Computing KSS Clustered SEs

October 8, 2023

This vignette describes a MATLAB code that permits to compute the leave-out clustered SEs introduced by Kline, Saggio, and Sølvsten (2020) (henceforth KSS) in a linear regression model.

Contents

1 Introduction

Consider a regression equation of the form

$$y_{ij} = x_{ij}'\beta + \varepsilon_{ij} \qquad j = 1, ..., J; \qquad i = 1, ...n_j; \tag{1}$$

where i indexes a particular observation which belongs to a cluster j and we have $N = \sum_j M_j$ total observations; x_{ij} is a vector of regressors of dimension $K \times 1$ and y_{ij} is the outcome of interest. The error terms, ε_{ij}, are assumed to be heteroskedastic and potentially correlated across observations belonging to the same cluster j with a block-diagonal variance-covariance matrix given by

$$\Omega = \begin{bmatrix} \mathbf{\Omega}_1 & 0 & 0 & 0 \\ 0 & \mathbf{\Omega}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{\Omega}_J \end{bmatrix}$$

Cluster robust standard errors for the OLS estimator of β, $\hat{\beta}$, in most software packages (e.g `reghdfe`) is based on the following well-known formula

$$\hat{\mathbb{V}}[\hat{\beta}] = d \left(\sum_{i,j}^{N} x_{ij} x_{ij}' \right)^{-1} \left[\sum_{j=1}^{J} \left(\sum_{i}^{n_j} x_{ij} \hat{e}_{ij} \right) \left(\sum_{i}^{n_j} x_{ij} \hat{e}_{ij} \right)' \right] \left(\sum_{i,j}^{N} x_{ij} x_{ij}' \right)^{-1}, \tag{2}$$

where d is some degrees of freedom adjustment and $\hat{e}_{ij} = y_{ij} - x_{ij}\hat{\beta}$ is the OLS residual.[1]

KSS introduces an unbiased estimate of the variance of $\mathbf{\Omega}_j$ given by

$$\hat{\mathbf{\Omega}}_j = \mathbf{y}_j (\mathbf{y}_j - \mathbf{x}_j \hat{\beta}_{-j})', \tag{3}$$

where $\hat{\beta}_{-j}$ is the OLS estimate of β obtained after fitting (1) leaving cluster j out; $\mathbf{y}_j$ is a $n_j \times 1$ vector that stacks the outcome variable y_{ij} for observations belonging in cluster j; similarly, $\mathbf{x}_j$ is a $n_j \times K$ matrix that stacks the regressors x_{ij} for the observations belonging to cluster j.

Let $\hat{\eta}_{ij}$, represent the leave-cluster out residual, i.e. $\hat{\eta}_{ij} \equiv y_{ij} - x_{ij}\hat{\beta}_{-j}$. An unbiased estimate of the sampling variability of $\hat{\beta}$ is therefore

$$\hat{\mathbb{V}}[\hat{\beta}] = \left(\sum_{i,j}^{N} x_{ij} x_{ij}' \right)^{-1} \left[\sum_{j=1}^{J} \left(\sum_{i}^{n_j} x_{ij} y_{ij} \right) \left(\sum_{i}^{n_j} x_{ij} \hat{\eta}_{ij} \right)' \right] \left(\sum_{i,j}^{N} x_{ij} x_{ij}' \right)^{-1}, \tag{4}$$

The software described in this vignette, `KSS_SE`, permits to compute $\hat{\mathbb{V}}[\hat{\beta}]$. We illustrate how the software works in an example where one is interested in fitting an event-study based on two-way fixed effects regression.

[1] For instance, `reghdfe` sets $d = \frac{J}{J-1}\frac{N}{N-K}$.

2 Computing KSS Cluster-Robust SEs

The function `KSS_SE` is a MATLAB function that takes as input the following:

- `y`: outcome variable. Dimension is $N \times 1$.
- `D`: treatment(s) of interest. Dimension is $N \times N_D$.
- `clusterID`: variable that indexes clusters. Dimension is $N \times 1$.
- `indexID`: variable that indexes observations within clusters. Dimension is $N \times 1$.
- `controls`: additional controls. Dimension is $N \times N_P$.

