# Incorporating V-Dem's Uncertainty Estimates in Regression Analysis[1]

Draft v.3

April 5, 2016

Fernando Bizzarro[2]

University of Notre Dame

Michael Coppedge

University of Notre Dame

Daniel Pemstein

North Dakota State University

[1] We thank Matthew Maguire, Gary Hollibaugh, Donald Brower, and Reid Boehm for their help at different stages of the preparation of this resource. We thank John Gerring, Matt Maguire, and their co-authors for the authorization to use their codes. Parts of the codes below replicate the "example analysis" by Melton, Meserve and Pemstein (2010), found at http://www.unified-democracy-scores.org/example.html.

[2] Authors ordered alphabetically.

# Contents

## Introduction

Uncertainty about data quality is part of a researcher's everyday morning thoughts. Whether the data one uses are true or not is a fundamental question of empirical research. We assume that no data point is "true", given that any tool to measure reality adds error to it, but we do believe that some data points are closer to the true than others are. Estimating how close to the true a given information is constitutes a necessary step in the validation of any empirical analysis.

V-Dem has taken estimating uncertainty seriously and has implemented many procedures to ensure that V-Dem Data is valid and reliable (see the Methodology documentation and Pemstein et al. 2016). Nonetheless, there is still uncertainty around V-Dem point estimates. It is impossible to tell without doubt that South Africa scored 1.31 and not a 1.30 on V-Dem's "Women's access to justice" indicator in 1985. Not only is this number an abstraction (it results from the transformation of coder scores into a single data point using V-Dem's measurement model); the original information (the scores provided by V-Dem coders themselves) is dependent on how confident coders were about their answers to a given question in a given year for a given country.

The degree of uncertainty about V-Dem data can be estimated, however. The outcome of the Measurement Model is not a single point estimate, but rather a distribution that provides information on where the "true" value is most likely to be. It takes into consideration the variation of the scores multiple coders give to the same country-year for a given question, and how these scores compare to similar scores given by the same coders, as well as by others, to different countries. The more coders diverged on the answer to one of V-Dem's questions, the less certain we are about the true score for that country-year. Conversely, the more agreement among coders there is, the more confident we are about our point-estimate.

Although estimating the uncertainty of our data is a good practice in itself, its main contribution to empirical analysis depends on our ability to incorporate uncertainty estimates in the tests and arguments we make about the topics researched. Following the lead of Melton, Meserve, and Pemstein (2010), in this tutorial we expose how researchers using V-Dem's data can incorporate the uncertainty estimates produced by the Measurement Model on regression analysis.

Paradoxically, we argue that by incorporating uncertainty estimates into one's analysis, researchers can actually feel more confident – more certain – about whether the empirical relationships they unveil with their research strategies are correct. This is because by adopting the method here suggested, researchers repeat their tests for many of the most likely values of their variable of interest. This reinforces the confidence one has on the empirical results of a given test, suggesting that it is robust to possible measurement error. Given that most regression analysis assume that there is no measurement error in the data, incorporating uncertainty estimates in the way we propose offer a powerful way of overcoming this very restrictive assumption.

In this document, we explain how you can download raw uncertainty estimates produced by the V-Dem Measurement Model, how you can transform them into a workable set of information, and how do you incorporate them in basic regression analysis. We offer one example for each technique. The strategies explored here, however, are robust and can be replicated on a variety

of different variables, indices, and estimation procedures. Although all examples use Stata codes, R codes for the same analyses are available.

## Data Types

The V-Dem Dataset contains four types of variables. "A" variables provide relatively objective information about countries and years and were coded by Project Managers and Research Assistants. "B" variables are similar to "A" variables but were coded by Country Coordinators who had better access to local information. "C" variables were coded by country-experts. "D" variables are Indices created by the aggregation of two or more A, B, or C variables. Each of those variables can be incorporated in the analysis when researchers try to account for measurement error, but there are some differences among them worth noticing:

> *A and B variables*: those variables are – we assume – free from measurement error. Strictly speaking, their measurement error is unknown because each score came from just one coder. However, as they are relatively objective variables, they are less subject to measurement error.
>
> *C variables*: are by definition not free from measurement error and most of the work done in the V-Dem Project to estimate error was done for these variables. There are two types of C variables with two different error estimation procedures:
>
> > *Measurement Model Variables*: are created using the V-Dem's Bayesian IRT Measurement Model that aggregates coder scores and translates them from categorical values into a single continuous scale for each variable.
> >
> > *Bootstrap Variables*: are created from the bootstrapping of coder scores. Most of these are variables that asked coders to provide continuous rather than categorical answers to questions, such as percentages.
>
> *D variables:* the uncertainty estimates of these indices are derived from the uncertainties of the indicators comprising them.

Available for download by users are the uncertainty estimates for two of those five groups of variables. Users can download information about the Measurement Model Variables and about the Indices (D variables). Uncertainties for Bootstrap variables need to be estimated, while A and B variables have no uncertainty estimates.

## Downloading the Data

V-Dem's uncertainty estimates ("posteriors") are stored in the CurateND Archive, a data repository hosted by the University of Notre Dame's Hesburgh Libraries system. To access it, users can click on the link to the V-Dem posteriors in the "Data" section of the V-Dem website, or access it directly by searching for it on CurateND at https://curate.nd.edu/.

The V-Dem posteriors collection holds a separate file for each Measurement Model C Variable. These files are stored in compressed files (.zip) that group five matrices produced by the V-Dem Measurement Model. The matrices included in each file are described on Table 1.

