# acreg: instructions and examples

Fabrizio Colella[*], Rafael Lalive[*], Seyhun Orcan Sakalli[†], and Mathias Thoenig[*]

## Contents



[*]Department of Economics, HEC University of Lausanne, 1015 Lausanne, Switzerland

[†]King's Business School, King's College London, WC2R-2LS London, United Kingdom

[‡]Our companion statistical package (acreg) can be downloaded at the following address https://acregstata.weebly.com. If you use the command please cite: Colella, Fabrizio; Lalive, Rafael; Sakalli, Seyhun Orcan; Thoenig, Mathias. (2019) Inference with Arbitrary Clustering, IZA Discussion Paper n. 12584.

# 1 Description of the command

acreg computes standard errors corrected for arbitrary cluster correlation in spatial and network settings. It implements a range of error correction methods for linear regression models: OLS and 2SLS.

## 1.1 Syntax

```
acreg depvar [varlist1] [(varlist2 = varlist_iv) [if] [in]] [fweight pweight] [, id(idvar)
time(timevar) spatial network latitude(latitudevar) longitude(longitudevar)
links_mat(varlist_links) dist_mat(varlist_distances) dist(distcutoff) lag(timecutoff)
storeweights storedistances weights(varlist_weights) cluster(varlist_cluster) pfe1(fe1var)
pfe2(fe2var) correctr2 dropsingletons hac nbclust(n_clusters) ak0 bartlett]
```

- `depvar` is the dependent variable.
- `varlist1` is the list of exogenous variables.
- `varlist2` is the list of endogenous variables.
- `varlist_iv` is the list of exogenous variables used with varlist1 as instruments for varlist2.

## 1.2 Options

### 1.2.1 Panel

- `idvar` is the unique identifier, required in panel databases.
- `timevar` is the time unit variable, required in panel databases.

### 1.2.2 Spatial Environment

- `spatial` specifies that the environment is a spatial environment, not required if no arbitrary cluster correction and if varlist_weights or varlist_cluster or network option is specified.
- `latitudevar` is the variable containing the latitude of each observation, decimal degrees: range[-180,180].
- `longitudevar` is the variable containing the longitude of each observation, decimal degrees: range[-180,180].
- `varlist_distances` is the list of N variables containing bilateral distances between observations. In the spatial environment, bilateral distance is the spatial distance between observations, i.e., physical distance between two locations (in the network environment, it is the network distance between observations, i.e., the number of links along the shortest path between two nodes).

- `distcutoff` specifies the distance cutoff in kilometers beyond which the correlation between error term of two observations is assumed to be zero, required if latitude and longitude are specified or dist_mat is specified.

- `timecutoff` specifies the time lag for observations with the same idvar, not required in cross-sectional environment, default in panel environment is 0, i.e. when id and time are specified. In panel environment when timecutoff is 0, or not specified, Standard Errors are automatically clustered at id-x-time cell level.

### 1.2.3 Network Environment

- `network` specifies that the environment is a network environment, not required if no arbitrary cluster correction and if varlist_weights or varlist_cluster or spatial option is specified.

- `varlist_links` is the list of N dummy variables specifying the links between observations, i.e., the adjacency matrix. If distcutoff>1 only the first observation in time of each individual will be used as input to compute the bil istance between two nodes.

- `varlist_distances` is the list of N variables containing bilateral distances between observations. In the network environment, bilateral distance is the network distance between observations, i.e., the number of links along the shortest path between two nodes (in the spatial environment, it is the spatial distance between observations, i.e., p distance between two locations).

- `distcutoff` specifies the distance cutoff (geodesic paths) beyond which the correlation between error term of two observations is assumed to be zero, required if dist_mat is specified, optional if links_mat is specified, default is 1. When links_mat is specified and distcutoff is grerater than 1, acreg will automatically computes the bilateral distance between two nodes.

- `timecutoff` specifies the time lag for observations with the same idvar, not required in cross-sectional environment, default in panel environment is 0, i.e. when id and time are specified. In panel environment when timecutoff is 0, or not specified, Standard Errors are automatically clustered at id-x-time cell level.

### 1.2.4 Multiway Clustering Environment

- `varlist_cluster` is the list of variables to use for multy-way clustered SEs, not required if no arbitrary cluster correction and if the option spatial or the option network or varlist_weights is specified.

#### 1.2.5 Arbitrary Clustering Environment

- `varlist_weights` is the list of NxT variables containing the weights that will be used for error correction, not required if no arbitrary cluster correction and if the option spatial or the option network or varlist_cluster is specified.

#### 1.2.6 High Dimensional Fixed Effects (partial out)

- `fe1var` identification of the first high dimensional fixed effects variable.
- `fe2var` identification of the second high dimensional fixed effects variable.
- `correctr2` computes the pre-partialling out R2 when pfe1 or pfe2 are specified. not allowed with fweights.
- `dropsingletons` drop singleton groups when pfe1 or pfe2 are specified.

#### 1.2.7 Storing

- `storeweights` stores the computed weights used to correct the VCV for arbitrary cluster correlation as a matrix under the name weightsmat, which may be used as input for the option varlist_weights, only if the option spatial or the option network or varlist_cluster is specified.
- `storedistances` stores the computed distances used to correct the VCV for arbitrary cluster correlation as a matrix under the name distancesmat, which may be used as input for the option varlist_distances, only if the option spatial or the option network is specified and varlist_distances is not specified.

#### 1.2.8 Other Options

- `hac` heteroskedastic and autocorrelation consistent (HAC) standard errors; lagcutoff will be the temporal decay, requires id, time and lagcutoff.
- `n_clusters` is the number of clusters used to compute the Kleibergen-Paap statistic in case of cluster correction. Default is 100.
- `ak0` Conservative Standard Errors. Residuals are replaced by the difference between the dependent variable and its mean while computing the VCV matrix.
- `bartlett` Impose a distance linear decay in the correlation structure.

## 1.3 Stored results

`acreg` stores the following in e():

**Scalars**

- `e(N)` number of observations
- `e(mss)` model sum of squares (centered)
- `e(mssu)` model sum of squares (uncentered)
- `e(rss)` residual sum of squares
- `e(tss)` total sum of squares (centered)
- `e(tssu)` total sum of squares (uncentered)
- `e(r2)` centered R2 (1-rss/tss)
- `e(r2u)` uncentered R2
- `e(widstat)` Kleibergen-Paap rk Wald F statistic

**Matrices**

- `e(b)` coefficient vector
- `e(V)` corrected variance-covariance matrix of the estimators

**Functions**

- `e(sample)` marks estimation sample

# 2 How to use `acreg` to correct SEs accounting for spatial correlation using latitude and longitude with cross-sectional data?

## 2.1 Database

For this example we use the data on the homicides in southern states of the U.S. `homicide_1960_1990.dta` available at the STATA website. Data contains, among others, the county-level homicide rate per year per 100,000 persons (`hrate`), the population in logs (`ln_population`), the logarithm of the average income (`ln_income`), the unemployment rate (`unemployment`) and the average age (`age`). The data is an extract of the data originally used by Messner *et al.* (1999) and concerns 4 different periods (1960, 1970, 1980, 1990).

## 2.2 Estimation

In this example we consider only the cross-sectional database for 1990 and we want to estimate the effect of *income* on *homicide rate*, controlling for *population* and *age*. We claim that income is endogenous and we assume that unemployment is a valid instrument for it[1]. Figure 1 shows the spatially dependency of the outcome variable, the endogenous regressor and the instrument.

