Leave Out Estimation of Variance Components in Two-Way Models using MATLAB

January 10, 2021

This notebook describes the MATLAB package that implements the leave-out correction of Kline, Saggio and Sølvsten (2020) – KSS henceforth – for two-way fixed effects models.

1 Introduction

Economists often study settings where units possess two or more group memberships, some of which can change over time. A prominent example comes from Abowd, Kramarz, and Margolis (1999) (henceforth AKM) who proposed a panel model of log wage determination that is additive in worker and firm fixed effects. This so-called "two-way" fixed effects or "AKM" model takes the form:

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + w_{gt}'\delta + \varepsilon_{gt} \qquad (g = 1, \dots, N,\ t = 1, \dots, T_g \geq 2) \tag{1}$$

where the function $j(\cdot,\cdot) : \{1, \dots, N\} \times \{1, \dots, \max_g T_g\} \to \{0, \dots, J\}$ allocates each of $n = \sum_{g=1}^{N} T_g$ person-year observations to one of $J + 1$ firms. Here α_g is a person effect, $\psi_{j(g,t)}$ is a firm effect, w_{gt} is a time-varying covariate, and ε_{gt} is a time-varying error.

Note that can we simply rewrite the original AKM model as:

$$y_i = x_i'\beta + \varepsilon_i \qquad i = 1, ..., n \tag{2}$$

where i indexes a particular person g year t observation, x_i is a vector that collects all the worker, firm dummies as well as the time-varying covariates w_{gt} so that $\beta = (\alpha, \psi, \delta)'$ is a $k \times 1$ vector that collects all the worker, firm fixed effects along with δ.

Interest in AKM models often centers on understanding how much of the variability in log wages is attributable to firms. It is common to summarize the firm contribution to wage inequality using the following two variance components parameters:

$$\sigma_\psi^2 = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}\left(\psi_{j(g,t)} - \bar{\psi}\right)^2 \qquad \text{and } \sigma_{\alpha,\psi} = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}\left(\psi_{j(g,t)} - \bar{\psi}\right)\alpha_g \tag{3}$$

where $\bar{\psi} = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}\psi_{j(g,t)}$. The variance component σ_ψ^2 measures the contribution of firm wage variability to inequality, while the covariance component $\sigma_{\alpha,\psi}$ measures the additional contribution of systematic sorting of high wage workers to high wage firms.

The function `leave_out_KSS` provides unbiased estimates of σ^2_ψ and $\sigma_{\alpha,\psi}$ as well as an estimate of $\sigma^2_\alpha = \frac{1}{n}\sum_{g=1}^{N}\sum_{t=1}^{T_g}(\alpha_g - \bar{\alpha})^2$ using the leave-out bias correction approach proposed by KSS. In what follows, we use the words "workers" and "firms" when describing the procedure for simplicity but the the function `leave_out_KSS` can be applied to any two-way models (e.g. patients and doctors, students and teachers, strata and treatment arms).

2 Running the KSS Correction

We now demonstrate the functioning of `leave_out_KSS` with a simple example.

2.1 Setup

We begin with some auxiliary lines of code that define the relevant paths, call the CMG package developed by Yiannis Koutis and set-up the parallel environment within MATLAB.

[1]:
```
%Setup Paths and Install CMG
clc
clear
cd '/Users/raffaelesaggio/Dropbox/LeaveOutTwoWay'
path(path,'codes'); %this contains the main LeaveOut Routines.
path(path,'CMG'); % CMG package http://www.cs.cmu.edu/~jkoutis/cmg.html
[result,output] = evalc('installCMG(1)'); %installs CMG routine (silently)
delete(gcp("nocreate")) %clear parallel envir.
c = parcluster('local');  %tell me # of available cores
nw = c.NumWorkers; %tell me # of available cores
pool=parpool(nw,'IdleTimeout', Inf); %all cores will be assigned to Matlab
```

```
Starting parallel pool (parpool) using the 'local' profile …
Connected to the parallel pool (number of workers: 6).
```

2.2 Importing the Data

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 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 1062 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 (Pii,Bii)…
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 23.689412 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: 0.010308
Covariance of Firm and Person Effects: 0.0046808
```

```
Variance of Person Effects: 0.085103
Correlation of Firm, Person Effects: 0.15804
```

3 Interpreting the Output

The code `leave_out_KSS` and associated output is composed by three sections.

Section 1: Here we provide info on leave-out connected set. This is the largest connected set of firms that remains connected after removing any worker from the associated graph, see Lemma 1 and the Computational Appendix of Kline, Saggio and Sølvsten (2020) for details. The code provides some summary statistics (e.g. # of movers, # of firms, mean and variance of the outcome, etc) for the leave-out connected set.

Section 2: After printing the summary statistics, the code computes the statistical leverages of the AKM model, denoted as P_{ii} , and the error influence weights, B_{ii}. Computation of $\{P_{ii}, B_{ii}\}_{i=1}^{n}$ 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). The fourth column reports the statistical leverages of the AKM model. 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, denoted as the Johnson– Lindenstrauss Approximation (JLA henceforth) algorithm in KSS for its connection to the work of Johnson and Lindenstrauss (1984), see also Achlioptas (2003).

The JLA algorithm is used to approximate the statistical leverages, P_{ii}, of the OLS fit used to compute the leave-out residual. To compute the bias correction, we also approximate the terms B_{ii} which measure the influence of the squared error terms for a given variance component. These terms can be written as

$$P_{ii} = x_i' S_{xx}^{-1} x_i, \qquad B_{ii} = x_i' S_{xx}^{-1} A S_{xx}^{-1} x_i. \tag{4}$$

where $S_{xx} = \sum_{i=1}^{n} x_i x_i'$ and A is based upon a particular variance component of interest (see example 3 of KSS).

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

Note: The user might want to pre-specify 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 (Pii,Bii)…
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 231.584528 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: 0.030397
Covariance of Firm and Person Effects: 0.01458
Variance of Person Effects: 0.048874
Correlation of Firm, Person Effects: 0.37827
```