Table 1. Description of the matrices created by the V-Dem Measurement Model

| Matrix | Description |
|---|---|
| Z | Uncertainty Estimates. Posteriors |
| B | Rater discrimination parameters |
| Gamma | Rater thresholds |
| Gamma c | Hierarchical parameter. Average threshold of raters within countries |
| Gamma u | Sample averages |

After downloading the .zip file for the variable of interest, users can decompress the file ("unzip"), which will create five new .csv (comma-separated) files, one for each of the matrices produced after the run of the Measurement Model. The "normal" matrix is the one we use for incorporating measurement error estimates in the analysis. You probably will not need the other matrices; we archive them because some measurement specialists will find them useful.

Users should download and decompress files for all the variables they will use.

Files are stored at CurateND by survey. If your variable is a "v2cs*" variable, i.e., a variable regarding the civil society, you can find it at CurateND by clicking on the link to the dataset containing the Civil Society variables. While most V-Dem datasets at CurateND include two surveys, one includes only Executive questions, and one includes only Civil Liberties variables.

For D Variables (Indices) CurateND stores a single matrix for each index. It contains the posterior distribution estimated by the Bayesian Factor Analysis method, which was used to create all higher-level indices in the V-Dem Dataset.

## Example 1. Uncertainty estimates for Measurement Model C variables

In the first example, we use just one Measurement Model C Variable as an independent variable in a regression analysis. We believe this is the most common setting in which users will take advantage of these tools. If the user has the V-Dem variable as a Dependent variable, the procedure does not change at all. We test whether freedom of discussion for women affects infant mortality rates. Freedom of discussion for women is one of the Measurement Model C Variables in the V-Dem dataset. Data for infant mortality rates come from Gapminder, with additional data imputed from Clio-Infra (for additional information, see the V-Dem Codebook).

Before we estimate the relationship between the two variables using the posterior files, we can do the same using the point-estimates available in the V-Dem Country-Year Dataset. We run a model with country fixed effects and standard errors clustered by countries. We add one control variable: GDP per capita (logged). All independent variables are lagged 1 year.

```
* DO-FILE: "Example1.do"

. *** Single C-Variables ***
.
. ** Mean Value analysis (Baseline) **
.
. * Loading the data
. clear

. use "C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta"
(V-Dem Country-Year Dataset v5. Team)

.
. * Setting up the dataset
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

.
. * Running the analysis with mean value
. xtreg F.e_peinfmor v2cldiscw e_migdppcln , fe vce(cluster country_id)

Fixed-effects (within) regression               Number of obs     =      9,358
Group variable: country_id                      Number of groups  =        150

R-sq:                                           Obs per group:
     within  = 0.5028                                         min =         13
     between = 0.6492                                         avg =       62.4
     overall = 0.5214                                         max =        111

                                                F(2,149)          =     184.78
corr(u_i, Xb)  = -0.5625                        Prob > F          =     0.0000

                          (Std. Err. adjusted for 150 clusters in country_id)
------------------------------------------------------------------------------
             |               Robust
F.e_peinfmor |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   v2cldiscw |  -6.081386   1.543519    -3.94   0.000     -9.1314   -3.031373
 e_migdppcln |  -49.18635    4.16452   -11.81   0.000   -57.41549    -40.9572
       _cons |   464.9086   32.57084    14.27   0.000    400.5482     529.269
-------------+----------------------------------------------------------------
     sigma_u |  28.995657
     sigma_e |  28.933936
         rho |  .50106544   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Freedom of discussion for women has a negative (-6.08) and significant (t = -3.94) effect on infant mortality rates after controlling for GDP per capita (logged) levels. Keep those numbers in mind because you will compare them to the results we obtain after we incorporate measures of uncertainty in the analysis.

Uncertainty estimates for this variable can be downloaded from the CurateND website. Users should find the link for the "Civil Liberties" variables and then download v2cldiscw from the list of variables inside this work. After download, users can "unzip" the file. We will use the v2cldiscw.10000.z.sample.csv file. After downloading and unzipping the posterior file, the first step is to import the CSV file into Stata.

```
* DO-FILE: "Example1.do"

. *** Incorporating Measurement Error ***
.
. *** Generating Posterior File in Stata ***
.
. *** Setting up Stata ***
. *Clears Data and Existing Matrices
. clear

. matrix drop _all

.
. *Changes Working Directory
. cd "C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs" /*INSERT
THE DIRECTORY W
> HERE YOU EXTRACTED THE DATA FILES*/
C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs

.
. *** Loading and preparing the data ***
. * load posteriors file
. import delimited "v2cldiscw.10000.z.sample.csv", clear
(901 vars, 2,117 obs)
```

After downloading and importing the matrix into Stata, the user may realize that the matrix is incomplete: not all country-years are represented in the matrix. This happens because the V-Dem Measurement Model estimates scores only for the first year of a sequence of years in which no coder changed their answers, or their confidence on their answer, for that country in that year (what we call "regimes"). In order to have a complete matrix, we need to "carry forward" the values for the first year of a "regime" to all the subsequent years.

```
* DO-FILE: "Example1.do"

. * generate year and country_text_id
. gen country_text_id = substr(v1,1,3)

. gen year = substr(v1,5,4)

. destring year, replace
year has all characters numeric; replaced as int

.
. * when there are multiple observations for a particular country-year, keep
the oldest observation (i.e. toward the end of the year rather than the
beginning)
. gen obs_sort =_n

. gsort -obs_sort

. duplicates drop country_text_id year, force

Duplicates in terms of country_text_id year

(19 observations deleted)

. sort obs_sort

. drop v1 obs_sort

.
. * add country_id and the point-estimate from V-Dem Country-Year dataset
. merge 1:1 country_text_id year using ///
"C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta", ///
keepusing(country_id v2cldiscw) nogenerate
(note: variable country_text_id was str3, now str6 to accommodate using
data's values)
(note: variable year was int, now float to accommodate using data's values)

    Result                           # of obs.
    -----------------------------------------
    not matched                        21,470
        from master                         0
        from using                     21,470

    matched                             2,098
    -----------------------------------------


