The Stata Journal (2023)
23, Number 1, pp. 119–147 DOI: 10.1177/1536867X231162031

acreg: Arbitrary correlation regression

Fabrizio Colella
University College London
London, U.K.
f.colella@ucl.ac.uk

Rafael Lalive
HEC Lausanne
Lausanne, Switzerland
rafael.lalive@unil.ch

Seyhun Orcan Sakalli
King's College London
London, U.K.
seyhun.sakalli@kcl.ac.uk

Mathias Thoenig
HEC Lausanne
Lausanne, Switzerland
mathias.thoenig@unil.ch

Abstract. We present `acreg`, a new command that implements the arbitrary clustering correction of standard errors proposed in Colella et al. (2019, IZA discussion paper 12584). Arbitrary here refers to the way observational units are correlated with each other: we impose no restrictions so that our approach can be used with a wide range of data. The command accommodates both cross-sectional and panel databases and allows the estimation of ordinary least-squares and two-stage least-squares coefficients, correcting standard errors in three environments: in a spatial setting using units' coordinates or distance between units, in a network setting starting from the adjacency matrix, and in a multiway clustering framework taking multiple clustering variables as input. Distance and time cutoffs can be specified by the user, and linear decays in time and space are also optional.

Keywords: st0703, acreg, spatial correlation, time correlation, inference, spatial data, network data

1 Introduction

Thanks to increasing computational power, databases have become more complex in the past decades. They now embed convoluted correlation structures between observational units that were not common before. For example, fueled by the growing availability of geocoded data and the integration of geographic information systems in the toolkit of economists, empirical works using spatial data are proliferating in fields like development economics, urban economics, and economic history. Other examples of new correlation structures pertain to network data: individuals are linked, and these links are now measurable through social networks, mobile data, coworking relations, or coauthorships.

Statistical inference in these environments is challenging because the underlying data-generating process is often unknown and researchers need to make assumptions about the relationship between observations. Available methods to address the correlation between objects build on the sandwich-type variance–covariance (VCV) estimator proposed by White (1980). The most common approach is standard clustering (Cameron and Miller 2015), which defines clusters as groups of linked observations that share a common characteristic. With spatial data, a frequently used approach has been developed by Conley (1999), who considers a circle around each unit, within which the

strength of the dependence between the unit and the surrounding ones is specified. In the case of network data, the practice is less developed; many studies simply do not correct for the potential correlation of unobserved shocks across linked observations.

In our companion article Colella et al. (2019), we explore pitfalls and provide guidelines for conducting inference in complex settings, allowing for any type of topological and temporal dependence between observational units in large samples. Our arbitrary clustering approach builds on the seminal insight by White (1980), using estimated regression errors and knowledge on the clustering structure to reconstruct estimates of the unknown elements of the sandwich formula. We perform extensive Monte Carlo simulations for both spatial and network data structures, for example, U.S. counties and coauthorship in economics. Our simulation results show that arbitrary clustering inference dominates inference based on conventional estimators.

In this article, we present our new community-contributed command `acreg`, which implements the arbitrary clustering correction of standard errors proposed in Colella et al. (2019). We also provide several examples of how to use it. Our command accommodates ordinary least-squares (OLS) and two-stage least-squares (2SLS) estimations and is designed to deal with several clustered covariance matrix estimators, including multiway clustering (Cameron, Gelbach, and Miller 2011), spatial clustering (Conley 1999; Bester, Conley, and Hansen 2011), network clustering, and heteroskedasticity- and autocorrelation-consistent (HAC) (Newey and West 1987).

In network settings, to the best of our knowledge, there is no Stata command designed to correct standard errors starting from the knowledge of the binary links between observations.

In spatial settings, three community-contributed commands are available (Conley 1999; Hsiang 2010; and Fetzer 2015): however, they suit only OLS estimation. In addition, they all have preset options that are not desirable in all settings. In particular, the commands by Conley (1999) and Fetzer (2015) impose a linear decay in the correlation structure between units (Bartlett), while Hsiang (2010) and Fetzer (2015) set a time decay (HAC) as the default.[1] Compared with those commands, `acreg` is more flexible because it enables the user to freely set the type of correlation structure and decay across observations and time. Moreover, in the presence of multiple cross-sectional observational units sharing the same geolocation, our command provides consistent standard errors, replicating the heteroskedasticity-robust standard errors from `ivreg2` (Baum, Schaffer, and Stillman 2003) when the distance correction is set to zero, while the programs by Conley (1999), Hsiang (2010), and Fetzer (2015) do not. Stata 15 introduced a series of commands named `sp` to model spatial relations between objects using spatial autoregressive models. These models allow for spatial lags of the dependent variable, which modifies point estimates, or for spatial autocorrelation in the errors. The command closest to ours, `spregress`, allows only for heteroskedasticity-robust and asymptotic maximum-likelihood theory-driven standard errors. Conversely, `acreg` does not modify the point estimates but improves inference by computing standard errors corrected for spatial correlation.

1. Conley (1999) allows correction only for cross-sectional dependence and not time dependence.

Concerning multiway cluster–robust standard errors, `ivreg2` and `xtivreg2` (Schaffer 2005) allow the user to specify up to two cluster variables (that is, two-way clustering). The community-contributed command by Gelbach and Miller (2009), `cgmreg`, instead accommodates multiway clustering but suits only OLS estimation and does not allow for the estimation of 2SLS models. `acreg` instead can be used to estimate both OLS and 2SLS coefficients, correcting standard errors for an infinite number of cluster dimensions.

