves an Excel file (`.xls`) containing $n(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})$, the EIF of the estimator. `replace` specifies that it is okay to replace *filename* if it already exists.

`hat(`*newvar*[, `replace`]) generates a new variable containing the diagonal elements of the relevant projection ("hat") matrix. This option is available only when *command* is specified as `regress` or `ivregress 2sls`. `replace` specifies that it is okay to replace *newvar* if it already exists.

`fehat(`*filename*[, `replace`]) saves an Excel file (`.xls`) containing as many sheets as the number of diagonal blocks of the relevant projection ("hat") matrix. `replace` specifies that it is okay to replace *filename* if it already exists. This option is available only when *command* is specified as `xtreg, fe` or `xtivreg, fe`. Notice that this option can be very time consuming when the number of clusters is large.

`presidual(`*newvar*[, `replace`]) generates a new variable containing the predictive residuals. `replace` specifies that it is okay to replace *newvar* if it already exists.

`irstudent(`*newvar*[, `replace`]) generates a new variable containing the internally Studentized residuals. This option is available only when *command* is specified as `regress`. `replace` specifies that it is okay to replace *newvar* if it already exists.

`erstudent(`*newvar*[, `replace`]) generates a new variable containing the externally Studentized residuals. This option is available only when *command* is specified as `regress`. `replace` specifies that it is okay to replace *newvar* if it already exists.

`cooksd(`*newvar*[`, replace`]`)` generates a new variable containing the value of Cook's D-statistic (Cook 1977) and its extension to IV, 2SLS, or FE estimators. This option is available only when *command* is specified as `regress`, `ivregress 2sls`, or `xtreg, fe`. `replace` specifies that it is okay to replace *newvar* if it already exists.

`bpd(`*newvar*[`, replace`]`)` generates a new variable containing the generalization (9) of Cook's D-statistic. `replace` specifies that it is okay to replace *newvar* if it already exists.

`dots(#)` displays dots every # replications. `dots(0)` is a synonym for `nodots`.

`nodots` suppresses replication dots.

6.2 Implementation

Both `jackknife` and `jackknife2` are built around a loop consisting of n iterations, one for each sample unit (cluster), but differ in the way the L1O estimate $\widehat{\boldsymbol{\theta}}_{(i)}$ is computed at each iteration.

`jackknife` computes $\widehat{\boldsymbol{\theta}}_{(i)}$ at each iteration by running the appropriate estimation command, for example, `regress`, on the subsample with the ith unit (cluster) removed. After exiting the loop, it then computes the jackknife estimate of variance using (7). This is computationally expensive because it involves solving the k OLS normal equations n times.

`jackknife2` instead computes $\widehat{\boldsymbol{\theta}}_{(i)}$ at each iteration using the L1O formula, for example, (1) for OLS. Within the loop, it also accumulates the ingredients for the final computation of the jackknife estimates of variance and bias, the L1OCV criterion discussed in section 4, and the options listed in section 3.2. This substantially reduces the computational burden because the only heavy computation, for example, the inversion of $\mathbf{X}^\top\mathbf{X}$ for OLS, is performed just once and outside the loop. Further, only the diagonal elements of certain high-dimensional matrices are needed, not the full matrices. For example, in the case of OLS, only the diagonal elements $h_i = \mathbf{X}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i$ of the $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ are needed, not the full matrix. To reduce the computational burden, `jackknife2` also exploits (2) when $r = k$ and (3) when $r > k$.

7 Examples

7.1 Computing time: jackknife2 versus jackknife

In this section, we provide a comparison of the effective computing time needed for estimating jackknife standard errors using `jackknife2` and `jackknife`.