We first estimate the model assuming that observations' errors are uncorrelated [2].

```
. use http://www.stata-press.com/data/r15/homicide1990.dta , clear
(S.Messner et al.(2000), U.S southern county homicide rates in 1990)

. acreg hrate ln_population age (ln_income=unemployment)
NO ARBITRARY CLUSTERING CORRECTION
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 990.487
                                                Number of obs =      1412
Total (centered) SS     =  69908.59003          Centered R2   =    0.1079
Total (uncentered) SS   =  198667.4579          Uncentered R2 =    0.6861
Residual SS             =  62363.84851
```

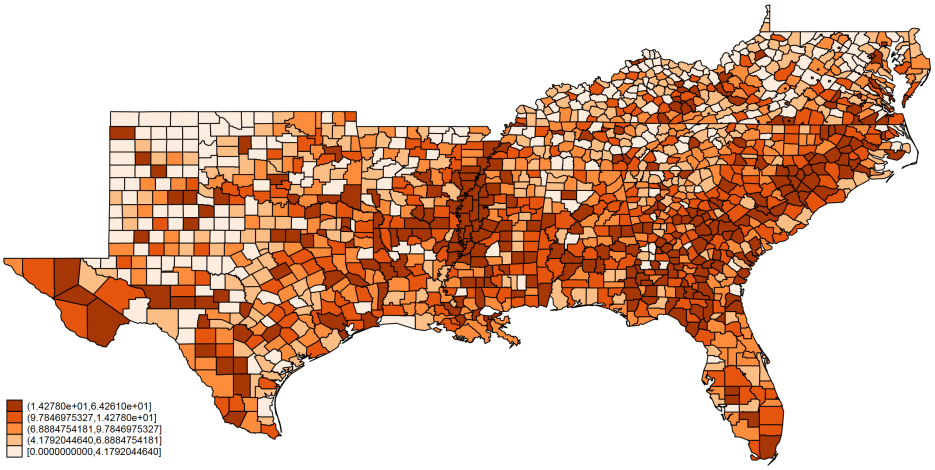
| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -8.822082 | 1.35491 | -6.51 | 0.000 | -11.47766 | -6.166507 |
| ln_population | 1.404433 | .2769494 | 5.07 | 0.000 | .861622 | 1.947244 |
| age | -.281615 | .050726 | -5.55 | 0.000 | -.381036 | -.1821939 |
| _cons | 94.4605 | 12.42859 | 7.60 | 0.000 | 70.10091 | 118.8201 |

[1] We do not test this assumption since this is outside the scope of this document.

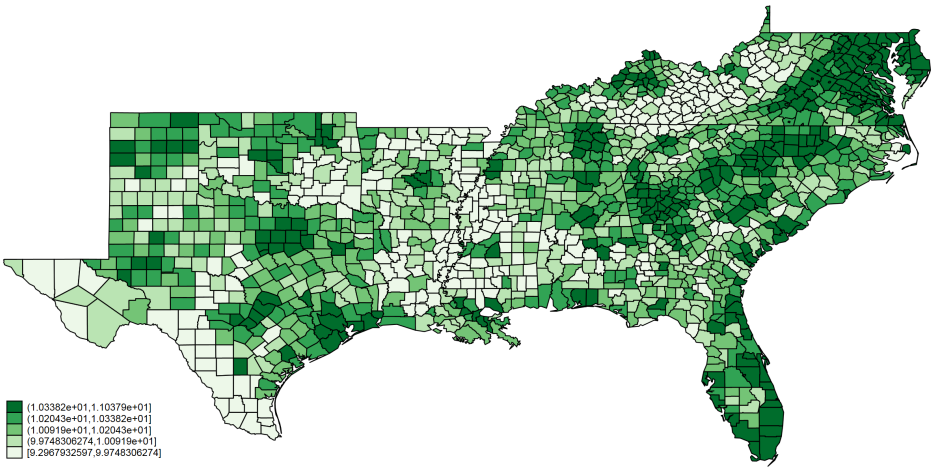
[2] This is equivalent of using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), robust`

Figure 1: Homicide rate, log income and unemployment in 1990 for southern U.S. counties

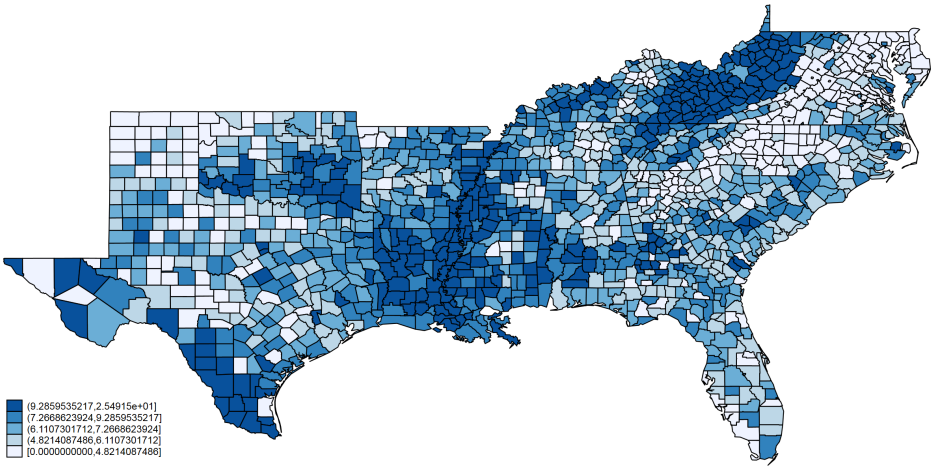
(a) Homicide rate




(b) Income




(c) Unemployment



We now estimate the model above clustering standard errors by state [3, 4].

```
. acreg hrate ln_population age (ln_income=unemployment), cluster(sfips)
MULTIWAY CLUSTERING CORRECTION
Cluster variable(s): sfips
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 143.959
                                                    Number of obs =     1412
Total (centered) SS     =  69908.59003              Centered R2   =   0.1079
Total (uncentered) SS   =  198667.4579              Uncentered R2 =   0.6861
Residual SS             =  62363.84851
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -8.822082 | 1.801762 | -4.90 | 0.000 | -12.35347 | -5.290693 |
| ln_population | 1.404433 | .3090553 | 4.54 | 0.000 | .7986955 | 2.01017 |
| age | -.281615 | .1303804 | -2.16 | 0.031 | -.5371558 | -.0260741 |
| _cons | 94.4605 | 17.89048 | 5.28 | 0.000 | 59.3958 | 129.5252 |

We now estimate the model above using the spatial correction proposed by Conley (1999), with a threshold of 100 kilometers. This means that the error of each county is assumed to be correlated with the ones of all the counties that are located within a radius of 100 kilometers from it[5].

```
. acreg hrate ln_population age (ln_income=unemployment), ///
> spatial latitude(_CX) longitude(_CY) dist(100)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 112.917
                                                    Number of obs =     1412
Total (centered) SS     =  69908.59003              Centered R2   =   0.1079
Total (uncentered) SS   =  198667.4579              Uncentered R2 =   0.6861
Residual SS             =  62363.84851
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -8.822082 | 2.357644 | -3.74 | 0.000 | -13.44298 | -4.201183 |
| ln_population | 1.404433 | .4689154 | 3.00 | 0.003 | .4853754 | 2.32349 |
| age | -.281615 | .109112 | -2.58 | 0.010 | -.4954706 | -.0677594 |
| _cons | 94.4605 | 21.86325 | 4.32 | 0.000 | 51.60932 | 137.3117 |

[3] We are aware that the number of states (clusters) is small and inference would suffer from it.

[4] This is equivalent of using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), cluster(sfips)`.

[5] In this example, there are in average 89 counties correlated with each county.

