ACRO: guide for system managers and output checkers

Contents

Version 0.1f (Pilot)

March 2021

Authors

Felix Ritchie, Elizabeth Green and Jim Smith, University of the West of England, Bristol.

For queries, please contact felix.ritchie@uwe.ac.uk.

1. Overview

ACRO (Automated disclosure Control of Research Outputs) is designed to reduce the burden of checking for confidentiality risks in the outputs of researchers and analysts in a secure environment.

Researchers identify statistical results that they would like to be released from the secure environment. At present, these include tables, summary statistics, estimates and graphs. These are checked for disclosure risks, and the results of the check sent to a sheet in an Excel workbook. If the output passes the disclosure control tests, or an exception is requested (see below) then the full original output is included in the spreadsheet. Output which is rejected (either automatically or after manual review) is not included on the spreadsheet. The researcher sees on screen which output is automatically refused, and can also check the spreadsheet. The researcher can then update his or her code as necessary.

Researchers can request for unacceptable values in tables to be automatically suppressed. They can also request that an 'exception' be granted; that is, a request would normally fail the checks but there might be reasons why the output can be released, such as the transformation of the data (or publication of weighted results only with high weights).

When the researcher has completed her work, then the Excel workbook can be sent to the output checking team, who will review the exception requests and release (or not) the output to the researcher.

The disclosure checking process operates in the background and does not affect the normal display of results. However, only results saved to the workbook by ACRO are part of the review process. Results which cannot be managed through this process need to be submitted for review through the normal channels.

The pilot version is currently coded for Stata only.

2. System manager role: installation

The ACRO zip files contains

- Several .ado files[1]
- Two do-files, **test_file_full.do** and **test_file_small.do**, which demonstrate how the code can be used
- A sample dataset (test_data.dta) on which the test files can be run
- The macro-enabled Excel workbook **ACRO_template.xlsm** which is the output template
- The user and manager guides

These should be unzipped and placed in a folder where the users have read-only permission. This allows the users to read all the files and see the parameters, but not change them. Stata can be configured by the system manager so that it finds this folder automatically. Alternatively, the folder where the files are stored can be notified to researchers so that they can include the folder on their ado-path.

In the examples below, and in the user guide, we assume that the ACRO files have been uploaded to the folder "H:\documentation\ACRO". If Stata is configured for users to find this folder, then no

[1] An ado-file is a text file that contains a Stata program. When you type a command that Stata does not know, it looks for an ado-file of that name. Stata loads and executes it, so it appears as if the ado-command is just another command built into Stata.

further action is needed by researchers to use ACRO. If however users need to configure their own code to use this folder, they should be told to include the following at the start of their program:

```
adopath + "H:\documentation\ACRO"
```

3. System manager role: setting parameters

The program parameters are set in the file **safe_globals.ado**. This is a plain text file, containing 'global macro' definitions. The code comes in two sections, highlighted as "Systems managers – edit these values" and "do not edit". The latter group control the functioning of ACRO and should not need to be changed.

The first value to be set is the location of the output template **ACRO_template.xlsm**:

```
global ACRO_output_template_file = "H:\documentation\ACRO
\ACRO output template.xlsm"
```

The next two parameters are the name to be given the default set of SDC parameters, and the default set of SDC tests to be run for tabular outputs:

```
global safe_SDC_set = "Default"

global safe_tests = "threshold nk pratio maxmin"
```

The default SDC name will appear in the outputs, to indicate that the default set of parameters is being used. The four tests (a threshold test, the two dominance tests, and a max/minimum test) are only for tabular outputs. For 'safe statistics' such as regressions, a degrees-of-freedom test is the only test and is always used. The manager can select which safe tests are run by editing the global for example:

```
global_safe_tests = "threshold nk maxmin"
```

Will result in the outputs being checked using threshold, nk, and maxmin parameters. These parameters can also be edited and set according to the statistical organisations rules :

Parameter	Check	Valid values (Initial settings)	Description
safe_threshold	Threshold	Any positive integer (10)	Minimum number of observations for tables and other simple statistics
safe_dof_threshold	Threshold	Any positive integer (10)	Minimum residual degrees of freedom for analytical outputs
safe_nk_n	Dominance	Any positive integer (2)	The sum of the largest n values do not exceed k% of the total of all values
safe_nk_k	Dominance	0-1 (0.90)	
safe_pratio_p	Dominance	0-1 (0.10)	The sum of the smallest N-2 values exceed at least p% of the largest value

The initial settings are for testing purposes and do not necessarily reflect any one organisation's or statistical domain's parameter values; they are values commonly used by several organisations.

To change the parameters, directly edit the file using Stata's do-file editor, or any text editor such as Notepad (if editing in Word or other complex programs, it must be saved as a plain text file.)

For example. Line 40 of the file currently reads

```
global safe_dof_threshold = 10
```

To change this to a threshold of 30 degrees of freedom (this means that some of the regressions in the test file will fail), simply edit the file to read

```
global safe_dof_threshold = 30
```

and save.

Dataset-specific thresholds can be specified. In the absence of dataset-specific rules, the default ones are used. In the original set up, rules for the CIS and ESS have been added as examples. To create a rule for a specific dataset, the system manager needs to

- Add the name of the dataset to the global macro "safe_SDC_variations"
- Specify the rules by adding the name of the dataset to the 'base rules

So, in the version of **safe_globals.ado** supplied, the default values are:

```
global safe_threshold = 10
global safe_dof_threshold = 10
global safe_nk_n = 2
global safe_nk_k = 0.90
global safe_pratio_p = 0.10
```

But the file also specifies that there are rules for the CIS and ESS:

```
global safe_SDC_variations = "CIS ESS"
```

And gives these rules as

```
global safe_thresholdCIS = 60
global safe_nk_nCIS = 5
global safe_nk_kCIS = 0.50
global safe_testsCIS = "threshold nk pratio"

global safe_thresholdESS = 15
global safe_dof_thresholdESS = 10
global safe_nk_nESS = 2
global safe_nk_kESS = 0.90
global safe_testsESS = "nk pratio"
```

Note that not all the rules are re-specified. These will use the default value. So, users of the CIS data will face the default p-ratio rules, as there is no specific threshold set for that. In the example above for ESS users, no threshold or max/min test is to be run, so setting the threshold has no effect.