The rest of this article is organized as follows. In section 2, we review the arbitrary clustering method proposed in Colella et al. (2019). In section 3, we provide a detailed description of the syntax of `acreg`. In section 4, we offer an illustration of our command with several examples in the spatial and the network settings: we show how options of our command can be used to suit many models of correlation structure. Finally, in section 5, we conclude.

2 Estimator for the VCV matrix

Here we present the estimator of the VCV proposed in Colella et al. (2019). The proposed estimator builds on the seminal insight from White (1980) and can be seen as an extension of the one-way or multiway clustering (Cameron, Gelbach, and Miller 2011) that also includes spatial clustering (Conley 1999; Bester, Conley, and Hansen 2011).[2]

In our setting, each observation can be correlated to any other, and the strength of their correlation is a function of both time and distance. We define a matrix $\mathbf{S}$, named pattern matrix, containing information on cross-observation correlations in errors. With spatial data, $\mathbf{S}$ is built from information on the geographic distance between spatial units, for example, regions, cities, and countries; in a network context, it reflects the direct links between observations at different degrees. `acreg` computes the matrix $\mathbf{S}$ starting from the position of objects in space, using their coordinates, or from the link structure in a network; it also allows the user to define the matrix $\mathbf{S}$ to accommodate more complex correlation structures. Entries of the $\mathbf{S}$ matrix range from 0 to 1: this measure represents the strength of the correlation between two units and is inversely proportional to their distance. The diagonal of $\mathbf{S}$ is a vector of ones, reflecting the self-links.

Consider n observations at each t instant of time T from the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where we observe each unit i several times in different periods t. $\mathbf{y}$ is a dependent variable, and $\mathbf{X}$ is a matrix of k linearly independent components. $\mathbf{X}$ could include a

2. We do not provide any theoretical or empirical validation of our approach here. In Colella et al. (2019), we show results of extensive Monte Carlo simulations based on real-life data on U.S. metropolitan areas or on coauthors in economics. We show that our arbitrary clustering estimator of the VCV yields inference at the correct significance level in moderately sized samples and that it always dominates other commonly used approaches to inference. We provide guidance to the applied practitioners on how to cluster and to make reasonable assumptions on the error distribution in absence of prior knowledge about the data-generating process.

long list of dummies for each unit in case we are interested in the within estimates in a panel dataset. The OLS estimator can be written as

$$\mathbf{b}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

and the theoretical VCV of the $\mathbf{b}_{\text{OLS}}$ is

$$\text{VCV}(\mathbf{b}_{\text{OLS}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where $\mathbf{\Omega} \equiv E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X})$ is the unknown VCV of $\boldsymbol{\epsilon}$.

The VCV is estimated by the sandwich estimator (White 1980)

$$\widehat{\text{VCV}}(\mathbf{b}_{\text{OLS}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left\{\mathbf{S} \times (\mathbf{e}\mathbf{e}')\right\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where $\mathbf{e} \equiv \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{OLS}}$ represents the vector of residuals, $\mathbf{S}$ is the pattern matrix, and $\times$ is element-by-element matrix multiplication. The key element of this estimator is the middle part $\mathbf{X}'\{\mathbf{S} \times (\mathbf{e}\mathbf{e}')\}\mathbf{X}$:

$$\mathbf{X}'\left\{\mathbf{S} \times (\mathbf{e}\mathbf{e}')\right\}\mathbf{X} = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T}\mathbf{x}_{it}e_{it}e_{js}\mathbf{x}'_{js}s_{itjs}$$

$\mathbf{x}_{it}$ is the (column) vector of regressors, and $\mathbf{x}'_{it}$ is the row it in matrix $\mathbf{X}$.

This framework can also be used in situations with endogeneity. We refer the reader to our article (Colella et al. 2019) for an illustration of the 2SLS version of the estimator.

3 The acreg command

`acreg` requires the installation of the latest versions of `ranktest`, `ivreg2` (Baum, Schaffer, and Stillman 2003), and `hdfe` (Correia 2016). It is possible to check whether the most up-to-date versions of these packages are installed (and to install them if they are not) by typing `acregpackcheck` after having installed `acreg`.

3.1 Syntax

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

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*.

`fweight`s and `pweight`s are allowed; see [U] **11.1.6 weight**.

3.2 Options

3.2.1 Panel

`id(`*idvar*`)` specifies the cross-sectional unit identifier named *idvar*; `id()` is required in a panel setting.

`time(`*timevar*`)` specifies the time unit variable named *timevar*; `time()` is required in a panel setting.

The model is assumed to be cross-sectional if `id()` and `time()` are not specified.

3.2.2 Spatial environment

`spatial` specifies that the environment is a spatial one; `spatial` is not required if arbitrary cluster correction is not performed or if the `weights()`, `cluster()`, or `network` option is specified.

`latitude(`*latitudevar*`)` sets the variable named *latitudevar*, which contains the latitude of each observation in decimal degrees: range$[-180.0, 180.0]$.