## 2.3 Additional Options

### 2.3.1 Thresholds

If we want to account for correlation between counties at a greater distance, we can increase the distance cutoff using the `dist()` option. In the following example we allow for a radius of 200 kilometers.

```
. acreg hrate ln_population age (ln_income=unemployment), ///
> spatial latitude(_CX) longitude(_CY) dist(200)
SPATIAL CORRECTION
DistCutoff: 200
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  92.540
                                                    Number of obs =     1412
Total (centered) SS     =  69908.59003              Centered R2   =   0.1079
Total (uncentered) SS   =  198667.4579              Uncentered R2 =   0.6861
Residual SS             =  62363.84851
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -8.822082 | 2.733507 | -3.23 | 0.001 | -14.17966 | -3.464507 |
| ln_population | 1.404433 | .4834539 | 2.90 | 0.004 | .4568805 | 2.351985 |
| age | -.281615 | .1223503 | -2.30 | 0.021 | -.5214172 | -.0418127 |
| _cons | 94.4605 | 24.78394 | 3.81 | 0.000 | 45.88487 | 143.0361 |

### 2.3.2 Bartlett

In previous examples the matrix used for the computation of the variance covariance matrix is binary: for each counties-pair it contains 1 if they are the two counties located within the distance threshold and 0 otherwise. *acreg* allows for weights in the matrix linearly decreasing as the distance increases with values very close to one for near counties and almost zero for counties close to the distance cutoff. To do that we only need to add the option `bartlett` to the syntax.

```
. acreg hrate ln_population age (ln_income=unemployment), ///
> spatial latitude(_CX) longitude(_CY) dist(200) bartlett
SPATIAL CORRECTION
DistCutoff: 200
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 132.260
                                                    Number of obs =     1412
Total (centered) SS     =  69908.59003              Centered R2   =   0.1079
```

```
Total (uncentered) SS    =  198667.4579                Uncentered R2 =   0.6861
Residual SS              =  62363.84851
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -8.822082 | 2.313018 | -3.81 | 0.000 | -13.35551 | -4.28865 |
| ln_population | 1.404433 | .4388646 | 3.20 | 0.001 | .5442741 | 2.264592 |
| age | -.281615 | .1015135 | -2.77 | 0.006 | -.4805778 | -.0826522 |
| _cons | 94.4605 | 21.23503 | 4.45 | 0.000 | 52.84061 | 136.0804 |

### 2.3.3 Partial out high dimensional fixed effects

`acreg` allows for adding high dimensional fixed effects and partial them out, using the `hdfe` command by Correia (2016): up to two fixed effects variables can be specified through the options `pfe1()` and `pfe2()`. In the example below we estimate the previous model adding state fixed effects.

```
. acreg hrate ln_population age (ln_income=unemployment), ///
> spatial latitude(_CX) longitude(_CY) dist(100) pfe1(sfips)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  0
No HAC Correction
Absorbed FE: sfips
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 307.097
                                                Number of obs =      1412
Total (centered) SS      =  58943.23761         Centered R2   =    0.1002
Total (uncentered) SS    =  58943.23761         Uncentered R2 =    0.1002
Residual SS              =  53037.54212
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -13.88229 | 1.835268 | -7.56 | 0.000 | -17.47935 | -10.28523 |
| ln_population | 1.649735 | .4000578 | 4.12 | 0.000 | .8656367 | 2.433834 |
| age | -.178832 | .0960779 | -1.86 | 0.063 | -.3671412 | .0094771 |
| _cons | 2.05e-15 | .241196 | 0.00 | 1.000 | -.4727355 | .4727355 |

```
nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2
```

# 3 How to use acreg to correct SEs accounting for spatial correlation using latitude and longitude with panel data?

## 3.1 Database

For this example we use the data on the homicides in southern states of the U.S. homicide_1960_1990.dta available at the STATA website. Data contains, among others, the county-level homicide rate per year per 100,000 persons (hrate), the population in logs (ln_population), the logarithm of the average income (ln_income), the unemployment rate (unemployment) and the average age (age). The data is an extract of the data originally used by Messner *et al.* (1999) and concerns 4 different periods (1960, 1970, 1980, 1990).

## 3.2 Estimation

In this example we want to estimate the effect of *income* on *homicide rate*, controlling for *population* and *age*. We assume that unemployment is a valid instrument for it[6].

### 3.2.1 Pooled model

In this section we consider a pooled model in which we do not include any Random or Fixed Effect. We first estimate the model assuming that observations' errors are uncorrelated[7].

```
. use http://www.stata-press.com/data/r15/homicide_1960_1990.dta , clear
(S.Messner et al.(2000), U.S southern county homicide rate in 1960-1990)

. acreg hrate ln_population age (ln_income=unemployment)
NO ARBITRARY CLUSTERING CORRECTION
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 289.132
                                                Number of obs =      5648
Total (centered) SS     =  286387.1082          Centered R2   =  -0.0447
Total (uncentered) SS   =  781008.6785          Uncentered R2 =   0.6169
Residual SS             =  299188.6495
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | 3.83872 | .7815313 | 4.91 | 0.000 | 2.306947 | 5.370494 |
| ln_population | -.4411802 | .1968992 | -2.24 | 0.025 | -.8270955 | -.055265 |
| age | -.4626917 | .0637006 | -7.26 | 0.000 | -.5875425 | -.3378408 |
| _cons | -7.265041 | 4.126029 | -1.76 | 0.078 | -15.35191 | .8218268 |

[6] We do not test this assumption since this is outside the scope of this document.

[7] This is equivalent of using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), robust`

We now estimate the same model, but we use the panel feature of `acreg` to account for correlation between observations from the same county[8]. We assume no correlation across counties. We fill the option `id()` with the county id, the option `time()` with the year variable and the option `lag()` with a number greater or equal than the maximum lag between observations, which in this case is 30[9].

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(30)
NO ARBITRARY CLUSTERING CORRECTION
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 210.438
                                                    Number of obs =     5648
Total (centered) SS     =  286387.1082              Centered R2   =  -0.0447
Total (uncentered) SS   =  781008.6785              Uncentered R2 =   0.6169
Residual SS             =  299188.6495
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | 3.83872 | .921289 | 4.17 | 0.000 | 2.033027 | 5.644414 |
| ln_population | -.4411802 | .2513095 | -1.76 | 0.079 | -.9337379 | .0513774 |
| age | -.4626917 | .0787756 | -5.87 | 0.000 | -.617089 | -.3082943 |
| _cons | -7.265041 | 4.832603 | -1.50 | 0.133 | -16.73677 | 2.206687 |

We now estimate the model above adding to the temporal correlation the spatial correction proposed by Conley (1999), with a threshold of 100 kilometers. This means that the error of each county at a given year is assumed to be correlated with the ones of all the counties observed at the same year that are located within a radius of 100 kilometers from it. Note that correlation between near counties but observed at different point in time is still assumed to be null.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(30) ///
> spatial latitude(_CX) longitude(_CY) dist(100)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  30
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  24.838
                                                    Number of obs =     5648
Total (centered) SS     =  286387.1082              Centered R2   =  -0.0447
Total (uncentered) SS   =  781008.6785              Uncentered R2 =   0.6169
Residual SS             =  299188.6495
```

[8] Note that the estimation of the betas does not change with respect to the previous model, `acreg` is only used for the computation of the standard errors.

[9] This is equivalent of using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), cluster(_ID)`

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | 3.83872 | 1.810937 | 2.12 | 0.034 | .2893488 | 7.388092 |
| ln_population | -.4411802 | .3871668 | -1.14 | 0.254 | -1.200013 | .3176528 |
| age | -.4626917 | .1425257 | -3.25 | 0.001 | -.742037 | -.1833464 |
| _cons | -7.265041 | 9.814094 | -0.74 | 0.459 | -26.50031 | 11.97023 |

### 3.2.2 FE model

In the following example we replicate the previous model, accounting for both spatial and temporal correlation, but we add to the specification the *counties fixed effects* using the option `pfe1`.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(30) ///
> spatial latitude(_CX) longitude(_CY) dist(100) pfe1(_ID)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  30
No HAC Correction
Absorbed FE: _ID
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  49.605
                                                    Number of obs =     5648
Total (centered) SS     =  144755.2058              Centered R2   =   0.0175
Total (uncentered) SS   =  144755.2058              Uncentered R2 =   0.0175
Residual SS             =  142223.0274
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | .2588154 | 1.149746 | 0.23 | 0.822 | -1.994645 | 2.512276 |
| ln_population | -1.630949 | 1.740873 | -0.94 | 0.349 | -5.042997 | 1.781099 |
| age | .1466193 | .2006033 | 0.73 | 0.465 | -.2465559 | .5397944 |
| _cons | -1.31e-17 | .1743959 | -0.00 | 1.000 | -.3418097 | .3418097 |