We consider the following data-generating process,

$$
\begin{aligned}
&Y_{it} = \alpha_i + \mathbf{X}_{it}^{\top}\beta + \epsilon_{1it} \\
&\begin{cases}
X_{1it}, \ldots, X_{kit} \text{ i.i.d. } \mathcal{N}(0,1), & \text{if } \rho = 0 \\
\\
X_{1it} = \delta_i + W_{1it}\gamma_1 + W_{2it}\gamma_2 + \epsilon_{2it}, \text{ and} & \\
X_{2it}, \ldots, X_{kit}, W_{1it}, W_{2it} \text{ i.i.d. } \mathcal{N}(0,1), & \text{if } \rho \neq 0
\end{cases} \\
&\begin{pmatrix} \epsilon_{1it} \\ \epsilon_{2it} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)
\end{aligned}
$$

with $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

This data-generating process encompasses all the cases covered by `jackknife2`, namely,

- `regress`: OLS estimator for cross-sectional data ($T = 1$) with $n = 10000$ or 1000000, $k = 10$ or 100 exogenous regressors, $\alpha_1 = \cdots = \alpha_n = 1$, and $\rho = 0$;[2]
- `ivregress 2sls`: 2SLS estimator for cross-sectional data ($T = 1$) with $n = 10000$ or 1000000, one endogenous regressor, $k - 1$ exogenous regressors ($k = 10$), two valid instruments, $\alpha_1 = \cdots = \alpha_n = 1$, $\delta_1 = \cdots = \delta_n = 1$, and $\rho \neq 0$;
- `xtreg, fe`: FE estimator for panel data with $T = 2$ or 10, $n = 10000$ or 1000000, $k = 10$ or 100 exogenous regressors, $\alpha_1, \ldots, \alpha_n$ i.i.d. $\mathcal{N}(0,1)$, and $\rho = 0$;
- `xtivreg, fe`: FE-2SLS estimator for panel data with $T = 2$ or 10, $n = 10000$ or 1000000, one endogenous regressor, $k-1$ exogenous regressors ($k = 10$), two valid instruments, $\alpha_1, \ldots, \alpha_n$ and $\delta_1, \ldots, \delta_n$ i.i.d. $\mathcal{N}(0,1)$, and $\rho \neq 0$.

When $\rho = 0$, $\beta_1, \ldots, \beta_k$ are all drawn independently from an $\mathcal{N}(0,1)$ distribution. When $\rho \neq 0$, $\beta_1 = 1$, $\beta_2, \ldots, \beta_k$ are drawn independently from an $\mathcal{N}(0,1)$ distribution, and $\gamma_1 = \gamma_2 = 0.5$. We consider 18 exercises (4 for `regress` and `xtivreg, fe`, 2 for `ivregress 2sls`, and 8 for `xtreg, fe`), each containing two sets of estimates, one for `jackknife` and one for `jackknife2`. We run all of them using Stata/MP8 15.1 on a x64 desktop with an Intel i7-7820X 8 Cores 3.60 GHz processor with 32 GB of RAM.

Results are reported in tables 1–4. The tables largely speak for themselves, showing substantial gains in computing time using `jackknife2`. When `jackknife2` is used as a prefix for the `regress` command, the estimation is up to 18,121 times faster compared with `jackknife` (this occurs when $n = 1000000$ and $k = 10$). A huge gain is obtained also for the case of the `ivregress 2sls` command, where the estimation is up to 148,162 times faster ($n = 1000000$ and $k = 10$). Similarly, the estimation is up to 507 times faster in the case of `xtivreg, fe` ($n = 10000$ $T = 2$ and $k = 10$), while smaller gains (around, on average, 37 times faster) are obtained when `jackknife2` is used with the `xtreg, fe` command.

2. The number of regressors, k, includes the constant term.

Table 1. `jackknife2` versus `jackknife` in the case of the `regress` command ($k = 10$ and $k = 100$)*

n	k	`jackknife`	`jackknife2`
10000	10	105.70	0.11
	100	577.85	1.52
1000000	10	127,101.31	7.01
	100	1,104,862.50	604.51

*Results are reported in seconds. Desktop x64 with Stata/MP8 15, Intel i7-7820X 8 Cores 3.60 GHz, 32 GB of RAM.

Table 2. `jackknife2` versus `jackknife` in the case of the `ivregress 2sls` command ($k = 10$)*

n	`jackknife`	`jackknife2`
10000	650.80	1.18
1000000	1,968,636.25	13.29

*See notes to table 1.

Table 3. `jackknife2` versus `jackknife` in the case of the `xtreg, fe` command ($k = 10$ and $k = 100$)*

n	T	`jackknife`	`jackknife2`
		$k = 10$	
10000	2	1,594.09	24.96
	10	4,263.13	91.04
100000	2	125,035.53	2,390.01
	10	332,554.56	9,706.91
		$k = 100$	
10000	2	4,103.94	133.18
	10	15,312.85	723.07
100000	2	292,890.03	12,033.85
	10	1,078,424.50	45,629.41

*See notes to table 1.