`longitude(`*longitudevar*`)` sets the variable named *longitudevar*, which contains the longitude of each observation in decimal degrees: range$[-180.0, 180.0]$.

`dist_mat(`*varlist_distances*`)` sets the N variables, listed in *varlist_distances*, containing bilateral distances between observations. In the spatial environment, bilateral distance is the spatial distance between observations, for example, physical or travel distance between two locations. If `dist_mat()` is specified, `latitude()` and `longitude()` may not be used.

`distcutoff(`#`)` specifies the distance cutoff, beyond which the correlation between the error terms of two observations is assumed to be zero; `distcutoff()` is required if `latitude()` and `longitude()` are specified or if `dist_mat()` is specified. The distance cutoff is in kilometers if `latitude()` and `longitude()` are specified. It can be in any other meaningful metric if bilateral distances are specified. # may be an integer or a float.

`lagcutoff(`#`)` specifies the time lag cutoff for those observations with the same *idvar*; `lagcutoff()` is not required in the cross-sectional environment. The default in the panel environment is `lagcutoff(0)`, that is, when the `id()` and `time()` options are

specified. In the panel environment when `lagcutoff(#)` is not specified, standard errors are clustered at *idvar* × *timevar* level. # must be an integer.

3.2.3 Network environment

`network` specifies that the environment is a network one; `network` is not required if arbitrary cluster correction is not performed and if the `weights()`, `cluster()`, or `spatial` option is specified.

`links_mat(`*varlist_links*`)` sets the N dummy variables, listed in *varlist_links*, specifying the links between observations, that is, the adjacency matrix. The links between two units can change over time. However, if `distcutoff()` is set to be greater than one, only the first observation in time of each individual will be used as input to compute the bilateral distance between two nodes.

`dist_mat(`*varlist_distances*`)` sets the N variables, listed in *varlist_distances*, containing bilateral distances between observations. In the network environment, bilateral distance is the network distance between observations, that is, the number of links along the shortest path between two nodes. If `dist_mat()` is specified, `links_mat()` may not be used.

`distcutoff(#)` specifies the distance cutoff (geodesic paths), beyond which the correlation between error terms of two observations is assumed to be zero; `distcutoff()` is required if `dist_mat()` is specified; it is optional if `links_mat()` is specified. The default is `distcutoff(1)` in the network environment. When `links_mat()` is specified and `distcutoff()` is greater than 1, `acreg` automatically computes the bilateral distance between two nodes. # may be an integer or a float.

`lagcutoff(#)` specifies the time lag cutoff for those observations with the same *idvar*. `lagcutoff()` is not required in the cross-sectional environment. The default in a panel environment is `lagcutoff(0)`, that is, when the `id()` and `time()` options are specified. In the panel environment when `lagcutoff(0)` is not specified, standard errors are clustered at *idvar* × *timevar* level. # must be an integer.

3.2.4 Multiway clustering environment

`cluster(`*varlist_cluster*`)` sets the variables, listed in *varlist_cluster*, to use for multiway clustered standard errors. `cluster()` is not required if arbitrary cluster correction is not performed and if the `spatial`, `network`, or `weights()` option is specified.

3.2.5 Arbitrary clustering environment

`weights(`*varlist_weights*`)` sets the $N \times T$ variables, listed in *varlist_weights*, containing the weights that will be used for error correction; `weights()` is not required if the `spatial`, `network`, or `cluster()` option is specified. The $N \times T$ variables need to follow the same order of the observations.

3.2.6 Correlation structure

`hac` reports HAC standard errors; `lagcutoff()` will be the temporal decay; `hac` requires `id()`, `time()`, and `lagcutoff()`.

`bartlett` imposes a distance linear decay between observations within the cutoff in the correlation structure.

`nbclust(#)` sets the number of clusters used to compute the Kleibergen–Paap statistic in case of arbitrary cluster correction; the default is `nbclust(100)`.

3.2.7 High-dimensional fixed effects

`pfe1(`*fe1var*`)` sets the categorical variable named *fe1var*, which identifies the first high-dimensional fixed effects to be absorbed.

`pfe2(`*fe2var*`)` sets the categorical variable named *fe2var*, which identifies the second high-dimensional fixed effects to be absorbed.

`correctr2` reports the R^2 of the overall model when `pfe1()` or `pfe2()` is specified, that is, the R^2 obtained before partialing out the high-dimensional fixed effects. The default reported R^2 is the R^2 of the within model when `pfe1()` or `pfe2()` is specified, that is, on the "partialed-out sample". `correctr2` is not allowed with `fweights`.

`dropsingletons` drops singleton groups when `pfe1()` or `pfe2()` is specified.

3.2.8 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 `weights()`; `storeweights` is optional only if the `spatial`, `network`, or `cluster()` option 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 `dist_mat()`; `storedistances` is optional only if the `spatial` option or `network` option is specified and `dist_mat()` is not specified.