.
. * Rename the mean point-estimate
. rename v2cldiscw cldiscw
```

```
. * carry forward and rename variables
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

. foreach var of varlist v2 - v901{
  2. qui bysort country_id: carryforward `var', replace
  3. qui replace `var' = . if cldiscw== .
  4. qui rename `var' cldiscw_`var'
  5. }

.
. * reorder the new dataset
. order country_text_id country_id year cldiscw

.
. * Drop observations for years in which values for the variable of interest
(with posteriors) is missing
. drop if cldiscw == .
(7,077 observations deleted)

.
. * Save the new file
. save
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\cldiscw.dta",
replace
```

Now that a dataset with all Country-years was created and filled, we merge this dataset with the original V-Dem dataset in order to include the other variables we will need for the analysis. By setting “xtset country_id year” we also tell Stata that we have panel data.

```
* DO-FILE: "Example1.do"

. *** Analysis Section ***
.
. *** Analysis Preparation ***
. *Merge with V-Dem Country-Year data
. merge m:m country_text_id year using
"C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta", nogenerate

    Result                           # of obs.
    -----------------------------------------
    not matched                         7,077
        from master                         0
        from using                      7,077

    matched                            16,491
    -----------------------------------------


.
. *Sets Data
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014, but with gaps
                delta:  1 unit

.
. * Set Matrix Size to fit analysis
. set matsize 5000
```

With the dataset ready for the analysis, we run the same regression we had for the baseline model above 900 times, i.e., the numbers of draws of the posterior distribution stored from the V-Dem Measurement Model. The user will realize that the code replaces "v2cldiscw" by each of the "v###" variables now in the dataset set sequentially. The code below follows a Monte Carlo Markov Chain strategy to run the multiple regressions and then group results. Please note that running the MCMC procedure may take some minutes.

```
* DO-FILE: “Example1.do”

. *** Analysis ***
. *** Monte Carlo Estimates Using V-Dem 900 Draw Posterior Distribution***
. *Run the monte carlo
.
. forvalues i = 2/901 {
  2.
. *Print out an iteration number
. display `i'
  3.
. *Fit the model, using the ith draw from the UDS posterior
. quietly xtreg F.e_peinfmor cldiscw_v`i' e_migdppcln, fe vce(cluster
country_id)
  4.
. *Extract the coefficients and variance-covariance matrix
. matrix b = e(b)
  5. matrix V = e(V)
  6. local blength = colsof(b)
  7. matrix rsq = e(r2)
  8.
. *Preserve the dataset, take a single multivariate normal draw from the
. *posterior distribution of the coefficients, and restore the dataset.
. *We use the capture command to catch possible errors in drawnorm
. *and drop these iterations gracefully.
. preserve
  9. capture quietly drawnorm b1-b`blength', double n(1) means(b) cov(V)
clear
 10. if _rc==0 {
 11. mkmat b1-b`blength', matrix(bsample)
 12. matrix posterior = nullmat(posterior) \ bsample
 13. matrix rsquared = nullmat(rsquared) \ rsq
 14. }
 15. else {
 16. display "Error drawing sample...iteration dropped"
 17. }
 18. restore
 19.
. *Closes the Monte Carlo Loop}
}
2
3
4
5
.
.
.
901
```

After the Monte Carlo procedure is finished, Stata will have stored the results of the 900 regressions we ran in the background. The code below retrieves the results. The first table has the mean and standard deviations of all the coefficients for the variables included in the regression command. E.g., in the regression we ran in this tutorial, the first coefficient is the v2cldiscw coefficient, the second is the e_migdppcln coefficient, and the third value is the

constant. In the first post-estimation table you see now, the values for "posterior1" are the mean coefficient and standard error of the coefficients for v2cldiscw, and similarly for the other variables. The second table reports similar results, but rather than mean coefficients and standard errors, it reports the upper and lower bounds of the distribution of the estimated coefficients ("centile" column), and the 95% confidence intervals for those bounds (estimated by Stata using a binomial-based method (See the Stata Manual for more information .

```
* DO-FILE: "Example1.do"


. *Calculate means and standard deviations
. tabstat posterior*, stat(mean sd)

   stats |  poster~1  poster~2  poster~3
---------+------------------------------
    mean | -4.927926 -50.19305  472.5253
      sd |  1.310335  3.901041  30.56771
----------------------------------------

.
. *Find the bounds of the 95 percent credible interval
. centile posterior*, centile(2.5, 97.5)

                                                      -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+------------------------------------------------------------
  posterior1 |       900        2.5   -7.255371       -7.450067   -7.129636
             |                 97.5   -2.241205       -2.533353   -1.950193
  posterior2 |       900        2.5   -57.91597       -59.31109   -57.27598
             |                 97.5   -42.60861       -43.32256   -41.80473
  posterior3 |       900        2.5     413.051         406.826     418.295
             |                 97.5     532.892        528.0778     544.035