Table 4. `jackknife2` versus `jackknife` in the case of the `xtivreg, fe` command ($k = 10$)*

n	T	`jackknife`	`jackknife2`
10000	2	14,038.09	27.67
	10	52,015.40	108.26
100000	2	865,672.19	3,069.46
	10	3,096,145.75	18,277.38

*See notes to table 1.

7.2 Leave-one-panel-out: jackknife2 versus cluster

In this section, we carry out a small Monte Carlo study comparing the performance of the jackknife in estimating the sampling variance of the FE estimator (see section 3.1) with that of its direct competitor, the clustered estimator (Stock and Watson 2008), implemented in Stata with the option `vce(cluster)`. To our knowledge, this is the first time such a comparison has been made.

We consider the simple Gaussian linear panel-data model

$$Y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it}$$

where the logarithm of X_{it} is distributed as normal with mean α_i and unit variance. Notice that the lognormal distribution for X_{it} tends to generate isolated high-leverage points. As for the simulation of the unit-specific effects, we consider two cases: i) α_i distributed as standard normal; and ii) α_i distributed as the normal mixture $0.95 \times \mathcal{N}(0,1) + 0.05 \times \mathcal{N}(5, 0.25)$.

Finally, we compare the cases of homoskedastic and heteroskedastic errors. In the first case, the ϵ_{it}'s are generated as i.i.d. $\mathcal{N}(0,1)$ pseudo–random variables, while in the second case, they are generated as independent $\mathcal{N}(0, \sigma_{it}^2)$ pseudo–random variables with means zero and variance $\sigma_{it}^2 = 0.1 + 0.2X_{it} + 0.3X_{it}^2$. Note that the latter ensures substantial heteroskedasticity, especially when the α_i's are generated according to the aforementioned mixture model.

We investigate the effect of varying the cross-sectional dimension ($n = 1000$ or 10000) or the panel length ($T = 2$ or 10). Each experiment involves $M = 2000$ replications, and there are 16 experiments in total (one for each combination of n and T, separately for homoskedastic and heteroskedastic errors and the two different models for the unit-specific effects α_i).

For each replication, we compute two "quasi-t" statistics for testing the hypothesis that β is equal to 1. These statistics, denoted by "Clustered" and "Jackknife", exploit the covariance matrices after which they are named. For each experiment, we calculated the sample mean, standard deviation, skewness, and kurtosis (over the 2,000 replications) for both test statistics, but because there was nothing in the simulation

results suggesting that they had a nonzero mean, or that their distributions were not symmetric, we report only the standard deviation ("Std.dev.") and the kurtosis.[3] To investigate how often we will be led to make invalid inferences by using the considered test statistics, we report rejection frequencies ("5%") of the form $\widehat{q} = R/M$, where R is the observed number of rejections, that is, the number of times the test statistic exceeds the 1.96 critical value, and M is the number of replications.

Simulation results for all experiments are reported in tables 5 and 6. As in MacKinnon and White (1985), we find that almost all the test statistics have standard deviations greater than one, so that rejection frequencies based on them almost always exceed their 5% nominal size. As expected, these standard deviations tend to one as n or T increases. Interestingly, the distribution of the test statistics is close to standard normal when the errors are homoskedastic (table 5). Overall, the standard deviation and the kurtosis of the test statistic based on the clustered variance estimator exceed those of the statistic based on the jackknife variance. The difference between the two test statistics is striking, especially in the presence of heteroskedasticity and when the unit-specific effects are distributed as a normal mixture. Table 5 clearly shows that, even with moderate sample sizes ($n = 1000$ regardless of the panel length) and homoskedasticity, using the clustered variance estimator could easily lead to serious errors of inference. With $n = 1000$ and substantial heteroskedasticity, the jackknife also does not perform well. Its worst performance is when $n = 1000$, $T = 2$, and the distribution of the unit-specific effects is characterized by heteroskedasticity and outliers. In this case, the jackknife-based test incorrectly rejects the null hypothesis 9.7% of the time at the nominal 5% level. Still, it performs much better than its competitor because the clustered-based test rejects the null 22.2% of the time.