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

One 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 (Pii,Bii)…
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 434.590252 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: 0.030395
Covariance of Firm and Person Effects: 0.014598
Variance of Person Effects: 0.048773
Correlation of Firm, Person Effects: 0.37913
```

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`=XXX.

We conclude this section by noting that the user can also calculate an exact version of $\{P_{ii}, B_{ii}\}_{i=1}^{n}$. 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 (Pii,Bii)…
Running Exact Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 167.132283 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 (Pii,Bii)…
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 2.432692 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: 0.01007
Covariance of Firm and Person Effects: 0.0048298
Variance of Person Effects: 0.084915
Correlation of Firm, Person Effects: 0.16517
```

The variance components estimated via JLA are extremely close to the `exact` ones but only take a fraction of the time to compute. If the input data has more than 10,000 obs, the code defaults to using the JLA algorithm unless the user specifies type_of_algorithm to "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 1062 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 (Pii,Bii)…
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 20.300395 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: 0.009752
Covariance of Firm and Person Effects: 0.0046773
Variance of Person Effects: 0.085602
Correlation of Firm, Person Effects: 0.16188
```

When controls are specified, the code proceeds by partialling them out. That is, it first estimates the AKM model in the leave-out connected set

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + w_{gt}'\delta + \varepsilon_{gt} \tag{5}$$

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 (unique combination of worker and firm identifier), see Remark 3 of KSS. We discuss the interpretation of the leave-out corrected variance of person effects when leaving a match out here.

The user can specify the function to run the KSS correction when leaving only an observation out using the option `leave_out_level`. When leaving a person-year observation out, the resulting KSS variance components are robust to unrestricted heteroskedasticity but not serial correlation within 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 1062 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 (Pii,Bii)…
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 20.986212 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: 0.010327
Covariance of Firm and Person Effects: 0.0045766
Variance of Person Effects: 0.085267
Correlation of Firm, Person Effects: 0.15423
```

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 is 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 for the variance of firm effects and 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. This implies that we cannot compute an unbiased estimate of $\sigma^2_{gt} = Var(\varepsilon_{gt})$ for stayers. An estimate of σ^2_{gt} for both stayers and movers is required in order to provide a bias correction for the variance of person effects, see equation (1) and Remark 3 in KSS.

The current implementation of the code provides an estimate of σ^2_{gt} for stayers that is robust to unrestricted heteroskedasticity while using an estimate of σ^2_{gt} for movers that is robust to both unrestricted heteroskedasticity and serial correlation within a match. This allows us to compute an estimate of the variance of person effects that represents 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:

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 leave person-year out estimator. 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 from the available

pairs randomly for each stayer with equal probability.

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 observables characteristics. Using the AKM model again as our leading example, suppose we are interested in the linear projection of the firm effects, ψ_{gt}, on some observables, Z_{gt}:

$$\psi_{j(g,t)} = Z'_{gt}\gamma + e_{gt} \tag{6}$$

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

$$\hat{\gamma} = \left(\sum_{g,t} Z_{gt} Z'_{gt\cdot} \right)^{-1} Z_{gt} \hat{\psi}_{gt} \tag{7}$$

KSS show that inference on $\hat{\gamma}$ needs to be adjusted. The reason is that the firm fixed effects $\hat{\psi}_{j(g,t)}$ are all potentially 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

$$\Delta y_g = \Delta f'_g \psi + \varepsilon_g \tag{8}$$

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,

$$\hat{\psi} = \psi + \underbrace{\sum_{g=1}^{N} (\Delta f_g \Delta f'_g)^{-1} \Delta f_g \varepsilon_g}_{\text{Correlated Noise}} \tag{9}$$

Notice how the dependence in the vector of estimated firm fixed effects, $\hat{\psi}$, is induced by the design $\sum_{g=1}^{N} (\Delta f_g \Delta f'_g)^{-1}$. Intuitively, conducting inference on $\hat{\gamma}$ is challenging as we have to provide standard errors in a context where all of our underlying observations are potentially correlated with one another. As shown in Table 3 of KSS, ignoring this correlation can easily lead standard errors to be underestimated by an order of magnitude in practice.

The package provides the correct standard errors on $\hat{\gamma}$ using the function `lincom_KSS` which is designed like 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} , here is 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 1123 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 (Pii,Bii)…
```

```
Running JLA Algorithm…
Time spent computing (Pii,Bii)
Elapsed time is 44.608516 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: 0.044213
Covariance of Firm and Person Effects: 0.0029136
Variance of Person Effects: 0.078861
Correlation of Firm, Person Effects: 0.049343
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
White Standard Error: 0.050155
KSS Standard error:  0.09534
T-stat: 2.7252
****************************************
```

We can see from the output above (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 "White" robust-standard error on this coefficient is around 0.05 while the KSS-adjusted standard error is roughly twice as large.