```
nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2
```

We now add to the previous model also time fixed effects using the option `pfe2`.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(30) ///
> spatial latitude(_CX) longitude(_CY) dist(100) pfe1(_ID)  pfe2(year)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  30
No HAC Correction
Absorbed FE: _ID and year
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:   3.895
                                                    Number of obs =     5648
```

```
Total (centered) SS     =   136166.339                Centered R2   =  -0.0793
Total (uncentered) SS   =   136166.339                Uncentered R2 =  -0.0793
Residual SS             =  146961.8234
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | -13.30126 | 17.5969 | -0.76 | 0.450 | -47.79055 | 21.18803 |
| ln_population | -1.602695 | 2.253785 | -0.71 | 0.477 | -6.020033 | 2.814642 |
| age | .0038921 | .0937463 | 0.04 | 0.967 | -.1798472 | .1876314 |
| _cons | -1.11e-15 | .128699 | -0.00 | 1.000 | -.2522454 | .2522454 |

```
nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2
```

## 3.3 Additional Options

### 3.3.1 Thresholds

Now we still account for spatial correlation between observations of the same year, but we do not account for any kind of temporal correlation. We do that by setting the lagcutoff at 0[10].

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(0) ///
> spatial latitude(_CX) longitude(_CY) dist(100)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  25.662
                                                Number of obs =     5648
Total (centered) SS     =  286387.1082          Centered R2   =  -0.0447
Total (uncentered) SS   =  781008.6785          Uncentered R2 =   0.6169
Residual SS             =  299188.6495
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | 3.83872 | 1.743993 | 2.20 | 0.028 | .4205571 | 7.256884 |
| ln_population | -.4411802 | .3542752 | -1.25 | 0.213 | -1.135547 | .2531864 |
| age | -.4626917 | .1347804 | -3.43 | 0.001 | -.7268564 | -.198527 |
| _cons | -7.265041 | 9.486122 | -0.77 | 0.444 | -25.8575 | 11.32742 |

Now we account for spatial correlation between observations of the same year, and also for temporal correlation between observations from the same county, but only between neighbor decades, i.e. two observations from the same county are assumed to be correlated only if they are observed with less than

[10] Note that the result will be different than the one in the cross sectional environment (`acreg hrate ln_population age (ln_income=unemployment) , spatial latitude(_CX) longitude(_CY) dist(100)`) because the spatial correlation is assumed only between observations from the same year.

10 years difference. We do that by setting the *lagcutoff* equal to 10.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(10) ///
> spatial latitude(_CX) longitude(_CY) dist(100)
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  10
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  24.722
                                                    Number of obs =     5648
Total (centered) SS     =  286387.1082              Centered R2   =  -0.0447
Total (uncentered) SS   =  781008.6785              Uncentered R2 =   0.6169
Residual SS             =  299188.6495
```

| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | 3.83872 | 1.801373 | 2.13 | 0.033 | .3080935 | 7.369347 |
| ln_population | -.4411802 | .377059 | -1.17 | 0.242 | -1.180202 | .2978418 |
| age | -.4626917 | .1403627 | -3.30 | 0.001 | -.7377975 | -.1875859 |
| _cons | -7.265041 | 9.822551 | -0.74 | 0.460 | -26.51689 | 11.9868 |

### 3.3.2 HAC

In the previous examples the matrix used for the computation of the variance covariance matrix is binary. We can use the option `hac` to have a linear decay in time and compute Heteroscedasticity-Autocorrelation-Consistent standard errors, following Newey and West (1987).

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year) lagcut(30) ///
> spatial latitude(_CX) longitude(_CY) dist(100) hac
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  30
HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  25.031
                                                    Number of obs =     5648
Total (centered) SS     =  286387.1082              Centered R2   =  -0.0447
Total (uncentered) SS   =  781008.6785              Uncentered R2 =   0.6169
Residual SS             =  299188.6495
```

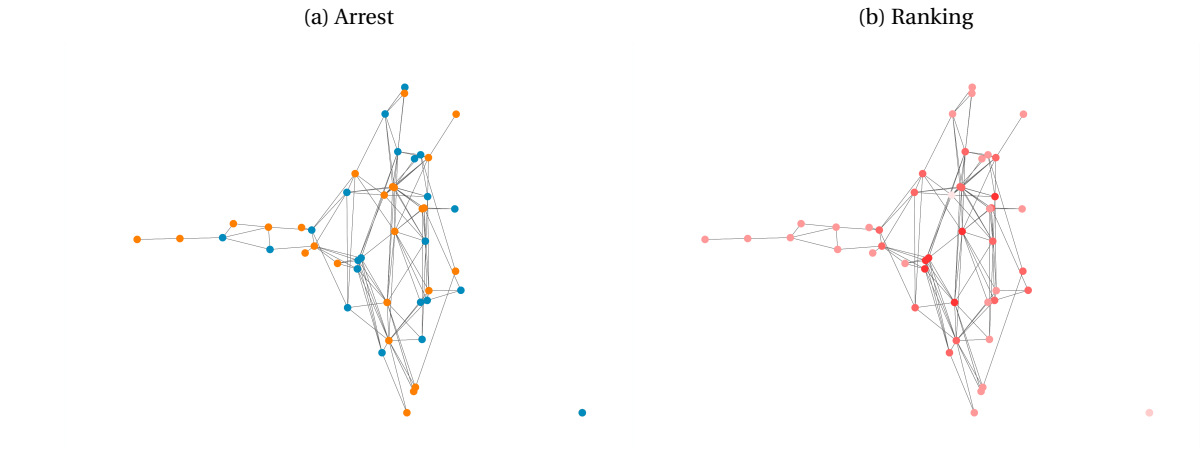
| hrate | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ln_income | 3.83872 | 1.785354 | 2.15 | 0.032 | .3394916 | 7.337949 |
| ln_population | -.4411802 | .3727145 | -1.18 | 0.237 | -1.171687 | .2893268 |
| age | -.4626917 | .139132 | -3.33 | 0.001 | -.7353854 | -.189998 |
| _cons | -7.265041 | 9.713499 | -0.75 | 0.455 | -26.30315 | 11.77307 |

# 4 How to use `acreg` to correct SEs accounting for correlation in network using an adjacency matrix with cross-sectional data?

## 4.1 Database

For this example we use a dataset of co-offending in a London-based youth gang. Data were collected by James Densley and Thomas Grund. The data has been used in Grund and Densley (2012) and Grund and Densley (2015). Information on 54 individuals are reported, two individuals are recorded to be linked if they committed at least a crime together. Data contains, among others, the age (`Age`), the birthplace (`Birthplace`), the number of arrests (`Arrests`), the number of convictions (`Convictions`), and the position in the gang's internal hierarchy( `Ranking`). The symmetric binary links constituting the co-offending network are stored in 54 variables (`_net2_1-_net2_54`). Figure 2 shows the distribution of the variables `Arrest` and `Ranking` within the network.

Figure 2: Gang Network




**Notes:** In figure A, blue dots represent arrested people. In figure B, darker red identify a greater position in the ranking

The code below is necessary to load the dataset (`webnwuse gang`), load the network (`nwload gang`) and replace the diagonal of the adjacency matrix with ones (*the loop*).

```
. webnwuse gang
Loading successful

(2 networks)
-------------------------

      gang_valued
      gang

. nwload gang

. forvalues j = 1(1)54 {
  2.         qui     replace _net2_`j´ = 1 in `j´
  3. }
```

## 4.2 Estimation

In this example we want to estimate the effect of *ranking* on *arrests*, controlling for *age*, *residence* and *birthplace FEs*. We first estimate the model assuming that observations' errors are uncorrelated[11].