The function `KSS_SE` computes the KSS leave-out standard errors on the regression coefficients associated with `D`. These SEs are clustered at the level defined by `clusterID`, after controlling for `controls`, `clusterID` fixed effects as well as `indexID` fixed effects.

We now demonstrate the functioning of `KSS_SE` in the context where one is interested in fitting an event study model of the form

$$y_{it} = \alpha_i + \lambda_t + \sum_{k=a}^{b} D_{it}^k \theta_k + X_{it}' \gamma + r_{it} \tag{5}$$

where α_i are, say, state fixed effects; λ_t are year fixed effects; R_{it} are event study indicators of the form $D_{it}^k \equiv \mathbf{1}\{t = t_i^* + k\}$ where t_i^* is the year in which the policy of interest is implemented in state i and X_{it} are some time-varying controls.

2.1 Building and Sending the Data To Matlab

We begin by loading the data in STATA

```
local mixtape https://raw.githubusercontent.com/Mixtape-Sessions

. use `mixtape'/Advanced-DID/main/Exercises/Data/ehec_data.dta, clear

. l in 1/5

     +-------------------------------------------------+
     |   stfips   year       dins   yexp2          W |
     |-------------------------------------------------|
  1. | alabama   2008   .6814122       .   613156 |
  2. | alabama   2009   .6580621       .   613156 |
  3. | alabama   2010   .6313651       .   613156 |
  4. | alabama   2011   .6563886       .   613156 |
  5. | alabama   2012   .6708115       .   613156 |
     +-------------------------------------------------+

. tab year

 Census/ACS |
survey year |      Freq.     Percent        Cum.
```

```
------------+-----------------------------------
       2008 |         46        8.33        8.33
       2009 |         46        8.33       16.67
       2010 |         46        8.33       25.00
       2011 |         46        8.33       33.33
       2012 |         46        8.33       41.67
       2013 |         46        8.33       50.00
       2014 |         46        8.33       58.33
       2015 |         46        8.33       66.67
       2016 |         46        8.33       75.00
       2017 |         46        8.33       83.33
       2018 |         46        8.33       91.67
       2019 |         46        8.33      100.00
------------+-----------------------------------
      Total |        552      100.00
```

```
. tab yexp2

    Year of |
   Medicaid |
  Expansion |      Freq.     Percent        Cum.
------------+-----------------------------------
       2014 |        264       73.33       73.33
       2015 |         36       10.00       83.33
       2016 |         24        6.67       90.00
       2017 |         12        3.33       93.33
       2019 |         24        6.67      100.00
------------+-----------------------------------
      Total |        360      100.00
```

This is a (balanced) state-year panel where the variable `dins` is the outcome of interest and `yexp2` measures the year in which Medicaid was expanded in a given state (it is missing for states that did not expand, like Alabama). Note that the panel runs from 2008 and 2019 and most states expanded in 2014.

We need to export to matlab the set of event study indicators. To do that, we set $a = -6$ and $b = 4$ in (5) and run the following few lines of code

```
rename yexp2 event_year
gen time_rel_event                = year-event_year
global lb = -6 // winsorize pre-event coefficients at -6
global ub = 4  // winsorize post-event coefficients at 4
replace time_rel_event  =$ub if time_rel_event>=$ub
& time_rel_event!=.
replace time_rel_event  =$lb if time_rel_event<=$lb
& time_rel_event!=.
qui forval k=$lb/$ub {
                local auxname=`k'-$lb
```

```
                gen     D`auxname'          = 0
                replace D`auxname'          = 1
if time_rel_event==`k' & treated == 1
}
local norma = -$lb - 1
```

Now that we have created the event study indicators, we can export the relevant information to MATLAB. Note that, to avoid multi-collinearity, we do not export D_{it}^{-1}, i.e. the event-study indicator relative to the year before medicaid expansion. Thus the event-study coefficients θ_k are all going to be expressed relative to θ_{-1}.

```
keep dins D* stfips year
order dins D* stfips year
drop D`norma' // to avoid collinearity issue.
export delimited using "data_MEDICAID.csv", replace novarnames nolabel
```

2.2 Importing the Data in Matlab

The GitHub Repo contains a matched employer-employee testing data where we observe the identity of the worker, the identity of the firm employing a given worker, the year in which the match is observed (either 1999 or 2001), and the associated log wage.