3.3 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 R^2 ($1 -$ `e(rss)/e(tss)`)
`e(r2u)`	uncentered R^2
`e(widstat)`	Kleibergen–Paap rk Wald F statistic

Matrices

`e(b)`	coefficient vector
`e(V)`	corrected VCV matrix of the estimators

Functions

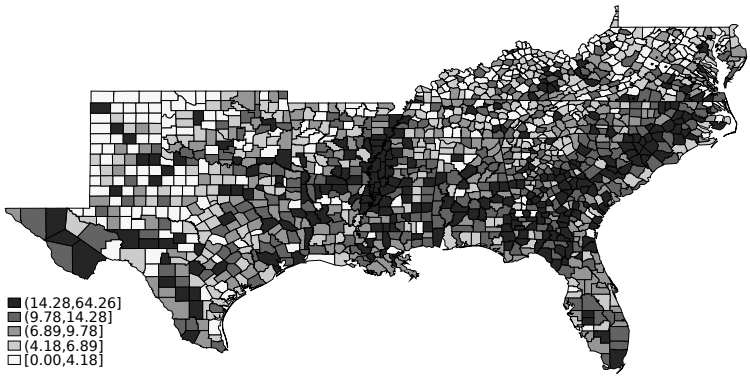
`e(sample)`	marks estimation sample

4 Examples

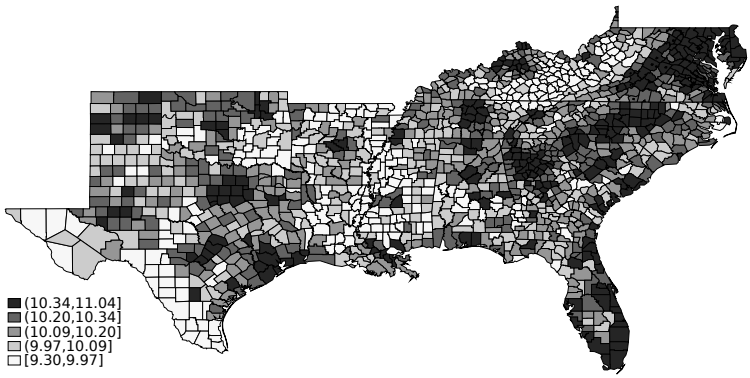
We illustrate the use of our command in five environments: spatial and network settings in both cross-sectional and panel contexts, and multiway clustering. In every environment, we estimate the same equation imposing different assumptions on the error correlation structure: independent and identically distributed, standard clustering, and arbitrary clustering.

4.1 Spatial environment, cross-sectional setting

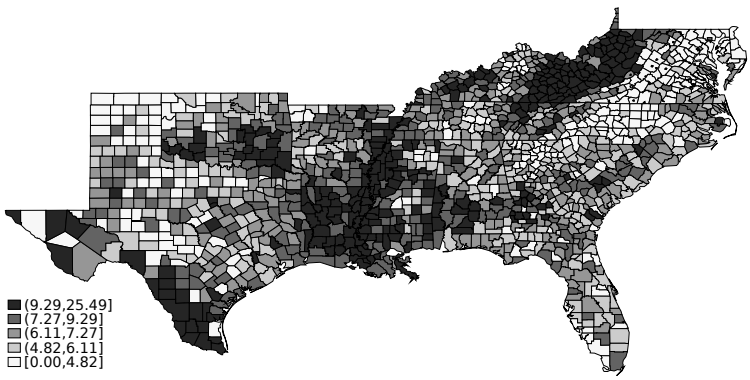
For this example, we use the data on the homicides in southern states of the United States. `homicide_1960_1990.dta` is available at the Stata website. The data contain, among others, the county-level homicide rate per year per 100,000 persons (`hrate`), the population in logs (`ln_population`), the logarithm of the average income (`ln_income`), the unemployment rate (`unemployment`), and the average age (`age`). This dataset is an extract of the data originally used by Messner et al. (1999) and concerns four different periods (1960, 1970, 1980, and 1990). We consider only the cross-sectional database for 1990, and we estimate the effect of income on homicide rate, controlling for population and age. For the sake of illustration, we claim that income is endogenous, and we assume that unemployment is a valid instrument for it. Figure 1 shows the spatial dependency of the outcome variable, the endogenous regressor, and the instrument.



(a) Homicide rate



(b) Income



(c) Unemployment

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

We first fit the model assuming that observations' errors are uncorrelated.[3]

```
. webuse homicide1990
(S.Messner et al.(2000), U.S southern county homicide rates in 1990)
. acreg hrate ln_population age (ln_income=unemployment)
HETEROSKEDASTICITY ROBUST STANDARD ERRORS
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	Coefficient	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

We then fit the model above clustering standard errors by state.[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	Coefficient	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

3. This is equivalent to using `ivreg2` (Baum, Schaffer, and Stillman 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), robust`.
4. This is equivalent to using `ivreg2` (Baum, Schaffer, and Stillman 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), cluster(sfips)`. We are aware that the number of states (clusters) is small and inference would suffer from it, but this is irrelevant to the scope of this exercise.

We also now fit the model above using a spatial correction following Conley (1999), with a threshold of 100 kilometers and without imposing a linear decay in the spatial correlation between units. This means that the error of each county is assumed to be correlated with the counties that are located within a radius of 100 kilometers.