.
. * Find the R-Squared
. tabstat rsquared*, stat(mean sd)

    variable |      mean        sd
-------------+--------------------
   rsquared1 |  .4995164  .0029321
----------------------------------
```

We can now compare the coefficient for the baseline model (-6.08) to the mean of the coefficients produced in the MCMC run of 900 regressions (-4.92). In the second case, as you notice, the coefficient is smaller but still significantly different from zero (t=-3.76). As expected, both the coefficient for e_migdppcln and the intercept change as well, given that the scores in the last analysis are the mean of their values across the 900 regressions. The second table reported by those codes brings the 2.5 and 97.5 percentile of the distribution of the coefficients, and the Stata function for extracting them estimates Confidence intervals for those values (reported in the right hand side of the equation). The last table provides the average overall R-

square and the standard deviation of the R-squares. If you need any other statistic from the model, you can update the code for the MCMC sequence similarly to the parts of it referring to the R-Square to store and then estimate values after the regressions.

## Example 2. Uncertainty estimates for D variables (Indices)

The procedure to incorporate the estimated uncertainty of V-Dem indices is identical to the previous procedure designed for C Measurement Model variables. The main differences are in the input file. Rather than a z.matrix, indices are named after a smaller version of the variable name. Posterior distributions for all indices can be downloaded from the CurateND collection, under the work labelled "Indices".

In this example, we run a similar analysis to the one we had in Example 1, but replace Freedom of discussion for Women with V-Dem's Clean Elections Index. Our expectations are similar in this model: countries with elections that are cleaner should have lower levels of infant mortality for many different reasons, but particularly because of accountability mechanisms that are part of electoral democratic settings. Again, before including the uncertainty estimates, we run the baseline model with the point-estimates in the V-Dem Dataset.

```
* DO-FILE: "Example2.do"

. *** Single C-Variables ***
.
. ** Mean Value analysis (Baseline) **
.
. * Loading the data
. clear

. use "C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta"
(V-Dem Country-Year Dataset v5. Team)

.
. * Setting up the dataset
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

.
. * Running the analysis with mean value
. xtreg F.e_peinfmor v2xel_frefair e_migdppcln , fe vce(cluster country_id)

Fixed-effects (within) regression               Number of obs     =      9,522
Group variable: country_id                      Number of groups  =        154

R-sq:                                           Obs per group:
     within  = 0.4891                                         min =         13
     between = 0.6410                                         avg =       61.8
     overall = 0.5067                                         max =        111

                                                F(2,153)          =     153.65
corr(u_i, Xb)  = -0.5922                        Prob > F          =     0.0000

                           Std. Err. adjusted for 154 clusters in country_id)
------------------------------------------------------------------------------
             |               Robust
 F.e_peinfmor |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
--------------+---------------------------------------------------------------
v2xel_frefair |  -26.28558   4.990621    -5.27   0.000  -36.14501   -16.42616
  e_migdppcln |  -49.86238   3.466855   -14.38   0.000  -56.71146   -43.01329
        _cons |   480.2628   26.87065    17.87   0.000   427.1774    533.3482
--------------+---------------------------------------------------------------
      sigma_u |  29.857156
      sigma_e |  29.258789
          rho |  .51012089   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Clean Elections have a negative (-26.28) and significant (t=-5.27) association with the levels of infant mortalilty, as we expected. The procedure to incorporate estimates of measurement error is identical to the previous one. First, we load the matrix with the estimated distributions for each country-year value. Then we expand the reduced matrix to cover all country-year combinations by using the carry forward command, merge this dataset to the V-Dem Dataset,

and run the Monte Carlo Markov Chain procedure to estimate and combine results of the analysis incorporating the uncertainty estimates.

One important detail in the next code is the "replace `var' = normal (`var')" command. V-Dem indices are first estimated as factor scores from a Bayesian Factor Analysis. Factor scores' range is unbounded, theoretically varying from minus to plus infinity. However, V-Dem Indices are converted to vary from 0 to 1 in relation to the Cumulative Distribution Function of the Factor Scores. In this sense, every score can be read as the percentage of cases that has a value smaller than the score for that country-year unit.

```
* DO-FILE: "Example2.do"

. clear
. matrix drop _all

.
. *Changes Working Directory
. cd "C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs" /*INSERT
THE DIRECTORY WHERE YOU EXTRACTED THE DATA FILES*/
C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs

.
. *** Loading and preparing the data ***
. * load posteriors file
. import delimited "frefair.csv", clear
(901 vars, 4,613 obs)

.
. * generate year and country_text_id
. gen country_text_id = substr(v1,1,3)

. gen year = substr(v1,5,4)

. destring year, replace
year has all characters numeric; replaced as int

.
. * when there are multiple observations for a particular country-year, keep
the oldest observation (i.e. toward the end of the year rather than the
beginning)
. gen obs_sort =_n

. gsort -obs_sort

. duplicates drop country_text_id year, force

Duplicates in terms of country_text_id year

(798 observations deleted)

. sort obs_sort

. drop v1 obs_sort
```

```
. * add country_id and the point-estimate from V-Dem Country-Year dataset
. merge 1:1 country_text_id year using
"C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta", ///
keepusing(country_id v2xel_frefair) nogenerate
(note: variable country_text_id was str3, now str6 to accommodate using
data's values)
(note: variable year was int, now float to accommodate using data's values)

    Result                           # of obs.
    -----------------------------------------
    not matched                        19,753
        from master                         0
        from using                     19,753

    matched                             3,815
    -----------------------------------------


.
. * Rename the V-Dem point-estimate
. rename v2xel_frefair frefair

.
. * carry forward and rename variables
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

. foreach var of varlist v2 - v901{
  2. qui bysort country_id: carryforward `var', replace
  3. qui replace `var' = . if frefair== .
  4. replace `var' = normal(`var') /* V-Dem Indices converted to 0 - 1 */
  5. qui rename `var' frefair_`var'
  6. }
.
. * reorder the new dataset
. order country_text_id country_id year frefair}

. * Drop observations for years in which values for the variable of interest
(with posteriors) is missing
. drop if frefair == .
(5,907 observations deleted)

.
. * Save the new file
. save
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\frefair.dta",
replace
file C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\frefair.dta
saved
```