```
. acreg Arrest Ranking  Age Residence i.Birthplace
NO ARBITRARY CLUSTERING CORRECTION
No HAC Correction
No Absorbed FEs
Included instruments: Ranking Age Residence 1b.Birthplace 2.Birthplace 3.Birthplace 4.Birthplace
                                                Number of obs =       54
Total (centered) SS     =  2196.537037          Centered R2   =   0.2442
Total (uncentered) SS   =         7497          Uncentered R2 =   0.7786
Residual SS             =  1660.198039
```

| Arrests | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Ranking | -2.168476 | .8207074 | -2.64 | 0.008 | -3.777033 | -.5599192 |
| Age | .7665194 | .3094139 | 2.48 | 0.013 | .1600793 | 1.372959 |
| Residence | -1.534665 | 1.561649 | -0.98 | 0.326 | -4.59544 | 1.526111 |
| Birthplace | | | | | | |
| Caribbean | 0 | (empty) | | | | |
| East Africa | -.2523035 | 2.869505 | -0.09 | 0.930 | -5.87643 | 5.371822 |
| UK | .7012659 | 2.228246 | 0.31 | 0.753 | -3.666016 | 5.068548 |
| West Africa | .8171717 | 2.012521 | 0.41 | 0.685 | -3.127297 | 4.76164 |
| _cons | 2.317286 | 7.506876 | 0.31 | 0.758 | -12.39592 | 17.03049 |

We now estimate the model above using the standard errors correction proposed in our paper (Colella *et al.*, 2019). We assume that the error of each individual is correlated with the one of an other individual if they are linked in the network. To implement this in `acreg` we provide as input in the `links_mat` option the variables containing the adjacency matrix and we set *distcutoff* equal to 1.

```
. acreg Arrest Ranking  Age Residence i.Birthplace, network links_mat(_net2_*) dist(1)
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: Ranking Age Residence 1b.Birthplace 2.Birthplace 3.Birthplace 4.Birthplace
                                                Number of obs =       54
Total (centered) SS     =  2196.537037          Centered R2   =   0.2442
Total (uncentered) SS   =         7497          Uncentered R2 =   0.7786
Residual SS             =  1660.198039
```

| Arrests | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Ranking | -2.168476 | .7132431 | -3.04 | 0.002 | -3.566407 | -.7705455 |

[11] This is equivalent of using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 Arrest Ranking Age Residence i.Birthplace , robust`

| | | | | | | |
|---|---|---|---|---|---|---|
| Age | .7665194 | .3730319 | 2.05 | 0.040 | .0353904 | 1.497648 |
| Residence | -1.534665 | 1.618858 | -0.95 | 0.343 | -4.707568 | 1.638239 |
| Birthplace | | | | | | |
| Caribbean | 0 | (empty) | | | | |
| East Africa | -.2523035 | 2.258789 | -0.11 | 0.911 | -4.679449 | 4.174842 |
| UK | .7012659 | 2.984775 | 0.23 | 0.814 | -5.148785 | 6.551317 |
| West Africa | .8171717 | 2.260143 | 0.36 | 0.718 | -3.612627 | 5.24697 |
| _cons | 2.317286 | 7.825902 | 0.30 | 0.767 | -13.0212 | 17.65577 |

## 4.3 Additional Options

### 4.3.1 Accounting for degree grater than one

Each node of a network has a certain number of links that connects it to other nodes. This number is called the degree $k$ of a node. `acreg` allows the user to accoun for correlation between two observations that are not necessarily directly linked but are linked through other observations. Starting from the same 0-1 adjacency matrix used in the previous example, we now want to allow for correlation also between individuals that are linked through one other individual (degree 2). To do that we will use the same syntax but we change the *distcutoff* to 2.

```
. acreg Arrest Ranking  Age Residence i.Birthplace, network links_mat(_net2_*) dist(2)
NETWORK CORRECTION
DistCutoff: 2
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: Ranking Age Residence 1b.Birthplace 2.Birthplace 3.Birthplace 4.Birthplace
                                                Number of obs =       54
Total (centered) SS     =  2196.537037          Centered R2   =   0.2442
Total (uncentered) SS   =         7497          Uncentered R2 =   0.7786
Residual SS             =  1660.198039
```

| Arrests | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Ranking | -2.168476 | .4801238 | -4.52 | 0.000 | -3.109502 | -1.227451 |
| Age | .7665194 | .4001636 | 1.92 | 0.055 | -.0177869 | 1.550826 |
| Residence | -1.534665 | 2.138931 | -0.72 | 0.473 | -5.726892 | 2.657563 |
| Birthplace | | | | | | |
| Caribbean | 0 | (empty) | | | | |
| East Africa | -.2523035 | . | . | . | . | . |
| UK | .7012659 | 2.48418 | 0.28 | 0.778 | -4.167637 | 5.570169 |
| West Africa | .8171717 | 1.141966 | 0.72 | 0.474 | -1.421041 | 3.055385 |
| _cons | 2.317286 | 8.948291 | 0.26 | 0.796 | -15.22104 | 19.85561 |

### 4.3.2 Bartlett

In previous examples the matrix used for the computation of the variance covariance matrix is binary: it contains values 1 for each pair of individuals that are first or second degree linked, and zeros otherwise. *acreg* allows for weights in the matrix linearly decreasing as the as the network distance[12] increases. To do that in our sample, i.e. having ones for first degree linked observations and 0.5 for second degree ones we will use the option `bartlett`.

```
. acreg Arrest Ranking  Age Residence i.Birthplace, network links_mat(_net2_*) dist(2) ///
> bartlett
NETWORK CORRECTION
DistCutoff: 2
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: Ranking Age Residence 1b.Birthplace 2.Birthplace 3.Birthplace 4.Birthplace
                                                Number of obs =       54
Total (centered) SS     =  2196.537037          Centered R2   =   0.2442
Total (uncentered) SS   =         7497          Uncentered R2 =   0.7786
Residual SS             =  1660.198039
```

| Arrests | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Ranking | -2.168476 | .7688551 | -2.82 | 0.005 | -3.675404 | -.6615479 |
| Age | .7665194 | .3427023 | 2.24 | 0.025 | .0948352 | 1.438204 |
| Residence | -1.534665 | 1.590511 | -0.96 | 0.335 | -4.652009 | 1.582679 |
| Birthplace | | | | | | |
| Caribbean | 0 | (empty) | | | | |
| East Africa | -.2523035 | 2.582265 | -0.10 | 0.922 | -5.31345 | 4.808843 |
| UK | .7012659 | 2.633815 | 0.27 | 0.790 | -4.460917 | 5.863448 |
| West Africa | .8171717 | 2.139917 | 0.38 | 0.703 | -3.376988 | 5.011331 |
| _cons | 2.317286 | 7.668048 | 0.30 | 0.762 | -12.71181 | 17.34638 |

### 4.3.3 Partial out high dimensional fixed effects

`acreg` allows for adding high dimensional fixed effects and partial them out, using the `hdfe` command by Correia (2016): up to two fixed effects variables can be specified through the options `pfe1()` and `pfe2()`. In the example below we estimate the previous model partialing out birthplace FEs instead of adding them as dummies in the main regression.