```
. acreg hrate ln_population age (ln_income=unemployment),
> spatial latitude(_CX) longitude(_CY) distcutoff(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	Coefficient	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

4.1.1 Additional options

Thresholds. If we want to account for correlation between counties at a greater distance, we can increase the distance cutoff using the `distcutoff()` 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) distcutoff(200)
  (output omitted)
. estimates store sp1
```

Bartlett. In previous examples, the matrix used for the computation of the VCV matrix is binary: for each county pair, it contains 1 if they are located within the distance threshold from each other and 0 otherwise. `acreg` allows for weights in the matrix to linearly decrease as the distance between units increases. To do that, we need to add only the option `bartlett` to the syntax.

```
. acreg hrate ln_population age (ln_income=unemployment),
> spatial latitude(_CX) longitude(_CY) distcutoff(200) bartlett
  (output omitted)
. estimates store sp2
```

Partial out high-dimensional fixed effects. `acreg` allows for adding high-dimensional fixed effects and partialing 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 fit the previous model by adding state fixed effects.

```
. acreg hrate ln_population age (ln_income=unemployment),
> spatial latitude(_CX) longitude(_CY) distcutoff(100) pfe1(sfips)
  (output omitted)
. estimates store sp3
```

The following code (Jann 2007, 2014) reports the results of the three estimations in this subsection:

```
. esttab sp1 sp2 sp3, cells(b se) keep(ln_income ln_population age)
> mtitles(spatial bartlett FE)
```

	(1) spatial b/se	(2) bartlett b/se	(3) FE b/se
ln_income	-8.822082	-8.822082	-13.88229
	2.733507	2.313018	1.835268
ln_populat~n	1.404433	1.404433	1.649735
	.4834539	.4388646	.4000578
age	-.281615	-.281615	-.178832
	.1223503	.1015135	.0960779
N	1412	1412	1412

4.2 Spatial environment, panel setting

Here we use the database we used in the previous section: `homicide_1960_1990.dta`. We again estimate the effect of income on homicide rate, controlling for population and age, and we assume that unemployment is a valid instrument for it. Compared with the previous section, here we use all four waves of the dataset.

4.2.1 Pooled model

We first consider a pooled model in which we do not include any random or fixed effects. We first fit the model assuming that observations' errors are uncorrelated.[5]

```
. webuse homicide_1960_1990, clear
(S.Messner et al.(2000), U.S southern county homicide rate in 1960-1990)
. acreg hrate ln_population age (ln_income=unemployment)
HETEROSKEDASTICITY ROBUST STANDARD ERRORS
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	Coefficient	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

We then fit the same model, but we use the panel feature of `acreg` to account for autocorrelation between observations from the same county over time.[6] We assume no correlation across counties. We specify the option `id()` with the county ID, the option `time()` with the `year` variable, and the option `lagcutoff()` with a number equal to or greater than the maximum lag between observations, which in this case is 30.[7]

5. This is equivalent to using `ivreg2` (Baum, Schaffer, and Stillman 2003) and the following syntax: `ivreg2 hrate ln_population age (ln_income=unemployment), robust`.
6. The estimation of the betas does not change with respect to the previous model. `acreg` is used only to compute the standard errors.
7. This is equivalent to using `ivreg2` (Baum, Schaffer, and Stillman 2003) and the syntax `ivreg2 hrate ln_population age (ln_income=unemployment), cluster(_ID)` or, alternatively, using `acreg` and the syntax `acreg hrate ln_population age (ln_income=unemployment), cluster(_ID)`.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(30)
TEMPORAL 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	Coefficient	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 then extend the model above, which accounts for autocorrelation over time, by adding the spatial correction proposed by Conley (1999), with a threshold of 100 kilometers. This means that the error term of each county at a given year is assumed to be correlated with those of all the counties that are located within a radius of 100 kilometers from it observed at the same year while simultaneously correcting for autocorrelation over time for each county. We assume the correlation between near counties but observed at different points in time to be zero.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(30) spatial latitude(_CX) longitude(_CY) distcutoff(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
```

hrate	Coefficient	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

4.2.2 Fixed-effects model

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

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(30) spatial latitude(_CX) longitude(_CY) distcutoff(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	Coefficient	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 also add time fixed effects to the previous model, using the option `pfe2()`.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(30) spatial latitude(_CX) longitude(_CY) distcutoff(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	Coefficient	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
```

4.2.3 Additional options

Thresholds. In the next example, we account for spatial correlation between observations of the same year without accounting for any temporal correlation. We do this with `lagcutoff(0)`.[8]

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(0) spatial latitude(_CX) longitude(_CY) distcutoff(100)
  (output omitted)
. estimates store spp1
```

Then, 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; that is, two observations from the same county are assumed to be correlated only if they are observed with a 10-year or less difference.[9] We do that by setting `lagcutoff(10)`.

8. The result differs from the one obtained in the cross-sectional environment (`acreg hrate ln_population age (ln_income=unemployment), spatial latitude(_CX) longitude(_CY) distcutoff(100)`) because the spatial correlation is assumed to be present only between observations from the same year.
9. This would allow an observation's error term to be correlated with all other observations within 10-year lags and 10-year leads from the same county.

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(10) spatial latitude(_CX) longitude(_CY) distcutoff(100)
  (output omitted)
. estimates store spp2
```

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

```
. acreg hrate ln_population age (ln_income=unemployment), id(_ID) time(year)
> lagcutoff(30) spatial latitude(_CX) longitude(_CY) distcutoff(100) hac
  (output omitted)
. estimates store spp3
```