We can now merge the V-Dem Dataset to this new matrix in order to add the other variables we will use in the analysis. Then, we run the Monte Carlo Markov Chain procedure to estimate the new set of regressions. Note that from the previous example, the only change we made was updating the regression command to change the name of the variables we are using.

```
* DO-FILE: "Example2.do"

. *** Analysis ***
. *** Monte Carlo Estimates Using V-Dem 900 Draw Posterior Distribution***
. *Run the monte carlo
.
. forvalues i = 2/901 {
  2.
. *Print out an iteration number
. display `i'
  3.
. *Fit the model, using the ith draw from the UDS posterior
. quietly xtreg F.e_peinfmor frefair_v`i' e_migdppcln, fe vce(cluster
country_id)
  4.
. *Extract the coefficients and variance-covariance matrix
. matrix b = e(b)
  5. matrix V = e(V)
  6. local blength = colsof(b)
  7. matrix rsq = e(r2)
  8.
. *Preserve the dataset, take a single multivariate normal draw from the
. *posterior distribution of the coefficients, and restore the dataset.
. *We use the capture command to catch possible errors in drawnorm
. *and drop these iterations gracefully.
. preserve
  9. capture quietly drawnorm b1-b`blength', double n(1) means(b) cov(V)
clear
 10. if _rc==0 {
 11. mkmat b1-b`blength', matrix(bsample)
 12. matrix posterior = nullmat(posterior) \ bsample
 13. matrix rsquared = nullmat(rsquared) \ rsq
 14. }
 15. else {
 16. display "Error drawing sample...iteration dropped"
 17. }
 18. restore
 19.
. *Closes the Monte Carlo Loop
. }
2
3
4
.
.
.
901
```

Results are collected after the MCMC procedure is completed, with each "posterior" in the new tables corresponding to one of the variables in the baseline model. In this case, "posterior1" stands for v2xel_frefair.

```
* DO-FILE: "Example2.do"


. *Get posterior ready to work with
. svmat posterior

. svmat rsquared

.
. *Calculate means and standard deviations
. tabstat posterior*, stat(mean sd)

   stats |  poster~1  poster~2  poster~3
---------+------------------------------
    mean | -26.76279 -50.18824  483.5814
      sd |  8.715293  3.505868   27.7675
----------------------------------------

.
. *Find the bounds of the 95 percent credible interval
. centile posterior*, centile(2.5, 97.5)

                                                      -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
  posterior1 |       900        2.5   -43.28243       -45.10047    -42.0427
             |                 97.5   -9.423181       -11.19669   -7.890188
  posterior2 |       900        2.5   -57.07853       -57.74767   -56.42854
             |                 97.5   -43.80626       -44.11761   -42.82604
  posterior3 |       900        2.5    431.2346        423.7401    434.6084
             |                 97.5    537.4703        533.1596    542.1565

.
. * Find the R-Squared
. tabstat rsquared*, stat(mean sd)

    variable |      mean        sd
-------------+--------------------
   rsquared1 |  .4971693   .002986
-----------------------------------
```

Estimates for the baseline model (-26.2) and for the model incorporating estimates of measurement error (-26.7) are very similar, and both are statistically different from 0 (t = -5.27, and t = -3.06). Remember that V-Dem indices incorporate two types of uncertainties: those coming from the individual indicators and the uncertainty produced by the aggregation method itself. Similar results like those are reassuring: they confirm the strength of the relationship uncovered.

## Example 3. Two V-Dem variables

The previous two examples include only one V-Dem variable. In the first case, we have a C variable with uncertainty estimates generated by the V-Dem Bayesian IRT Measurement Model, and in the second example, we have an index with uncertainty estimates being the posterior distribution of the Bayesian Factor analysis used for aggregation. In this third example, we show how to use multiple V-Dem variables in the same analysis. In order to make things simpler, we run a model similar to the previous ones (DV = Infant Mortality Rates), and use both Freedom of Discussion for Women and the Clean Elections Index as predictors. We also control for GDP per capita (logged).

The procedure to include multiple variables is identical to the previous one. The user must pay attention though to two crucial steps: 1. Merging all the correct posteriors in the using dataset; 2. Updating the regression command in order to ensure that Stata (or another statistical program) runs the regression with the right variables as many times as needed.

This procedure is robust to including more than two variables or including variables in both the right and left-hand sides of the equation. All you need to do is to include the "var_v`l'" (var being your variable of interest) command in the right place of the regression equation.

Before we delve into the procedure to incorporate uncertainty estimates, we run the baseline model using the point-estimates in the V-Dem Dataset.

```
* DO-FILE: "Example3.do"

. *** Two Variables Example (1 C-var, 1 index) ***
.
. ** Mean Value analysis (Baseline) **
.
. * Loading the data
. clear

. use "C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta"
(V-Dem Country-Year Dataset v5. Team)

.
. * Setting up the dataset
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

.
. * Running the analysis with mean value
. xtreg F.e_peinfmor v2xel_frefair v2cldiscw e_migdppcln , fe vce(cluster
country_id)

Fixed-effects (within) regression               Number of obs     =      9,282
Group variable: country_id                      Number of groups  =        150

R-sq:                                           Obs per group:
     within  = 0.5003                                         min =         13
     between = 0.6479                                         avg =       61.9
     overall = 0.5163                                         max =        111
                                                F(3,149)          =     116.94
corr(u_i, Xb)  = -0.5818                        Prob > F          =     0.0000

                           (Std. Err. adjusted for 150 clusters in
country_id)
------------------------------------------------------------------------------
             |               Robust
 F.e_peinfmor |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
--------------+---------------------------------------------------------------
--
v2xel_frefair |  -13.86806   6.017958    -2.30   0.023  -25.75962   -1.976492
    v2cldiscw |  -4.180604   1.826811    -2.29   0.024  -7.790405    -.570802
  e_migdppcln |  -48.43413   4.097581   -11.82   0.000  -56.53101   -40.33726
        _cons |   464.5324   32.77322    14.17   0.000   399.7721    529.2927
--------------+---------------------------------------------------------------
      sigma_u |  29.436939
      sigma_e |  28.802839
          rho |  .51088646   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Both the Clean Elections index (v2xel_frefair) and Freedom of Discussion for women (v2cldiscw) have negative coefficients (-13.8 and -4.1, respectively), and statistically distinct from zero (t = -

2.3 and t = -2.29, respectively), suggesting that granting civil rights to women might affects infant mortality rates independently from how clean elections are, and that cleaner elections, in themselves, tend also to predict lower rates of infant mortality.

To save space, we do not show how to extract and generate .dta files for those two variables again. Users can review the initial procedures included in the previous two examples to estimate that. In this code, we load the already saved posterior files, merge them, and prepare the data for the analysis.