```
. acreg Arrest Ranking  Age Residence, network links_mat(_net2_*) dist(1) pfe1(Birthplace)
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  0
No HAC Correction
Absorbed FE: Birthplace
```

[12] With distance here we refer to the strength of the link: first degree is distance 1, second degree is distance 2, etc...

Included instruments: Ranking Age Residence

| | | | |
|---|---|---|---|
| | | Number of obs | = 54 |
| Total (centered) SS | = 2131.041667 | Centered R2 | = 0.2209 |
| Total (uncentered) SS | = 2131.041667 | Uncentered R2 | = 0.2209 |
| Residual SS | = 1660.198039 | | |

| Arrests | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Ranking | -2.168476 | .7132431 | -3.04 | 0.002 | -3.566407 | -.7705455 |
| Age | .7665194 | .3730319 | 2.05 | 0.040 | .0353904 | 1.497648 |
| Residence | -1.534665 | 1.618858 | -0.95 | 0.343 | -4.707568 | 1.638239 |
| _cons | -3.59e-16 | .7728722 | -0.00 | 1.000 | -1.514802 | 1.514802 |

nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2

# 5 How to use `acreg` to correct SEs accounting for correlation in network using an adjacency matrix with panel data?

## 5.1 Database

For this section we use an ad-hoc database that can be downloaded from our website. It is a balanced panel dataset on 1000 observations (*NT*) referring to 100 (*N*) individuals at 10 (*T*) points in time. Individuals are identified through the variable `id`, while time through the variable `time`. Database also contains, among others, the following variables `Y_it`, `X1_it`, `Z_it`, `IV_it`. The symmetric binary links constituting the network are stored in 100 (*N*) variables (`clus_1-clus_100`).

## 5.2 Estimation

In this example we want to estimate the effect of `Z_it` on `Y_it`, controlling for `X_it`. We claim that *Z_it* is endogenous and we assume that *IV_it* is a valid instrument for it.

### 5.2.1 Pooled model

In this section we consider a pooled model in which we do not include any Random or Fixed Effects. We first estimate the model assuming that observations' errors are uncorrelated[13].

```
. use https://acregstata.weebly.com/uploads/2/9/1/6/29167217/acregfakedata.dta , clear

. acreg Y_it X1_it  (Z_it=IV_it)
NO ARBITRARY CLUSTERING CORRECTION
No HAC Correction
No Absorbed FEs
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  37.874
                                                Number of obs =      1000
Total (centered) SS     =  2834382.139          Centered R2   =    0.4913
Total (uncentered) SS   =    4195421.4          Uncentered R2 =    0.6563
Residual SS             =  1441795.144
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.02863 | .2409828 | 4.27 | 0.000 | .5563128 | 1.500948 |
| X1_it | 1.228864 | .3320382 | 3.70 | 0.000 | .5780809 | 1.879647 |
| _cons | 11.61852 | 3.013075 | 3.86 | 0.000 | 5.713007 | 17.52404 |

We now estimate the same model accounting for correlation between errors from observations of the same individual (`id`), but we still assume no correlation between individuals, we do not consider the network structure yet. To do that we use the panel features (options `id()` and `time`) and we set the `lag()`

[13] This is equivalent of using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 Y_it X1_it (Z_it=IV_it) , robust`

option to be greater or equal than the maximum distance in time between observations, which in this case is 10footnoteThis is equivalent of clustering by individuals using ivreg2 (Baum *et al.*, 2003) and the following syntax: `ivreg2 Y_it X1_it (Z_it=IV_it) , cluster(id)`, or acreg: `acreg Y_it X1_it (Z_it=IV_it), cluster(id)`. `acreg` allows also to account for temporal correlation up to a certain lag, we decide this maximum distance in time through the option `lag()`.

```
. acreg Y_it X1_it  (Z_it=IV_it) , id(id) time(time) lag(10)
NO ARBITRARY CLUSTERING CORRECTION
No HAC Correction
No Absorbed FEs
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  30.295
                                                    Number of obs =     1000
Total (centered) SS     =  2834382.139              Centered R2   =   0.4913
Total (uncentered) SS   =    4195421.4              Uncentered R2 =   0.6563
Residual SS             =  1441795.144
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.02863 | .2720916 | 3.78 | 0.000 | .4953406 | 1.56192 |
| X1_it | 1.228864 | .3779895 | 3.25 | 0.001 | .4880181 | 1.96971 |
| _cons | 11.61852 | 3.042037 | 3.82 | 0.000 | 5.656242 | 17.58081 |

We now estimate the model above adding to the temporal correlation the correction proposed in our paper (Colella *et al.*, 2019). We assume that the error of each individual is correlated with the one of an other individual if they are linked in the network. To implement this in `acreg` we provide as input in the `links_mat` option the variables containing the adjacency matrix and we set *distcutoff* equal to 1[14]. Note that correlation between linked individuals but observed at different point in time is still assumed to be null.

```
. eststo: acreg  Y_it X1_it  (Z_it=IV_it),  id(id) time(time) lag(10) ///
> network  links_mat(clus*) dist(1)
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  10
No HAC Correction
No Absorbed FEs
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  22.720
                                                    Number of obs =     1000
Total (centered) SS     =  2834382.139              Centered R2   =   0.4913
Total (uncentered) SS   =    4195421.4              Uncentered R2 =   0.6563
Residual SS             =  1441795.144
```

[14]Note that the total number of observations in the database is *NT* (1000), but the total number of individuals is *N* (100). Since we are using the panel feature, `acreg` will require a link matrix formed by *N* variables, not *NT*.

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.02863 | .3842782 | 2.68 | 0.007 | .2754589 | 1.781802 |
| X1_it | 1.228864 | .4495232 | 2.73 | 0.006 | .3478147 | 2.109913 |
| _cons | 11.61852 | 4.743084 | 2.45 | 0.014 | 2.32225 | 20.9148 |

```
(est1 stored)
```

### 5.2.2 FE model

In the following example we replicate the previous model, accounting for both spatial and temporal correlation, but we add to the specification the *individual fixed effects* using the option `pfe1`.

```
. eststo: acreg  Y_it X1_it  (Z_it=IV_it),  id(id) time(time) lag(10) ///
> network  links_mat(clus*) dist(1) pfe1(id)
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  10
No HAC Correction
Absorbed FE: id
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  38.899
                                                    Number of obs =     1000
Total (centered) SS     =  2331112.842              Centered R2   =   0.4938
Total (uncentered) SS   =  2331112.842              Uncentered R2 =   0.4938
Residual SS             =  1180104.818
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.368636 | .346849 | 3.95 | 0.000 | .6888244 | 2.048448 |
| X1_it | .7942328 | .3663375 | 2.17 | 0.030 | .0762245 | 1.512241 |
| _cons | 9.58e-17 | 1.266864 | 0.00 | 1.000 | -2.483007 | 2.483007 |

```
nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2
(est2 stored)
```

We now add to the previous model also time fixed effects using the option `pfe2`.

```
. eststo: acreg  Y_it X1_it  (Z_it=IV_it),  id(id) time(time) lag(10) ///
> network  links_mat(clus*) dist(1) pfe1(id) pfe2(tim)
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  10
No HAC Correction
Absorbed FE: id and time
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  39.988
                                                    Number of obs =     1000
```

```
Total (centered) SS     =  2226516.365                Centered R2   =   0.4935
Total (uncentered) SS   =  2226516.365                Uncentered R2 =   0.4935
Residual SS             =  1127664.807
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.327506 | .3119844 | 4.26 | 0.000 | .7160278 | 1.938984 |
| X1_it | .8232877 | .3574087 | 2.30 | 0.021 | .1227796 | 1.523796 |
| _cons | -7.70e-17 | .9797572 | -0.00 | 1.000 | -1.920289 | 1.920289 |

```
nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2
(est3 stored)
```

## 5.3 Additional Options

### 5.3.1 Thresholds

Now we still account for spatial correlation between observations of the same year, but we do not account for any kind of temporal correlation. We do that by setting the lagcutoff at 0.