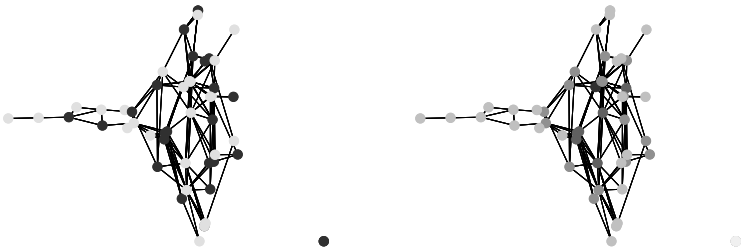
The following code reports the result of the three estimations in this subsection.

```
. esttab spp1 spp2 spp3, cells(b se) keep(ln_income ln_population age)
> mtitles(lag0 lag10 hac)
```

	(1) lag0 b/se	(2) lag10 b/se	(3) hac b/se
ln_income	3.83872	3.83872	3.83872
	1.743993	1.801373	1.785354
ln_populat~n	-.4411802	-.4411802	-.4411802
	.3542752	.377059	.3727145
age	-.4626917	-.4626917	-.4626917
	.1347804	.1403627	.139132
N	5648	5648	5648

4.3 Network environment, cross-sectional setting

In this section, we use a dataset of cooffending in a London-based youth gang. Data were collected by James Densley and Thomas Grund. The data have been used in Grund and Densley (2012, 2015). Information on 54 individuals is reported, and 2 individuals are recorded to be linked if they committed at least one crime together. The data contain, 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 cooffending network are stored in 54 variables (`_net2_1`–`_net2_54`). Figure 2 presents the distribution of the variables `Arrest` and `Ranking` within the network. In this example, we want to estimate the effect of ranking on arrests, controlling for age, residence, and birthplace fixed effects.



(a) Arrest (b) Ranking

Figure 2. Gang network. NOTES: In panel (a), dark dots represent arrested people. In panel (b), darker dots 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), which is needed because the original database does not contain self-links. `webnwuse` and `nwload` were written by Grund (2015).

```
. webnwuse gang, clear
Loading successful
(4 networks)
─────────────────────

     gang_valued
     gang
     gang_valued_1
     gang_1
. nwload gang
. forvalues j = 1(1)54 {
  2. quietly replace _net2_`j' = 1 in `j'
  3. }
```

We first fit the model assuming that observations' errors are uncorrelated.[10]

```
. acreg Arrest Ranking Age Residence i.Birthplace
HETEROSKEDASTICITY ROBUST STANDARD ERRORS
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	Coefficient	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 fit the same model using the standard error correction proposed in our article (Colella et al. 2019). We assume that the error term of each observation is correlated with that of another if they are linked in the network. To implement this in `acreg`, we provide the variables containing the adjacency matrix as input in the `links_mat()` option and set `distcutoff(1)`.

10. This is equivalent to `ivreg2 Arrest Ranking Age Residence i.Birthplace, robust`.

```
. acreg Arrest Ranking Age Residence i.Birthplace, network links_mat(_net2_*)
> distcutoff(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	Coefficient	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
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

4.3.1 Additional options

Accounting for degree greater 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 account 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 also want to allow for correlation between individuals that are linked through another individual (degree 2). To do that, we will use the same syntax but change `distcutoff(2)`.

```
. acreg Arrest Ranking Age Residence i.Birthplace, network links_mat(_net2_*)
> distcutoff(2)
  (output omitted)
. estimates store ne1
```

Bartlett. In previous examples, the matrix used for the computation of the VCV matrix is binary: it contains values 1 for each pair of individuals that are first- or second-degree linked, and 0s otherwise. `acreg` allows for weights in the matrix to linearly decrease as the network distance increases. To do that in our sample, that is, having 1 for first-degree linked observations and 0.5 for second-degree linked observations, we will use the option `bartlett`.

```
. acreg Arrest Ranking Age Residence i.Birthplace, network links_mat(_net2_*)
> distcutoff(2) bartlett
  (output omitted)
. estimates store ne2
```

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 fit the previous model partialing out birthplace fixed effects instead of adding them as dummies in the main regression.

```
. acreg Arrest Ranking Age Residence, network links_mat(_net2_*) distcutoff(1)
> pfe1(Birthplace)
  (output omitted)
. estimates store ne3
```

The following code reports the result of the three estimations in this subsection.

```
. esttab ne1 ne2 ne3, cells(b se) keep(Ranking Age Residence)
> mtitles(degree2 bartlett FE)
```

	(1) degree2 b/se	(2) bartlett b/se	(3) FE b/se
Ranking	-2.168476	-2.168476	-2.168476
	.4801238	.7688551	.7132431
Age	.7665194	.7665194	.7665194
	.4001636	.3427023	.3730319
Residence	-1.534665	-1.534665	-1.534665
	2.138931	1.590511	1.618858
N	54	54	54

4.4 Network environment, panel setting

In this section, we use an ad hoc database, which can be downloaded from our command's website. It is a balanced panel dataset of 1,000 observations (NT) referring to 100 (N) individuals at 10 (T) points in time. Individuals are identified through the variable `id`, while time is identified through the variable `time`. The database also contains, among others, the following variables: `Y_it`, `X1_it`, `End_it`, and `IV_it`. The symmetric binary links constituting the network are stored in 100 (N) variables (`clus_1`–`clus_100`). In this example, we want to estimate the effect of `End_it` on `Y_it`, controlling for `X_it`. We claim that `End_it` is endogenous, and we assume that `IV_it` is a valid instrument for it.