```
* DO-FILE: "Example3.do"

. *** Incorporating Measurement Error ***
.
. *** Setting up Stata ***
. *Clears Data and Existing Matrices
. clear

. matrix drop _all

.
. *Changes Working Directory
. cd "C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs" /*INSERT
THE DIRECTORY WHERE YOU EXTRACTED THE DATA FILES*/
C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs

.
. * Loading posteriors dta
. * pubcorr
. use "frefair.dta", clear

.
. * Merge the second posterior
. merge m:m country_text_id country_id year using
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\cldiscw.dta",
nogenerate

    Result                           # of obs.
    -----------------------------------------
    not matched                         1,598
        from master                     1,384
        from using                        214

    matched                            16,277
    -----------------------------------------
```

```
. * Merge with V-Dem Country-Year data
. merge m:m country_text_id year using
"C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta", nogenerate

    Result                           # of obs.
    -----------------------------------------
    not matched                         5,693
        from master                         0
        from using                      5,693

    matched                            17,875
    -----------------------------------------

.
. *Sets Data
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

.
. * Set Matrix Size to fit analysis
. set matsize 50000
```

After the matrices are loaded and merged, we can run the MCMC procedure one more time. Note that the code loops through the regression 900 times and that at each time, the $i^{th}$ draw of both frefair and cldiscw are included in the analysis.

```
* DO-FILE: “Example3.do”

. *** Analysis ***
. *** Monte Carlo Estimates Using V-Dem 900 Draw Posterior Distribution***
. *Run the monte carlo
.
. forvalues i = 2/901 {
  2.
. *Print out an iteration number
. display `i'
  3.
. *Fit the model, using the ith draw from the UDS posterior
. quietly xtreg F.e_peinfmor frefair_v`i' cldiscw_v`i' e_migdppcln, fe
vce(cluster country_id)
  4.
. *Extract the coefficients and variance-covariance matrix
. matrix b = e(b)
  5. matrix V = e(V)
  6. local blength = colsof(b)
  7. matrix rsq = e(r2)
  8.
. *Preserve the dataset, take a single multivariate normal draw from the
. *posterior distribution of the coefficients, and restore the dataset.
. *We use the capture command to catch possible errors in drawnorm
. *and drop these iterations gracefully.
. preserve
  9. capture quietly drawnorm b1-b`blength', double n(1) means(b) cov(V)
clear
 10. if _rc==0 {
 11. mkmat b1-b`blength', matrix(bsample)
 12. matrix posterior = nullmat(posterior) \ bsample
 13. matrix rsquared = nullmat(rsquared) \ rsq
 14. }
 15. else {
 16. display "Error drawing sample...iteration dropped"
 17. }
 18. restore
 19.
. *Closes the Monte Carlo Loop
}
2
3
4
.
.
.
901
```

Using the same matrix operations from the previous examples, we can extract the mean of the 900 estimated coefficients for each variable included in the analysis, as well as the upper and lower bounds of those estimates. In this example, posterior1 refers to the estimates for Clean Elections (frefair), and posterior2 refer to the estimates for Freedom of Discussion for Women (v2cldiscw).

```
* DO-FILE: "Example3.do"

. *Get posterior ready to work with
. svmat posterior

. svmat rsquared

.
. *Calculate means and standard deviations
. tabstat posterior*, stat(mean sd)

   stats |  poster~1  poster~2  poster~3  poster~4
---------+----------------------------------------
    mean | -4.870637 -3.222026 -47.92444  453.4457
      sd |  3.039264  1.620669  3.696745  29.09013
------------------------------------------------

.
. *Find the bounds of the 95 percent credible interval
. centile posterior*, centile(2.5, 97.5)

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
  posterior1 |       900        2.5   -10.93175       -11.98236    -10.3376
             |                 97.5    1.116468        .6094904    1.803772
  posterior2 |       900        2.5   -6.345521       -6.688132   -6.078082
             |                 97.5    -.034826       -.3028529    .3201572
  posterior3 |       900        2.5    -55.6458       -56.76825   -55.04263
             |                 97.5   -40.92642       -41.52695   -39.73515
  posterior4 |       900        2.5    398.4656         388.459    403.2658
             |                 97.5     514.255        509.6558    522.7474

.
. * Find the R-Squared
. tabstat rsquared*, stat(mean sd)