Important! The original data must be sorted by individual identifiers (id). For instance, one can see that the testing data is sorted by individual identifiers (and by year, using `xtset id year` in Stata)

[2]:
```
%% Import Data
namesrc='data/test.csv'; %path to original testing data
data=importdata(namesrc); %import data
id=data(:,1); %worker identifiers
firmid=data(:,2); %firm identifiers
y=data(:,4); % outcome variable
clear data
```

2.3 Calling the Main Function

The function `leave_out_KSS` relies on three mandatory inputs: `(y,id,firmid)`. We can obtain an unbiased variance decomposition of the associated AKM model by simply calling

[3]:
```
%% Run KSS!
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  = leave_out_KSS(y,id,firmid);
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
Algorithm for Computation of Statistical Leverages: JLA with 200 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7636
variance of wage: 0.1245
# of Movers: 6414
# of Firms: 1684
# of Person Year Observations: 56044
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
Running JLA Algorithm...
Done!
Elapsed time is 5.602229 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.019821
Covariance of Firm, Person Effects: -0.0039091
Variance of Person Effects: 0.10354
Correlation of Firm, Person Effects: -0.08629
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.010218
Covariance of Firm and Person Effects: 0.0047795
Variance of Person Effects: 0.085005
Correlation of Firm, Person Effects: 0.16217
```

3 Interpreting the Output

The code starts by printing its two key inputs: the algorithm used to compute the statistical leverages (exact vs. JLA) — we explain this distinction in Section **??** — and the level at which the leave-out correction is carried (observation vs. match) — we explain this in more detail in Section **??**.

The output printed by `leave_out_KSS` is composed of three sections.

Section 1: Here we provide info on the leave-out connected set. This is the largest connected set of firms that remains connected after any worker from the associated graph is removed, see Lemma 1 and the Computational Appendix of KSS for details. The code provides some summary statistics

(e.g. number of movers, number of firms, mean and variance of the outcome, etc.) of the leave-out connected set.

Section 2: After printing the summary statistics, the code computes the statistical leverages of the design, denoted by P_{ii}. Computation of $\{P_{ii}\}_{i=1}^{n}$ represents the main computational bottleneck of the routine.

Section 3: The code then enters its third, and final stage, where the main results are printed. The code starts by reporting the — biased — estimates of the variance components that result from the "plug-in" approach of treating OLS estimates as measured without error. Finally, the code prints the bias-corrected variance of firm effects and the covariance of worker and firm effects.

4 What Does the Code Save?

`leave_out_KSS` saves three scalars: the variance of firm effects (`sigma2_psi` in [4]), the covariance of worker and firm effects (`sigma_psi_alpha`), and the variance of person effects (`sigma2_alpha`).

`leave_out_KSS` also saves on disk one .csv file. This .csv contains information on the leave-out connected set. This file has 4 columns. First column reports the outcome variable, second and third columns the worker and the firm identifiers (as originally inputted by the user) and the fourth column reports the statistical leverages of the regression design. If the code is reporting a leave-out correction at the match-level, the .csv will be collapsed at the match level. By default, the .csv file is going to be saved in the main directory under the name `leave_out_estimates`. The user can specify an alternative path using the option `filename` when calling `leave_out_KSS`.