3. Standard deviation and kurtosis are equal to one and three, respectively, if the test statistic of interest is distributed as standard normal.

Table 5. Homoskedastic errors⋆

n	T	Clustered			Jackknife		
		Std.dev.	Kurtosis	5%†	Std.dev.	Kurtosis	5%†
$\alpha_i \sim \mathcal{N}(0,1)$							
1000	2	1.13	3.72	0.075*	1.04	3.42	0.062
	10	1.11	6.82	0.069*	1.01	3.19	0.052
10000	2	1.01	2.88	0.051	0.99	2.90	0.046
	10	1.02	3.10	0.052	1.00	3.13	0.049
$\alpha_i \sim 0.95 * \mathcal{N}(0,1) + 0.05 * \mathcal{N}(5,0.25)$							
1000	2	1.30	5.62	0.112*	1.07	4.26	0.065*
	10	1.22	4.17	0.105*	1.07	4.12	0.069*
10000	2	1.05	3.00	0.058	1.01	2.90	0.052
	10	1.04	3.23	0.056	1.01	3.23	0.049

⋆Numbers under Std.dev. and Kurtosis are the standard deviation and kurtosis of the quasi-t statistic.
†Numbers under 5% are the estimated rejection probabilities at this nominal level. An asterisk indicates they differ at the 1% level from what they should be if the quasi-t statistic was distributed as $\mathcal{N}(0,1)$.

Table 6. Heteroskedastic errors⋆

n	T	Clustered			Jackknife		
		Std.dev.	Kurtosis	5%	Std.dev.	Kurtosis	5%
$\alpha_i \sim \mathcal{N}(0,1)$							
1000	2	1.49	7.78	0.141*	1.15	3.69	0.064*
	10	1.36	4.09	0.132*	1.11	3.03	0.068*
10000	2	1.16	3.85	0.075*	1.06	2.90	0.056
	10	1.14	2.76	0.078*	1.06	2.68	0.056
$\alpha_i \sim 0.95 * \mathcal{N}(0,1) + 0.05 * \mathcal{N}(5,0.25)$							
1000	2	2.66	82.68	0.222*	1.34	11.61	0.097*
	10	1.59	5.68	0.179*	1.21	4.21	0.093*
10000	2	1.21	3.29	0.094*	1.07	2.96	0.062
	10	1.14	3.17	0.084*	1.04	3.02	0.057

⋆See notes to table 5.

8 Conclusions

Although the jackknife is potentially very useful, its current implementation in Stata is very general and also inefficient for linear estimators. In this article, we described the new prefix command `jackknife2`, which computes jackknife standard errors and other useful statistics, such as CV criteria, predictive residuals, and measures of influence and leverage, much faster than the official `jackknife` command. The new prefix command can be used when the model is fit via the `regress`, `ivregress 2sls`, `xtreg, fe`, and `xtivreg, fe` official Stata commands. We reported a comparison of the effective computing time needed for the estimation of the jackknife standard errors using `jackknife` and `jackknife2`, documenting the huge benefits in terms of computing time obtainable using the new prefix command. We also reported Monte Carlo evidence comparing the performance of the jackknife and its direct competitor, the clustered estimator, in estimating the sampling variance of the FE estimator.

9 Acknowledgments

We thank Roberto Rocci and Alwyn Young for useful discussions and an anonymous referee for very detailed comments. Franco Peracchi acknowledges financial support from MIUR PRIN 2015FMRE5X.