4.4.1 Pooled model

We first consider a pooled model in which we do not include any random or fixed effects. We first fit the model assuming that observations' errors are uncorrelated.[11]

```
. use https://acregstata.weebly.com/uploads/2/9/1/6/29167217/acregfakedata.dta,
> clear
. acreg Y_it X1_it (Z_it=IV_it)
HETEROSKEDASTICITY ROBUST STANDARD ERRORS
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 | Coefficient  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 then fit the same model accounting for correlation between errors from observations of the same individual (`id`). We still assume that there is no correlation between individuals and do not consider the network structure yet. To do this, we use the panel features (options `id()` and `time()`), and we set the `lagcutoff()` option to be equal to or greater than the maximum distance in time between observations, which in this case is 10.[12]

11. This is equivalent to using `ivreg2` (Baum, Schaffer, and Stillman 2003) and the syntax `ivreg2 Y_it X1_it (End_it=IV_it), robust`.
12. This is equivalent to clustering by individuals using `ivreg2` (Baum, Schaffer, and Stillman 2003) and the syntax `ivreg2 Y_it X1_it (End_it=IV_it) cluster(id)`, or `acreg`: `acreg Y_it X1_it (End_it=IV_it), cluster(id)`.

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(10)
TEMPORAL 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	Coefficient	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 further fit the model above adding to the temporal correlation the correction for network links, as proposed in our article (Colella et al. 2019). We assume that the error term of each individual is correlated with that of another individual observed in the same year if they are linked in the network while accounting for autocorrelation between errors from observations of the same individual over time. To implement this in `acreg`, we provide the variables containing the adjacency matrix as input in the `links_mat()` option and set `distcutoff(1)`.[13] The correlation between individuals that are linked but observed at different points in time is still assumed to be null.

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(10)
> network links_mat(clus*) distcutoff(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
```

Y_it	Coefficient	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

13. The total number of observations in the database is T (1,000), but the total number of individuals is N (100). Because we are using the panel feature, `acreg` will require a link matrix formed by N variables, not NT.

4.4.2 Fixed-effects 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()`.

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(10)
> network links_mat(clus*) distcutoff(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	Coefficient	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	1.10e-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
```

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

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(10)
> network links_mat(clus*) distcutoff(1) pfe1(id) pfe2(time)
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	Coefficient	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	-7.43e-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
```

4.4.3 Additional options

Thresholds. In the next example, we still account for network correlation between observations of the same year, but we do not account for any kind of temporal correlation. We do that by setting `lagcutoff(0)`.

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(0)
> network links_mat(clus*) distcutoff(1)
  (output omitted)
. estimates store nep1
```

Next we account for network correlation between observations of the same year, and also for temporal correlation between observations from the same individual, but only if they are observed with a difference of three years or less. We do that by setting `lagcutoff(3)`.

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(3)
> network links_mat(clus*) distcutoff(1)
  (output omitted)
. estimates store nep2
```

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

```
. acreg Y_it X1_it (Z_it=IV_it), id(id) time(time) lagcutoff(3)
> network links_mat(clus*) distcutoff(1) hac
  (output omitted)
. estimates store nep3
```

The following code reports the result of the three estimations in this subsection.

```
. esttab nep1 nep2 nep3, cells(b se) keep(X1_it Z_it) mtitles(lag0 lag10 hac)
```

	(1) lag0 b/se	(2) lag10 b/se	(3) hac b/se
Z_it	1.02863	1.02863	1.02863
	.3629168	.3783906	.3756538
X1_it	1.228864	1.228864	1.228864
	.4116362	.4578899	.4442984
N	1000	1000	1000

4.5 Multiway clustering

In this section, we illustrate the multiway clustering environment. `acreg` allows for the traditional one-dimension clustering, two-way clustering, and multiway cluster-

ing; in the last two scenarios, two observations are considered to be correlated if they share at least one cluster dimension (Cameron and Miller 2015). For this example, we again use the data on the homicides in southern states of the United States. `homicide_1960_1990.dta` is available at the Stata website. As in section 4.1, we consider only the cross-sectional database for 1990, and we estimate the effect of income on homicide rate, controlling for population and age. For the sake of illustration, we claim that income is endogenous, and we assume that unemployment is a valid instrument for it.

4.5.1 Two-way clustering

In this first example, we cluster standard errors following two dimensions: state and age.

```
. webuse homicide1990.dta, clear
(S.Messner et al.(2000), U.S southern county homicide rates in 1990)
. acreg hrate ln_population age (ln_income=unemployment), cluster(sfips age)
MULTIWAY CLUSTERING CORRECTION
Cluster variable(s): sfips age
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 141.918
                                                    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	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ln_income	-8.822082	1.818954	-4.85	0.000	-12.38717	-5.256998
ln_population	1.404433	.2960066	4.74	0.000	.8242705	1.984595
age	-.281615	.1315389	-2.14	0.032	-.5394265	-.0238035
_cons	94.4605	17.95901	5.26	0.000	59.26148	129.6595

4.5.2 Multiway clustering

The example above can also be replicated using the `ivreg2` command by simply typing `ivreg2 hrate ln_population age (ln_income=unemployment), cluster(sfips age)`. However, `ivreg2` accommodates a maximum of two cluster variables, while `acreg` allows for clustering any number of variables.[14] In the following and last example, we cluster standard errors following three dimensions: state, age, and homicide count.

14. The `cgmreg` command developed by Collin Cameron allows for multiway clustering but is not suitable for 2SLS estimation.