5 Scaling to Large Datasets

`leave_out_KSS` can be used on extremely large datasets. The code uses a variant of the random projection method, known as the Johnson– Lindenstrauss approximation (JLA) algorithm in KSS for its connection to the work of Johnson and Lindenstrauss (1984); see also Achlioptas (2003). We now discuss briefly the main computational bottleneck of the procedure and the JLA algorithm.

5.1 Computational Bottleneck

Recall from the discussion in Section 1 that the KSS leave-out bias correction procedure relies on leave-out estimates of σ_i^2,

$$\hat{\sigma}_i^2 = y_i(y_i - x_i'\hat{\beta}_{-i}), \tag{6}$$

where $\hat{\beta}_{-i}$ is the OLS estimate of β from the AKM model in equation (2) after leaving observation i out.

Clearly, reestimating $\hat{\beta}_{-i}$ by leaving a particular observation i for n times, is infeasible computationally. Fortunately, one can rewrite $\hat{\sigma}_i^2$ as

$$\hat{\sigma}_i^2 = y_i\frac{(y_i - x_i'\hat{\beta})}{1 - P_{ii}}, \tag{7}$$

where P_{ii} measures the influence or leverage of observation i, i.e., $P_{ii} = x_i' S_{xx}^{-1} x_i$. The above expression highlights that all that is needed for computation of $\hat{\sigma}_i^2$ are the n statistical leverages $\{P_{ii}\}_{i=1}^n$. However, exact computation of P_{ii} may remain prohibitive when n is in the order of tens of millions or higher.

5.2 Approximating the Statistical Leverages

The JLA algorithm introduced by KSS provides a stochastic approximation to $\{P_{ii}\}_{i=1}^n$ using the random projection ideas developed by Johnson and Lindenstrauss (1984). We refer the reader to the Computational Appendix of KSS for further details.

The number of simulations underlying the JLA algorithm is governed by the input `simulations_JLA` (which is denoted by p in the computational appendix). Intuitively, more simulations imply a higher accuracy – but also higher computation time — when estimating $\{P_{ii}, B_{ii}\}_{i=1}^n$.

Note: The user might want to prespecify a random-number generator seed to ensure replicability when calling the function `leave_out_KSS`.

We now demonstrate the performance of the code on a large dataset.

[4]:
```
%% Running KSS on a large dataset
websave('large_fake.csv', 'https://www.dropbox.com/s/ny5tef29ij7ran2/
 ↪large_fake_data.csv?dl=1'); %downloads and saves to disk a fake, large matched␣
 ↪employer employee data
namesrc='large_fake.csv'; %path to the large data
data=importdata(namesrc); %import data
id=data(:,1); %worker identifiers
firmid=data(:,2); %firm identifiers
y=data(:,4); % outcome variable
clear data
delete('large_fake.csv'); %delete original .csv data from disk

%Run Leave Out Correction (50 simulations)
type_of_algorithm='JLA'; %run random projection algorithm
simulations_JLA=50;
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  =␣
 ↪leave_out_KSS(y,id,firmid,[],[],type_of_algorithm,simulations_JLA);
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
Algorithm for Computation of Statistical Leverages: JLA with 50 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7304
variance of wage: 0.16248
# of Movers: 916632
# of Firms: 165360
# of Person Year Observations: 13860616
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
Running JLA Algorithm...
Done!
Elapsed time is 251.168051 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.039448
Covariance of Firm, Person Effects: 0.0084313
Variance of Person Effects: 0.080329
Correlation of Firm, Person Effects: 0.14978
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.030371
Covariance of Firm and Person Effects: 0.014639
Variance of Person Effects: 0.048803
Correlation of Firm, Person Effects: 0.38024
```

We can see from the output that the leave-out connected set has almost 14 million person-year observations. The code is able to complete in around 4 minutes (on a 2020 Macbook Pro with 6 cores and 16GB of RAM).

The computational appendix in KSS shows that the JLA algorithm can cut computation time by a factor of 100 while introducing an approximation error of roughly 10^{-4}.