```
. eststo: acreg  Y_it X1_it  (Z_it=IV_it),  id(id) time(time) lag(0) ///
> network  links_mat(clus*) dist(1)
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  26.732
                                                      Number of obs =     1000
Total (centered) SS     =  2834382.139                Centered R2   =   0.4913
Total (uncentered) SS   =    4195421.4                Uncentered R2 =   0.6563
Residual SS             =  1441795.144
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.02863 | .3629168 | 2.83 | 0.005 | .3173265 | 1.739934 |
| X1_it | 1.228864 | .4116362 | 2.99 | 0.003 | .4220717 | 2.035656 |
| _cons | 11.61852 | 4.724562 | 2.46 | 0.014 | 2.358554 | 20.8785 |

```
(est4 stored)
```

Now we account for spatial correlation between observations of the same year, and also for temporal correlation between observations from the same individual, but only if they were observed with a lag lower than 3 years. We do that by setting the *lagcutoff* equal to 10.

```
. eststo: acreg  Y_it X1_it  (Z_it=IV_it),  id(id) time(time) lag(3) ///
> network  links_mat(clus*) dist(1)
NETWORK CORRECTION
```

```
DistCutoff: 1
LagCutoff:  3
No HAC Correction
No Absorbed FEs
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  21.277
                                                    Number of obs =     1000
Total (centered) SS     =  2834382.139              Centered R2   =   0.4913
Total (uncentered) SS   =    4195421.4              Uncentered R2 =   0.6563
Residual SS             =  1441795.144
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.02863 | .3783906 | 2.72 | 0.007 | .2869983 | 1.770262 |
| X1_it | 1.228864 | .4578899 | 2.68 | 0.007 | .3314161 | 2.126312 |
| _cons | 11.61852 | 4.824297 | 2.41 | 0.016 | 2.163077 | 21.07397 |

```
(est5 stored)
```

### 5.3.2 HAC

In the previous examples the matrix used for the computation of the variance covariance matrix is binary. We can use the option `hac` to have a linear decay in time and compute Heteroscedasticity-Autocorrelation-Consistent standard errors, following Newey and West (1987).

```
. eststo: acreg  Y_it X1_it  (Z_it=IV_it),  id(id) time(time) lag(3) ///
> network  links_mat(clus*) dist(1) hac
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  3
HAC Correction
No Absorbed FEs
Included instruments: X1_it
Instrumented: Z_it
Excluded instruments: IV_it
Kleibergen-Paap rk Wald F statistic:  22.477
                                                    Number of obs =     1000
Total (centered) SS     =  2834382.139              Centered R2   =   0.4913
Total (uncentered) SS   =    4195421.4              Uncentered R2 =   0.6563
Residual SS             =  1441795.144
```

| Y_it | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Z_it | 1.02863 | .3756538 | 2.74 | 0.006 | .2923624 | 1.764898 |
| X1_it | 1.228864 | .4442984 | 2.77 | 0.006 | .3580549 | 2.099673 |
| _cons | 11.61852 | 4.812064 | 2.41 | 0.016 | 2.187053 | 21.05 |

```
(est6 stored)
```

# 6 How does the acreg syntax compares with previously available commands for spatial?

In this section we show how our syntax compares with previous commands correcting standard errors for spatial correlation (Conley, 1999). We refer to two common commands: ols_spatial_HAC by Solomon Hsiang (Hsiang, 2010) and reg2hdfespatial by Thiemo Fetzer (Fetzer, 2015). The goal of this section is to provide users a better understanding of the differences between commands and syntax and facilitate the process to users that want to switch from one command to another.

## 6.1 Database

For this section we use an ad-hoc database that can be downloaded from our website. It is a balanced panel dataset on 1000 observations ($NT$) referring to 100 ($N$) individuals at 10 ($T$) points in time. Individuals are identified through the variable id, while time through the variable time. Database also contains, among others, the following variables Y_it, X1_it, Z_it, IV_it. Each individuals is located at one point in the space identified through the variables latitude (lat) and longitude (lon). In addition, we create dummies identifying individuals and time (id_d and time_d) variables and three additional variables necessary for some comparison: const is a constant variable taking always the value 1, fakeid and faketime are necessary to mimic a cross-sectional environment.

```
. use https://acregstata.weebly.com/uploads/2/9/1/6/29167217/acregfakedata.dta , clear
. qui tab time, g(time_d)
. qui tab id, g(id_d)
. gen const = 1
. gen fakeid = _n
. gen faketime = 1
```

## 6.2 Estimation

Before estimate the models, we need to install the three commands, acreg can be download by typing net install acreg, from(https://acregstata.weebly.com/uploads/2/9/1/6/29167217) replace, while the other two can be downloaded from Solomon Hsiang and Thiemo Fetzer websites. In addition, we have to install a few necessary packages through ssc: estout, ivreg2, ranktest, reg2hdfe, tmpdir. We do that by typing ssc install and the name of the command for each of them. The three codes will be compared in four different settings, for each setting we first estimate the same model using the three commands and we report a table comparing the results. Note that all regressions will be estimated through an OLS methodology, since previous commands do not support 2SLS regressions.

### 6.2.1 No correction

In this section we consider a pooled model in which we do not include any Random or Fixed Effects. We first estimate the model assuming that observations' errors are uncorrelated, we compare the three commands to the output of ivreg2 (Baum *et al.*, 2003). Note that previous commands require id and time to be specified, therefore we use the variables `fakeid` taking a different value for each observation and `faketime` which is a constant. To avoid the spatial correction we set the cutoffs at 0. The table below shows beta coefficients and standard errors for each of the 4 regressions. Column 1 reports ivreg2 results (`ivreg2`), column 2 refers to acreg (`acreg`), column 3 to ols_spatial_hac (`ols_spat`) and column 4 to reg2hdfespatial (`2hdfespat`).

```
. * ivreg2
. qui eststo: ivreg2 Y_it X1_it Z_it , robust

. * acreg
. qui eststo: acreg Y_it X1_it  Z_it

. * ols_spatial_HAC (ols_spat)
. qui eststo: ols_spatial_HAC Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(fakeid) timevar(time) distcutoff(0.000001) lagcutoff(0)

. * reg2hdfespatial (2hdfespat)
. qui eststo: reg2hdfespatial Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(fakeid) timevar(time) distcutoff(0.000001) lagcutoff(0) ///
> altfetime(faketime) altfepanel(faketime)

. * table of results
. esttab, cells(b se) keep( X1_it  Z_it) mtitles(ivreg2 acreg ols_spat 2hdfespat)
```

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | ivreg2 | acreg | ols_spat | 2hdfespat |
| | b/se | b/se | b/se | b/se |
| X1_it | .7389741 | .7389741 | .7389741 | .7389741 |
| | .188245 | .188245 | .188245 | .188245 |
| Z_it | 1.452726 | 1.452726 | 1.452726 | 1.452726 |
| | .0474316 | .0474316 | .0474316 | .0474316 |
| N | 1000 | 1000 | 1000 | 1000 |

### 6.2.2 Spatial Correction

We now estimate the model above using the spatial correction proposed by Conley (1999), with a threshold of 500 kilometers. This means that the error of each individual is assumed to be correlated with the ones of all the individuals that are located within a radius of 100 kilometers from it. Note that we keep on ignoring the panel dimension of the database assuming a cross-sectional environment. We estimate a model with a binary weighting matrix and a model allowing for weights in the weighting matrix linearly decreasing as the distance increases with values very close to one for near counties and almost zero for counties close to the distance cutoff using the option `bartlett`. This option is optional in `acreg` and in

*ols_spatial_hac*, but it is default in `reg2hdfespatial`. Therefore we run 5 regressions. The table below shows beta coefficients and standard errors for each of the 5 regressions. Columns 1-2 present estimates of a model with binary weights, columns 3-5 refer to a model with a decay in weights (`bartlett`).