    variable |      mean        sd
-------------+--------------------
   rsquared1 |  .5038968  .0030608
---------------------------------
```

The two variables of interest change the size of their coefficient when we include estimates of measurement error: the mean of the coefficients for Clean Elections is much larger (-4.8 vs -13.1), although still statistically distinct from zero, and Freedom of Discussion for Women (posterior2) is somewhat larger (-3.2 vs -4.1) after we incorporate measurement error estimates in our analysis. Note that despite those changes, there is not a statistically significant difference between the coefficients in the baseline model and in the posterior models, given that baseline estimates fall before the Confidence Intervals of the Posteriors.

## Example 4. Aggregating V-Dem Indicators' uncertainty measures

One of the advantages of the V-Dem Dataset is that it provides fine-grained indicators of many dimensions, practices, and institutions of political regimes that can be combined into higher-level indices to represent concepts that are more complex. This is the logic behind the V-Dem indices, for example. However, scholars may want to combine variables in ways not yet anticipated by V-Dem and create their own indices. Using aggregated indices need not to prevent scholars from incorporating uncertainty measures in their analyses. In this section, we explain how aggregation strategies can be replicated using the posterior distributions in order to provide a good estimation of measurement error for aggregate indices as well.

The principle behind this aggregation is simple. Everything you would do to aggregate the point estimates in the main V-Dem Dataset can also be done to each of the "i" draws of the posterior distribution to generate aggregate estimates of measurement uncertainty. In other words, if for the baseline model the researcher builds an additive index of four V-Dem indicators, the index's uncertainty estimates are a matrix of 900 variables, in which each "v" variable is the sum of the "v" columns of the four matrices of the four indicators.

In order to demonstrate this procedure, we adapt the test done by Gerring et al. (2016) in which they explore the impacts of democracy on human development. Their main dependent variable is infant mortality rates (logged) as in our previous examples, while the independent variable is a "Multiplicative Electoral Democracy Index (MEDI)" described as follows:

> "Our chosen index draws on indicators that tap into the institutional procedures emphasized by Dahl (1989) in connection with the concept of polyarchy. Specifically, it is intended to measure the extent of responsiveness and accountability between leaders and citizens through the mechanism of competitive elections. This is presumed to be maximized when (a) elections are clean and not marred by fraud or systematic irregularities, (b) the chief executive of a country is selected (directly or indirectly) through elections, (c) suffrage is extensive, (d) political and civil society organizations operate freely, and (e) there is freedom of expression, including access to alternative information." (Gerring et al., 2016, 16).

In accordance with this reasoning, the index combines, by multiplication, five V-Dem indices from two different sources: three indices (Free and Fair Elections [v2xel_frefair], Freedom of association [v2x_frassoc_thick], and Freedom of Expression [v2x_freeexp_thick]), which were created from the aggregation of C variables; and two indices created from A and B variables (Elected Executive Index (de jure) [v2x_accex] and Share of the population with suffrage [v2x_suffr]). While the first three indices are created using a Bayesian Factor Analysis model – for which, therefore, we have estimated posterior distributions – the last two have no measurement error estimates associated with them.

Before we start constructing the uncertainty matrix for their index, we can run their baseline model.

```
* DO-FILE: "Example4.do"

. *** Combined Index ***
.
. ** Mean Value analysis (Baseline) **
.
. * Loading the data
. clear

. use "C:\Users\fbizz\Dropbox\Dissertation\Data\vdem_cy_v5.dta"
(V-Dem Country-Year Dataset v5. Team)

.
. * Setting up the dataset
. xtset country_id year
       panel variable:  country_id (unbalanced)
        time variable:  year, 1900 to 2014
                delta:  1 unit

.
. * Generating Multiplicative Index
. qui gen edi_mult = v2xel_frefair * v2x_frassoc_thick * v2x_freexp_thick *
v2x_accex * v2x_suffr

.
. * Baseline Model
. xtreg F.e_peinfmor edi_mult e_migdppcln, fe vce(cluster country_id)

Fixed-effects (within) regression               Number of obs     =      9,112
Group variable: country_id                      Number of groups  =        150

R-sq:                                           Obs per group:
     within  = 0.5084                                         min =         13
     between = 0.6211                                         avg =       60.7
     overall = 0.5222                                         max =        111

                                                F(2,149)          =     158.42
corr(u_i, Xb)  = -0.5385                        Prob > F          =     0.0000

                          (Std. Err. adjusted for 150 clusters in country_id)
------------------------------------------------------------------------------
             |               Robust
F.e_peinfmor |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-
    edi_mult |  -46.51259   7.903093    -5.89   0.000    -62.12921   -30.89597
 e_migdppcln |  -43.32959   4.279629   -10.12   0.000    -51.78619   -34.87299
       _cons |   427.9336   32.67252    13.10   0.000     363.3722    492.4949
-------------+----------------------------------------------------------------
     sigma_u |   29.32406
     sigma_e |  28.244513
         rho |  .51874577   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The coefficient for the MEDI in this first test is negative (-46.5) and statistically significant (t=--5.89), in line with the authors' general finding that their multiplicative index of democracy predicts lower levels of infant mortality.

In order to replicate this analysis and incorporate measurement error, all the steps of this first analysis need to be replicated for the posterior distributions of the indices. First, we load and prepare the posteriors matrices. Please notice that this code loops through multiple files in a same folder to automatically generate .dta files with the posterior distributions for all z.matrices stored there. So in order to reproduce it, you can download all files in one same folder (and make sure they are the only .csv files in this folder) and run this model to generate all matrices .dta at once).

```
* DO-FILE: “Example4.do”

. *** Incorporating Measurement Error ***
. * Make sure you downloaded all the posteriors in the same folder
. clear

. local files : dir
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\ind_posteriors"
files "*.csv"

. cd
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\ind_posteriors"

. foreach file in `files' {
  2.         import delimited using `file', clear
  3.
.         * generate year and country_text_id
.         gen country_text_id = substr(v1,1,3)
  4.         gen year = substr(v1,5,4)
  5.         destring year, replace
  6.
.         * when there are multiple observations for a particular country-
year, keep the oldest observation (i.e. toward the end of the year rather
than the beginning)
.         gen obs_sort =_n
  7.         gsort -obs_sort
  8.         duplicates drop country_text_id year, force
  9.         sort obs_sort
 10.         drop v1 obs_sort
 11.
.     local varname = substr("`file'",1,length("`file'")-4)
 12.         if "`varname'" == "frefair" local varname2 = "v2xel_`varname'"
/* change this if you have other variables */
 13.         else if "`varname'" == "frefair" local varname2 =
"v2x_`varname'"
 14.
.         * add country_id
.         merge 1:1 country_text_id year using
"C:/Users/fbizz/Dropbox/Dissertation/Data/vdem_cy_v5.dta",
keepusing(country_id `varname2')
 15.         drop _merge
 16.
.         * carry forward and rename variables
.         xtset country_id year
 17.         foreach var of varlist v2 - v901{
 18.         qui bysort country_id: carryforward `var', gen (`varname'_`var')
 19.         drop `var'
 20.         replace `varname'_`var' = normal(`varname'_`var')
 21.         qui replace `varname'_`var' = . if missing(`varname2')
 22.         }
 23.         order country_text_id country_id year `varname2'
 24.
.         save `varname'_post.dta, replace
 25.
}
```