The current code uses an improved estimator of both P_{ii} and $M_{ii} = 1 - P_{ii}$, which are both guaranteed to lie in $[0, 1]$. These improved estimators are then combined to derive an asymptotically unbiased JLA estimator of a given variance component provided that $\frac{n}{p^4} = o(1)$; see this document for further details..

We can check the stability of the estimates for different values of `simulations_JLA`. For instance, if we double `simulations_JLA` from 50 to 100 and run the code again on the same data:

[5]:
```
%% Compute estimates while doubling number of simulations
simulations_JLA=100;
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  =␣
 ↪leave_out_KSS(y,id,firmid,[],[],type_of_algorithm,simulations_JLA); %check␣
 ↪stability of variance components
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
Algorithm for Computation of Statistical Leverages: JLA with 100 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7304
variance of wage: 0.16248
# of Movers: 916632
# of Firms: 165360
# of Person Year Observations: 13860616
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
Running JLA Algorithm...
Done!
Elapsed time is 458.093692 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.039448
Covariance of Firm, Person Effects: 0.0084313
Variance of Person Effects: 0.080329
Correlation of Firm, Person Effects: 0.14978
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.030382
Covariance of Firm and Person Effects: 0.014598
Variance of Person Effects: 0.048738
```

```
Correlation of Firm, Person Effects: 0.37937
```

We obtain virtually the same variance components as when `simulations_JLA=50` while significantly increasing the computational time! If the user does not specify a value for `simulations_JLA`, the code defaults to `simulations_JLA=200`.

We conclude this section by noting that the user can also calculate an exact version of $\{P_{ii}\}_{i=1}^{n}$. This can be done by setting the option `type_of_algorithm` to `exact`.

Warning! Calling the option `exact` in large datasets can be very time-consuming! We now load again the original, smaller, testing data and then compare the exact and JLA-based estimates of the variance components,

[6]:
```
%% Compare Exact vs. JLA Estimates
namesrc='data/test.csv'; %path to original testing data
data=importdata(namesrc); %import data
id=data(:,1); %worker identifiers
firmid=data(:,2); %firm identifiers
y=data(:,4); % outcome variable
clear data

%Run Leave Out Correction with exact
type_of_algorithm='exact'; %run random projection algorithm;
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  =␣
 ↪leave_out_KSS(y,id,firmid,[],[],type_of_algorithm);

%Run Leave Out Correction with JLA
simulations_JLA=100;
type_of_algorithm='JLA'; %run random projection algorithm;
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  =␣
 ↪leave_out_KSS(y,id,firmid,[],[],type_of_algorithm,simulations_JLA);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
Algorithm for Computation of Statistical Leverages: Exact
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7636
variance of wage: 0.1245
# of Movers: 6414
# of Firms: 1684
# of Person Year Observations: 56044
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
Running Exact Algorithm...
Done!
Elapsed time is 162.242224 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.019821
Covariance of Firm, Person Effects: -0.0039091
Variance of Person Effects: 0.10354
Correlation of Firm, Person Effects: -0.08629
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.010289
Covariance of Firm and Person Effects: 0.0046293
Variance of Person Effects: 0.085204
Correlation of Firm, Person Effects: 0.15635
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
Algorithm for Computation of Statistical Leverages: JLA with 100 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7636
variance of wage: 0.1245
# of Movers: 6414
# of Firms: 1684
# of Person Year Observations: 56044
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
```

```
Running JLA Algorithm...
Done!
Elapsed time is 2.434287 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.019821
Covariance of Firm, Person Effects: -0.0039091
Variance of Person Effects: 0.10354
Correlation of Firm, Person Effects: -0.08629
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.01044
Covariance of Firm and Person Effects: 0.0044957
Variance of Person Effects: 0.085326
Correlation of Firm, Person Effects: 0.15063
```

The variance components estimated using JLA are extremely close to the `exact` estimates but only take a fraction of the time to compute. If the input data has more than 10,000 observations, the code defaults to using the JLA algorithm unless the user specifies type_of_algorithm as "exact".