```
. ***** binary weighting matrix
. * acreg
. qui eststo: acreg Y_it X1_it  Z_it , ///
> spatial latitude(lat) longitude(lon) dist(500)

. * ols_spatial_HAC (ols_spat)
. qui  eststo: ols_spatial_HAC Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(fakeid) timevar(faketime)  distcutoff(500) lagcutoff(0)

. ***** linear bartlett window
. * acreg
. qui eststo: acreg Y_it X1_it  Z_it , ///
> spatial latitude(lat) longitude(lon) dist(500)  bartlett

. * ols_spatial_HAC (ols_spat)
. qui  eststo: ols_spatial_HAC Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(fakeid) timevar(faketime)  distcutoff(500) lagcutoff(0) bartlett

. * reg2hdfespatial (2hdfespat)
. qui eststo: reg2hdfespatial Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(fakeid) timevar(time) distcutoff(500) lagcutoff(0) ///
> altfetime(faketime) altfepanel(faketime)

. * table of results
. esttab, cells(b se) keep( X1_it  Z_it) mtitles(acreg ols_spat acreg_b ols_spat_b 2hdfespat_b)
```

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | acreg | ols_spat | acreg_b | ols_spat_b | 2hdfespat_b |
| | b/se | b/se | b/se | b/se | b/se |
| X1_it | .7389741 | .7389741 | .7389741 | .7389741 | .7389741 |
| | .2212767 | .2212767 | .207783 | .207783 | .1960521 |
| Z_it | 1.452726 | 1.452726 | 1.452726 | 1.452726 | 1.452726 |
| | .0486929 | .0486929 | .0487222 | .0487222 | .0502401 |
| N | 1000 | 1000 | 1000 | 1000 | 1000 |

### 6.2.3 Spatial and Temporal Correction

We now estimate the model above in a panel setting[15] using both the spatial correction proposed by Conley (1999), with a threshold of 500 kilometers and also the temporal correlation between errors from observations of the same individual (`id`). Note that the option `hac` used to have a linear decay in time and compute Heteroscedasticity-Autocorrelation-Consistent standard errors, following Newey and West (1987) is default in all the commands but `acreg`, therefore we specify it in the following regressions. At the same way as above, we estimate a model with a binary weighting matrix and a model using the `bartlett` option. The table below shows beta coefficients and standard errors for each of the 5 regressions. Columns

[15] Note that we do not include any type of fixed or random effects, but we consider individuals and time in the computation fo the standard errors.

1-2 present estimates of a model with binary weights, columns 3-5 refer to a model with a decay in weights (`bartlett`).

```
. ***** binary weighting matrix
. * acreg
. qui eststo: acreg Y_it X1_it  Z_it , ///
> spatial latitude(lat) longitude(lon) dist(500) id(id) time(time) lag(10) hac

. * ols_spatial_HAC (ols_spat)
. qui  eststo: ols_spatial_HAC Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(id) timevar(time)  distcutoff(500) lagcutoff(10)

. ***** linear bartlett window
. * acreg
. qui eststo: acreg Y_it X1_it  Z_it , ///
> spatial latitude(lat) longitude(lon) dist(500) id(id) time(time) lag(10) hac  bartlett

. * ols_spatial_HAC (ols_spat)
. qui  eststo: ols_spatial_HAC Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(id) timevar(time)  distcutoff(500) lagcutoff(10)  bartlett

. * reg2hdfespatial (2hdfespat)
. qui eststo: reg2hdfespatial Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(id) timevar(time) distcutoff(500) lagcutoff(10) ///
> altfetime(faketime) altfepanel(faketime)

. * table of results
. esttab, cells(b se) keep( X1_it  Z_it) mtitles(acreg ols_spat acreg_b ols_spat_b 2hdfespat_b)
```

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | acreg | ols_spat | acreg_b | ols_spat_b | 2hdfespat_b |
| | b/se | b/se | b/se | b/se | b/se |
| X1_it | .7389741 | .7389741 | .7389741 | .7389741 | .7389741 |
| | .209021 | .209021 | .1952831 | .1952831 | .1952831 |
| Z_it | 1.452726 | 1.452726 | 1.452726 | 1.452726 | 1.452726 |
| | .0548461 | .0548461 | .0509812 | .0509812 | .0509812 |
| N | 1000 | 1000 | 1000 | 1000 | 1000 |

### 6.2.4 Spatial and Temporal Correction and high dimensional FEs

In the following example we replicate the previous model, accounting for both spatial and temporal correlation, and we add to the specification the *individual fixed effects* using the option `pfe1` and the time fixed effects using the option `pfe2`. At the same way as above, we estimate a model with a binary weighting matrix and a model using the `bartlett` option. The table below shows beta coefficients and standard errors for each of the 5 regressions. Columns 1-2 present estimates of a model with binary weights, columns 3-5 refer to a model with a decay in weights (`bartlett`).

```
. ***** binary weighting matrix
. * acreg
. qui eststo: acreg Y_it X1_it  Z_it ,  pfe2(id) pfe1(time) ///
> spatial latitude(lat) longitude(lon) dist(500) id(id) time(time) lag(10) hac
```

```
. * ols_spatial_HAC (ols_spat)
. qui  eststo: ols_spatial_HAC Y_it X1_it Z_it const id_d2-id_d100 time_d2-time_d10 , ///
> lat(lat) lon(lon) panelvar(id) timevar(time)  distcutoff(500) lagcutoff(10)

. ***** linear bartlett window
. * acreg
. qui eststo: acreg Y_it X1_it  Z_it ,  pfe2(id) pfe1(time) ///
> spatial latitude(lat) longitude(lon) dist(500) id(id) time(time) lag(10) hac  bartlett

. * ols_spatial_HAC (ols_spat)
. qui  eststo: ols_spatial_HAC Y_it X1_it Z_it const id_d2-id_d100 time_d2-time_d10 , ///
> lat(lat) lon(lon) panelvar(id) timevar(time)  distcutoff(500) lagcutoff(10)  bartlett

. * reg2hdfespatial (2hdfespat)
. qui eststo: reg2hdfespatial Y_it X1_it Z_it const , ///
> lat(lat) lon(lon) panelvar(id) timevar(time) distcutoff(500) lagcutoff(10)

. * table of results
. esttab, cells(b se) keep( X1_it  Z_it) mtitles(acreg ols_spat acreg_b ols_spat_b 2hdfespat_b)
```

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | acreg | ols_spat | acreg_b | ols_spat_b | 2hdfespat_b |
| | b/se | b/se | b/se | b/se | b/se |
| X1_it | .7398598 | .7398598 | .7398598 | .7398598 | .7398598 |
| | .2183089 | .2183089 | .1972198 | .1972198 | .1972198 |
| Z_it | 1.418412 | 1.418412 | 1.418412 | 1.418412 | 1.418412 |
| | .0491794 | .0491794 | .0488122 | .0488122 | .0488122 |
| N | 1000 | 1000 | 1000 | 1000 | 1000 |

# References


Baum, C. F., Schaffer, M. E., and Stillman, S. (2003). Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, **3**(1), 1–31.

Colella, F., Lalive, R., Sakalli, S. O., and Thoenig, M. (2019). Inference with arbitrary clustering. *IZA Discussion Paper*.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, **92**(1), 1–45.

Correia, S. (2016). A feasible estimator for linear models with multi-way fixed effects.

Fetzer, T. (2015). Conley spatial hac standard errors for models with fixed effects.

Grund, T. U. and Densley, J. A. (2012). Ethnic heterogeneity in the activity and structure of a black street gang. *European Journal of Criminology*, **9**(4), 388–406.

Grund, T. U. and Densley, J. A. (2015). Ethnic homophily and triad closure: Mapping internal gang structure using exponential random graph models. *Journal of Contemporary Criminal Justice*, **31**(3), 354–370.

Hsiang, S. M. (2010). Temperatures and cyclones strongly associated with economic production in the caribbean and central america. *Proceedings of the National Academy of Sciences*, **107**(35), 15367–15372.

Messner, S. F., Anselin, L., Baller, R. D., Hawkins, D. F., Deane, G., and Tolnay, S. E. (1999). The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. *Journal of Quantitative criminology*, **15**(4), 423–450.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, **55**(3), 703–708.