After the three .dta files (one for each of the posteriors) were created, we can merge them with the rest of the variables we are using in this analysis.

```
* DO-FILE: "Example4.do"

* Merging with V-Dem Dataset
. clear

. matrix drop _all

. local files : dir
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\ind_posteriors"
files "*.dta"

. cd
"C:\Users\fbizz\Dropbox\ShamrockSeries\Fernando\tutorial_CIs\ind_posteriors"

. use "C:/Users/fbizz/Dropbox/Dissertation/Data/vdem_cy_v5.dta", clear
(V-Dem Country-Year Dataset v5. Team)

. keep country_id country_name year e_migdppcln e_peinfmor v2x_accex
v2x_suffr

. foreach file in `files' {
  2. merge 1:1 country_id year using `file', nogenerate
  3. }

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            23,568
    -----------------------------------------

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            23,568
    -----------------------------------------

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            23,568
    -----------------------------------------
. }
```

The next step is to generate the Multiplicative Electoral Democracy Index. Note that this index includes both indices for which we have uncertainty estimates as well as indices without uncertainty estimates (v2x_suffr and v2x_accex). In order to create the multiplication, we multiply the $n^{th}$ draw of each posterior, and the single values we have for v2x_accex and v2x_suffr 900 times.

```
* DO-FILE: "Example4.do"

. * generate multiplicative indices
. forvalues i = 2(1)901 {
  2. qui gen edi_mult_`i' = frefair_v`i' * frassoc_thick_v`i' *
freexp_thick_v`i' * v2x_accex * v2x_suffr
  3. drop frefair_v`i' frassoc_thick_v`i' freexp_thick_v`i'
* Drop to avoid too many variables in the Dataset
  4. }
```

With the 900 "edi_mult_i" columns created, we can now repeat the same code for the MCMC procedure described in the other examples.

```
* DO-FILE: "Example4.do"

. ***Monte Carlo Estimates Using 900 Draw Posterior Distribution***
. *Run the monte carlo
.
. forvalues i = 2(1)901 {
  2.
. *Print out an iteration number
. display `i'
  3.
. *Fit the model, using the ith draw from the UDS posterior
. quietly xtreg F.e_peinfmor edi_mult_`i' e_migdppcln, fe vce(cluster
country_id)
  4.
. *Extract the coefficients and variance-covariance matrix
. matrix b = e(b)
  5. matrix V = e(V)
  6. local blength = colsof(b)
  7. matrix rsq = e(r2)
  8.
. *Preserve the dataset, take a single multivariate normal draw from the
. *posterior distribution of the coefficients, and restore the dataset.
. *We use the capture command to catch possible errors in drawnorm
. *and drop these iterations gracefully.
. preserve
  9. capture quietly drawnorm b1-b`blength', double n(1) means(b) cov(V)
clear
 10. if _rc==0 {
 11. mkmat b1-b`blength', matrix(bsample)
 12. matrix posterior = nullmat(posterior) \ bsample
 13. matrix rsquared = nullmat(rsquared) \ rsq
 14. }
 15. else {
 16. display "Error drawing sample...iteration dropped"
 17. }
 18. restore
 19. *Closes the Monte Carlo Loop }
```

As in the previous examples, Stata has stored all the regression results in the background. We can retrieve the mean, standard deviation, and upper and lower bounds of the coefficients with the codes:

```
* DO-FILE: "Example4.do"

. *Get posterior ready to work with
. svmat posterior

. svmat rsquared

.
. *Calculate means and standard deviations
. tabstat posterior*, stat(mean sd)

   stats |  poster~1  poster~2  poster~3
---------+------------------------------
    mean | -44.83145 -42.54992  420.3559
      sd |  7.012106  3.954331  30.45544
----------------------------------------

.
. *Find the bounds of the 95 percent credible interval
. centile posterior*, centile(2.5, 97.5)

                                                      -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+------------------------------------------------------------
  posterior1 |       900        2.5   -58.30316       -59.42383   -56.66825
             |                 97.5   -31.34565       -32.17088     -29.824
  posterior2 |       900        2.5   -51.12466       -51.94955   -50.13505
             |                 97.5   -35.13378       -35.78853   -34.13374
  posterior3 |       900        2.5    364.2492        354.9419    368.2915
             |                 97.5    486.4699        478.9015    495.1632

.
. * Find the R-Squared
. tabstat rsquared*, stat(mean sd)

    variable |      mean        sd
-------------+--------------------
   rsquared1 |  .5145205  .0018353
-----------------------------------
```

The mean coefficient for MEDI (-44.8) is similar to the coefficient in the baseline model (-46.5) and both are statistically distinct from 0.

This method is robust to all different strategies of aggregation. Users that intend, for instance, to build an indicator using factor analysis can generate n factor scores for analysis combining the $n^{th}$ draw of each variable of interest and replicate the tests here described. Similarly, users can

include the index as a dependent variable in the analysis by just switching the position of the "variable_i" command in the regression equation.

## Example 5. Bootstrap C Variables

*This is still under construction.*