6 Adding Controls

We have demonstrated the functioning of `leave_out_KSS` using a simple AKM model with no controls ($w_{gt} = 0$). It is easy to add a matrix of controls to the routine. Suppose for instance that we want to add year fixed effects to the original AKM model. This can be done as follows.

[7]:
```
%% How to add controls
namesrc='data/test.csv'; %path to original testing data
data=importdata(namesrc); %import data
id=data(:,1); %worker identifiers
firmid=data(:,2); %firm identifiers
year=data(:,3); %year identifier
y=data(:,4); % outcome variable
clear data

%Specify year fixed effects as controls
[~,~,controls] = unique(year);
controls            = sparse((1:size(y,1))',controls',1,size(y,1),max(controls));
controls       = controls(:,1:end-1); %to avoid collinearity issues, omit last␣
 ↪year fixed effects.

%Call KSS with matrix of controls
```

```
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  = leave_out_KSS(y,id,firmid,controls);
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
Algorithm for Computation of Statistical Leverages: JLA with 200 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7636
variance of wage: 0.1245
# of Movers: 6414
# of Firms: 1684
# of Person Year Observations: 56044
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
pcg converged at iteration 58 to a solution with relative residual 8.7e-11.
Calculating the statistical leverages of the AKM model...
Running JLA Algorithm...
Done!
Elapsed time is 4.271175 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.019479
Covariance of Firm, Person Effects: -0.004008
Variance of Person Effects: 0.10404
Correlation of Firm, Person Effects: -0.089031
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.0097812
Covariance of Firm and Person Effects: 0.0046239
Variance of Person Effects: 0.08578
Correlation of Firm, Person Effects: 0.15963
```

When controls are specified, the code proceeds by partialling them out. That is, it first estimates

by OLS the AKM model in the leave-out connected set

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + w'_{gt}\delta + \varepsilon_{gt} \tag{8}$$

from which we obtain $\hat{\delta}$. We then work with a residualized model where the outcome variable is now defined as $y_{gt}^{new} = y_{gt} - w'_{gt}\hat{\delta}$ and project this residualized outcome on worker and firm indicators and report the associated (bias-corrected) variance components.

7 Leaving Out a Person-Year Observation vs. Leaving Out a Match

By default, the code reports leave-out corrections for the variance of firm effects and the covariance of firm and worker effects that are robust to unrestricted heteroskedasticity and serial correlation of the error term within a given match (defined as the unique combination of the worker and firm identifier); see Remark 3 of KSS. Intuitively, leaving out matches is analogous to "clustering" the standard error estimates at the match level. Section **??** discusses the interpretation of the leave-out bias-corrected variance of person effects when leaving a match out

The user can specify the function to run the KSS correction when leaving only an observation out using the option `leave_out_level`. When the user leaves a person-year observation out, the resulting KSS variance components are robust to unrestricted heteroskedasticity but not to serial correlation within a match. Below we demonstrate how to compute KSS- adjusted variance components when leaving a single (person-year) observation out.

[8]:
```
%% Leaving out a Person-Year Observation vs. Leaving Out a Match

leave_out_level='obs'; %leave a single person-year observation out
[sigma2_psi,sigma_psi_alpha,sigma2_alpha]  =␣
↪leave_out_KSS(y,id,firmid,[],leave_out_level);
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave person-year observation out
Algorithm for Computation of Statistical Leverages: JLA with 200 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7636
variance of wage: 0.1245
# of Movers: 6414
# of Firms: 1684
# of Person Year Observations: 56044
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
```

```
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
Running JLA Algorithm...
Done!
Elapsed time is 4.700531 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.019821
Covariance of Firm, Person Effects: -0.0039091
Variance of Person Effects: 0.10354
Correlation of Firm, Person Effects: -0.08629
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.010306
Covariance of Firm and Person Effects: 0.0046087
Variance of Person Effects: 0.085253
Correlation of Firm, Person Effects: 0.15548
```

When $T = 2$ (i.e., the underlying matched employer-employee data spans only two years), as in this example, it turns out that the KSS-adjusted variance of firm effects and covariance of firm and worker effects are robust to any arbitrary correlation between ε_{g2} and ε_{g1}.