10 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 20-4
. net install st0617     (to install program files, if available)
. net get st0617         (to install ancillary files, if available)
```

11 References

Ackerberg, D. A., and P. J. Devereux. 2009. Improved JIVE estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics* 91: 351–362. https://doi.org/10.1162/rest.91.2.351.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, ed. B. N. Petrov and F. Csáki, 267–281. Budapest, Hungary: Akademiai Kiado.

Angrist, J. D., G. W. Imbens, and A. B. Krueger. 1999. Jackknife instrumental variables estimation. *Journal of Applied Econometrics* 14: 57–67. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1⟨57::AID-JAE501⟩3.0.CO;2-G.

Banerjee, M., and E. W. Frees. 1997. Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association* 92: 999–1005. https://doi.org/10.2307/2965564.

Blomquist, S., and M. Dahlberg. 1999. Small sample properties of LIML and jackknife IV estimators: Experiments with weak instruments. *Journal of Applied Econometrics* 14: 69–88. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1⟨69::AID-JAE521⟩3.0.CO;2-7.

Cameron, A. C., and D. L. Miller. 2015. A practitioner's guide to cluster–robust inference. *Journal of Human Resources* 50: 317–372. https://doi.org/10.3368/jhr.50.2.317.

Chesher, A. 1989. Hájek inequalities, measures of leverage and the size of heteroskedasticity robust Wald tests. *Econometrica* 57: 971–977. https://doi.org/10.2307/1913779.

Chesher, A., and I. Jewitt. 1987. The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica* 55: 1217–1222. https://doi.org/10.2307/1911269.

Cook, R. D. 1977. Detection of influential observation in linear regression. *Technometrics* 19: 15–18. https://doi.org/10.2307/1268249.

Cook, R. D., and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman & Hall.

Davidson, R., and J. G. MacKinnon. 2007. Moments of IV and JIVE estimators. *Econometrics Journal* 10: 541–553. https://doi.org/10.1111/j.1368-423X.2007.00221.x.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Hahn, J., J. Hausman, and G. Kuersteiner. 2004. Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *Econometrics Journal* 7: 272–306. https://doi.org/10.1111/j.1368-423X.2004.00131.x.

Hinkley, D. V. 1977. Jackknifing in unbalanced situations. *Technometrics* 19: 285–292. https://doi.org/10.2307/1267698.

Horn, S. D., R. A. Horn, and D. B. Duncan. 1975. Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association* 70: 380–385. https://doi.org/10.2307/2285827.

MacKinnon, J. G., and H. White. 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 29: 305–325. https://doi.org/10.1016/0304-4076(85)90158-7.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics* 15: 661–675. https://doi.org/10.2307/1267380.

Miller, R. G. 1974. The jackknife—A review. *Biometrika* 61: 1–15. https://doi.org/10.1093/biomet/61.1.1.

Owen, A. D., and G. D. A. Phillips. 1975. Bias reduction and approximate confidence intervals for the jackknifed 2SLS estimator. Paper presented to the World Congress of the Econometric Society, Toronto.

Peracchi, F. 2001. *Econometrics*. Chichester, UK: Wiley.

Phillips, G. D. A. 1977. Recursions for the two-stage least-squares estimators. *Journal of Econometrics* 6: 65–77. https://doi.org/10.1016/0304-4076(77)90055-0.

Quenouille, M. H. 1956. Notes on bias in estimation. *Biometrika* 43: 353–360. https://doi.org/10.1093/biomet/43.3-4.353.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464. https://doi.org/10.1214/aos/1176344136.

Shao, J., and C. F. J. Wu. 1989. A general theory for jackknife variance estimation. *Annals of Statistics* 17: 1176–1197. https://doi.org/10.1214/aos/1176347263.

Stock, J. H., and M. W. Watson. 2008. Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* 76: 155–174. https://doi.org/10.1111/j.0012-9682.2008.00821.x.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* 36: 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

———. 1977. Asymptotics for and against cross-validation. *Biometrika* 64: 29–35. https://doi.org/10.1093/biomet/64.1.29.

Tukey, J. W. 1958. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics* 29: 614. https://doi.org/10.1214/aoms/1177706647.

Varian, H. R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28: 3–28. https://doi.org/10.1257/jep.28.2.3.

Young, A. 2020. Consistency without inference: Instrumental variables in practical application. http://personal.lse.ac.uk/YoungA/CWOI.pdf.

About the authors

Federico Belotti is an associate professor of econometrics in the Department of Economics and Finance at the University of Rome Tor Vergata, Rome, Italy. His research interests are microeconometric theory and methods, health economics, and production economics.

Franco Peracchi is a professor of econometrics in the Department of Economics and Finance at the University of Rome Tor Vergata, Rome, Italy, and a professor of the practice at Georgetown University, Washington, DC. He is also a Fellow of the Einaudi Institute for Economics and Finance (EIEF), Rome, Italy. His research interests are econometric theory and methods, labor and health economics, and the economics of social security and pensions.