```
. acreg hrate ln_population age (ln_income=unemployment),
> cluster(sfips age hcount)
MULTIWAY CLUSTERING CORRECTION
Cluster variable(s): sfips age hcount
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic: 128.582
                                                    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	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ln_income	-8.822082	2.240027	-3.94	0.000	-13.21245	-4.431709
ln_population	1.404433	.7062929	1.99	0.047	.0201242	2.788741
age	-.281615	.1261689	-2.23	0.026	-.5289014	-.0343285
_cons	94.4605	21.90178	4.31	0.000	51.53379	137.3872

5 Conclusion

In this article, we presented the `acreg` command, a new community-contributed command that allows for standard error correction in OLS and 2SLS estimation of models with complex correlation structures. `acreg` can flexibly accommodate dependence of the errors between units in space or in a network and across time. This command includes most of the standard options present in previous commands to estimate regression coefficients. The correlation structure can be inputted by the user in a matrix form or built from information on the geographic distance between spatial units or from the links between observations. We also provided a broad collection of examples with both cross-section and panel data.

6 Programs and supplemental materials

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

```
. net sj 23-1
. net install st0703     (to install program files, if available)
. net get st0703         (to install ancillary files, if available)
```

Our statistical package (`acreg`) can be installed directly from the Statistical Software Components Archive by typing `ssc install acreg`. Complementary material may be found at the dedicated website: https://acregstata.weebly.com.

7 References

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3: 1–31. https://doi.org/10.1177/1536867X0300300101.

Bester, C. A., T. G. Conley, and C. B. Hansen. 2011. Inference with dependent data using cluster covariance estimators. *Journal of Econometrics* 165: 137–151. https://doi.org/10.1016/j.jeconom.2011.01.007.

Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2011. Robust inference with multiway clustering. *Journal of Business and Economic Statistics* 29: 238–249. https://doi.org/10.1198/jbes.2010.07136.

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

Colella, F., R. Lalive, S. O. Sakalli, and M. Thoenig. 2019. Inference with arbitrary clustering. IZA Discussion Paper No. 12584, Institute of Labor Economics (IZA). https://docs.iza.org/dp12584.pdf.

Conley, T. 1999. GMM estimation with cross sectional dependence. *Journal of Econometrics* 92: 1–45. https://doi.org/10.1016/S0304-4076(98)00084-0.

Correia, S. 2016. A feasible estimator for linear models with multi-way fixed effects. http://scorreia.com/research/hdfe.pdf.

Fetzer, T. 2015. Conley spatial HAC standard errors for models with fixed effects. http://www.trfetzer.com/conley-spatial-hac-errors-with-fixed-effects/.

Gelbach, J. B., and D. L. Miller. 2009. The community-contributed command cgmreg version 3.0.0. http://cameron.econ.ucdavis.edu/research/cgmreg.ado.

Grund, T. E. 2015. nwcommands: Software tools for statistical modeling of network data in Stata. http://nwcommands.org.

Grund, T. U., and J. A. Densley. 2012. Ethnic heterogeneity in the activity and structure of a Black street gang. *European Journal of Criminology* 9: 388–406. https://doi.org/10.1177/1477370812447738.

———. 2015. Ethnic homophily and triad closure: Mapping internal gang structure using exponential random graph models. *Journal of Contemporary Criminal Justice* 31: 354–370. https://doi.org/10.1177/1043986214553377.

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: 15367–15372. https://doi.org/10.1073/pnas.1009510107.

Jann, B. 2007. Making regression tables simplified. *Stata Journal* 7: 227–244. https://doi.org/10.1177/1536867X0700700207.

———. 2014. Software Updates: st0085_2: Making regression tables from stored estimates. *Stata Journal* 14: 451. https://doi.org/10.1177/1536867X1401400217.

Messner, S. F., L. Anselin, R. D. Baller, D. F. Hawkins, G. Deane, and S. E. Tolnay. 1999. The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. *Journal of Quantitative Criminology* 15: 432–450. https://doi.org/10.1023/A:1007544208712.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708. https://doi.org/10.2307/1913610.

Schaffer, M. E. 2005. xtivreg2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML, and k-class regression for panel-data models. Statistical Software Components S456501, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456501.html.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838. https://doi.org/10.2307/1912934.

About the authors

Fabrizio Colella is a senior research officer at CReAM, Department of Economics of the University College London, and an assistant professor of economics at USI Lugano.

Rafael Lalive is a professor of economics at the Faculty of Business and Economics of University of Lausanne.

Seyhun Orcan Sakalli is an assistant professor in economics at the King's Business School of the King's College London.

Mathias Thoenig is a professor of economics at the Faculty of Business and Economics of University of Lausanne.