8 Variance of Person Effects When Leaving Out a Match

By leaving a match-out, we can bias-correct the variance of firm effects and the covariance of worker and firm effects while allowing for unrestricted hetoreskedasticity and serial correlation of the error term ε_{gt} within each worker-firm match.

However, the person effects, α_g, of "stayers" — workers that never leave a particular firm — are not leave-match-out estimable.[2] This implies that we cannot compute an unbiased estimate of $\Omega_g = Var(\varepsilon_{g1}, ..., \varepsilon_{gT_g})$ for stayers. An estimate of Ω_g for both stayers and movers is required in order to provide a bias-correction for the variance of person effects; see Section 1 and Remark 3 in KSS.

The current implementation of the code estimates Ω_g for stayers by leaving only a single observation out, that is, by assuming Ω_g is diagonal. This approach yields an upper bound estimate on the variance of person effects (computed across both stayers and movers).

There are several alternatives that the user can explore:

[2] This is because leaving a match-out means leaving *all* the observations associated with a stayer and therefore we cannot estimate her α_g.

1. Estimate a variance decomposition in a sample of movers only: For movers, it is possible to estimate a leave-out bias-corrected variance of person effects that is robust to both unrestricted heteroskedasticity and serial correlation in the error term of the AKM model within a given match. Therefore, one can provide an unbiased variance decomposition of all the three components of the two-way fixed effects model by simply feeding to the function `leave_out_KSS` a movers-only sample.
2. Drop adjacent wage observations for stayers: Under the assumption that the errors are serially independent after m periods, it suffices to keep every mth stayer observation and apply the estimator after leaving a person-year observation out. For example, if $m = 2$ and we have a balanced panel with $T = 5$, we can restore independence of the errors in the stayer sample by keeping any of the following pairs of stayer time periods: (1,4), (2,5), (1,5). One can choose randomly from the available pairs for each stayer with equal probability.
3. Drop interior wage observations for stayers: To minimize concerns regarding serial correlation, the user can drop all but the first and last wage observations of each stayer. Note that dropping stayer wage observations reduces their weight in the variance components. Future versions of the code will allow the variance components to be defined in terms of weights other than the number of micro-observations.

9 Regressing Firm Fixed Effects on Observables

It is common in empirical applications to regress the fixed effects estimated from the two-way model on some observable characteristics. Using the AKM model again as our leading example, suppose that we are interested in the linear projection of the firm effects ψ_{gt} on some observables Z_{gt}

$$\psi_{j(g,t)} = Z'_{gt}\gamma + e_{gt}. \tag{9}$$

The standard practice is to estimate γ using a simple regression where the estimated firm effects, $\hat{\psi}_{j(g,t)}$, are regressed on Z_{gt}

$$\hat{\gamma} = \left(\sum_{g,t} Z_{gt} Z'_{gt} \right)^{-1} Z_{gt} \hat{\psi}_{gt}. \tag{10}$$

KSS show that inference on $\hat{\gamma}$ needs to be adjusted because the estimated firm fixed effects $\{\hat{\psi}_j\}_{j=1}^{J}$ are correlated with one another.

To see this, suppose that we have a simple AKM model with only two time periods, set $w_{gt} = 0$, and take first differences $\Delta y_g \equiv y_{g2} - y_{g1}$ to eliminate the worker fixed effects so that the AKM model becomes

$$\Delta y_g = \Delta f'_g \psi + \varepsilon_g, \tag{11}$$

where $\Delta f_g = f_{g,2} - f_{g,1}$ and $f_{gt} = \{\mathbf{1}_{j(g,t)=1}, .., \mathbf{1}_{j(g,t)=J}\}$ is the vector containing the firm dummies.

In this model,

$$\hat{\psi} = \psi + \underbrace{\sum_{g=1}^{N} (\Delta f_g \Delta f_g')^{-1} \Delta f_g \varepsilon_g}_{\text{Correlated Noise}}. \tag{12}$$

Note how the dependence in the vector of estimated firm fixed effects, $\hat{\psi}$, is induced by the regressor design $\sum_{g=1}^{N} (\Delta f_g \Delta f_g')^{-1}$. As shown in Table 3 of KSS, ignoring this correlation can easily lead to underestimating standard errors by an order of magnitude in practice.

The package provides the HU standard errors on $\hat{\gamma}$ using the function `lincom_KSS`, which is designed to emulate the Stata function lincom and therefore works as a post-estimation command. We demonstrate the functioning of `lincom_KSS` with an example.

In this example, we are interested in testing whether the difference in person-year weighted mean firm effects between region 1 and region 2 is statistically different from zero. This amounts to running a regression where the dependent variable is the vector of estimated firm effects and the set of observables, Z_{gt} , is here represented by a constant and a dummy for whether the firm of worker g in year t belongs to region 2.

The resulting coefficient (and standard error) can be computed by calling the function `leave_out_KSS` specifying that we want to run the `lincom` option and using the region dummy as Z_{gt} (the constant is automatically added by the code).

[9]:
```
%Regressing firm effects on observables
namesrc='data/lincom.csv'; %testing data for the lincom function
data=importdata(namesrc);
id=data(:,1);
firmid=data(:,2);
y=data(:,5);
region=data(:,4); %Region indicator. Value -1 for region 1, Value 1 for region 2;
region_dummy=region;
region_dummy(region_dummy==-1)=0; %Make it a proper dummy variable

%Run the KSS correction and "lincom"
labels_lincom={'Region 2 Dummy'}; %give me the label of the columns of Z.
lincom_do=1; %tell the function leave_out_KSS that we want to project the firm␣
 ↪effects on some Z.
Z=region_dummy; %we're going to project the firm effects on a constant +  the␣
 ↪region dummy. Constant automatically added by the code

%Ready to call KSS with lincom option!
[sigma2_psi,sigma_psi_alpha,sigma2_alpha] =␣
 ↪leave_out_KSS(y,id,firmid,[],[],[],[],lincom_do,Z,labels_lincom);
```

```
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Running KSS Correction with the following options
Leave Out Strategy: Leave match out
```

```
Algorithm for Computation of Statistical Leverages: JLA with 200 simulations.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 1
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Info on the leave one out connected set:
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
mean wage: 4.7047
variance of wage: 0.14653
# of Movers: 9972
# of Firms: 2974
# of Person Year Observations: 89666
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 2
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Calculating the statistical leverages of the AKM model...
Running JLA Algorithm...
Done!
Elapsed time is 9.324213 seconds.
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
SECTION 3
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
PLUG-IN ESTIMATES (BIASED)
Variance of Firm Effects: 0.060695
Covariance of Firm, Person Effects: -0.012603
Variance of Person Effects: 0.10318
Correlation of Firm, Person Effects: -0.15926
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
BIAS CORRECTED ESTIMATES
Variance of Firm Effects: 0.044613
Covariance of Firm and Person Effects: 0.0025688
Variance of Person Effects: 0.079191
Correlation of Firm, Person Effects: 0.043218
Regressing the firm effects on observables...
pcg converged at iteration 115 to a solution with relative residual 8.6e-11.
****************************************
****************************************
RESULTS ON LINCOM
****************************************
****************************************
Coefficient on Region 2 Dummy:  0.25982
Robust "White" Standard Error: 0.050155
```

```
KSS Standard error:  0.088374
T-stat: 2.94
****************************************
```

We can see from the above output (make sure to scroll until the end) that the difference in person-year weighted mean firm effects between the two regions is equal to 0.26. The traditional HC or "robust" standard errors on this coefficient is around 0.05 while the HU standard error derived in KSS is roughly twice as large (0.09).