Dataset names should be specified in upper case (that is, "CIS" not "Cis"). In the user code, dataset names are converted to upper case, so they are not case sensitive as far as the user is concerned.

Once **safe_globals** has been run, and the ado path set, the SDC parameters can be checked by running the command **safe_show_SDC_parameters**. If a dataset name is included, then the SDC parameters for that dataset are printed. For example, on the test data set up, running

```
safe_show_SDC_parameters ess
```

will display

```
*********************************************************
SDC parameters for dataset [ESS]
*********************************************************
Threshold for descriptive statistics: 15
Threshold for degrees of freedom:     10
Dominance (n, k model):               2 largest units contribute no more than .9 of the total
Dominance (p-ratio model):            N-3 smallest units account for at least .15 of the largest unit
SDC rules being applied:              nk pratio
```

If the users combine datasets, it is not clear what rules should apply. The system manager will need to provide advice to users on which value to use. If users are likely to combine datasets, it might clarify to label the variants by *type* of dataset eg

```
global safe_SDC_variations = "social business"
```

Note that these are just labels for ACRO to use; one could also define SDC rules as

```
global safe_SDC_variations = "hi_risk med_risk low_risk"
```

4. Output checker role: reviewing and approving output

The program **test_file_small.do** created the example output used below.

4.1 User action: storing and submitting results for checking

When running ACRO, the user will be sending the output to an Excel workbook. Once the user has finished, the output checker needs to know that this file is ready for checking. How this is done will depend on the local safe centre procedures. In some systems, for example, the user copies the workbook to a specific folder and notifies the checker. In others, the user has the ability to mail messages to the checker. It should be noted that the workbook is not to be renamed as this will result in failure of the macro checks.

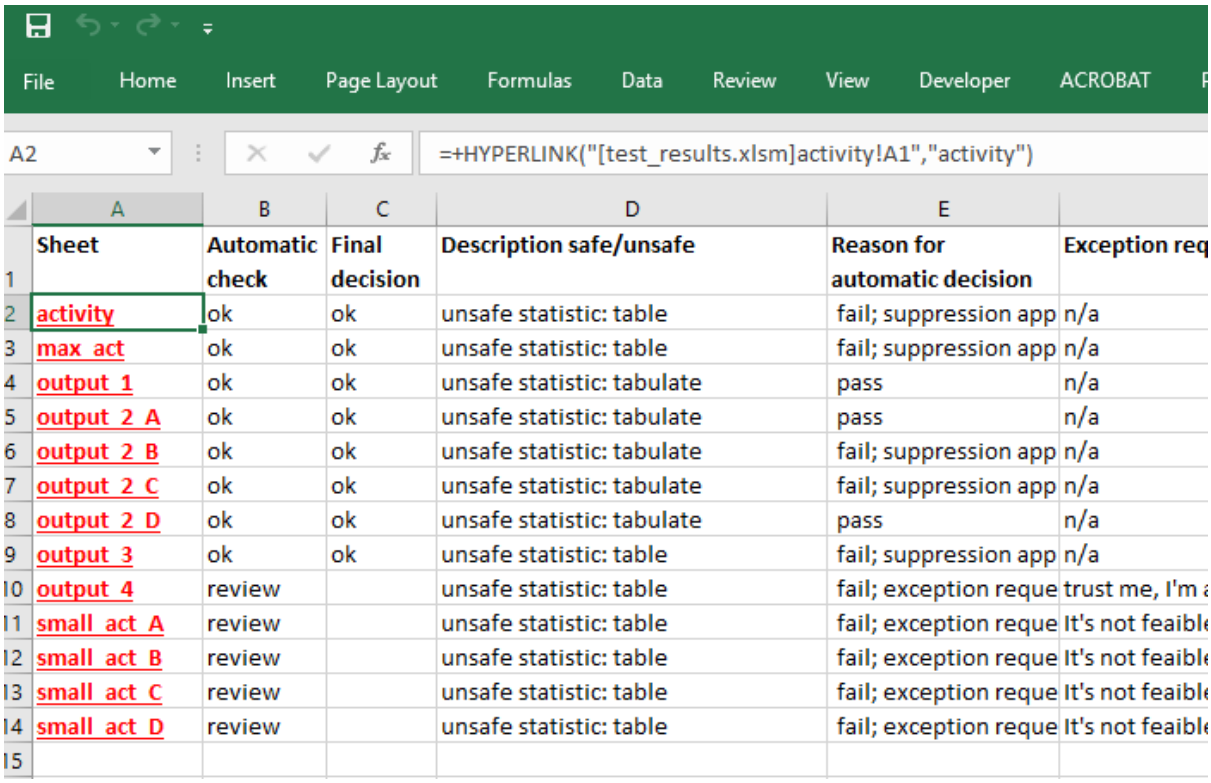
4.2 Checker action: reviewing 'output for review'

The checker should open the workbook, selecting "enable macros" if prompted by Excel.

Once the workbook is opened, the first 'description' sheet summarises all the information in the workbook: results that have passed, commands that have failed, and results which require review (see the User Guide for more details on these different outcomes and how they appear in the workbook). The 'final decision' comment is prefilled with the results of the automatic checking where this has come to a definite conclusion (pass or not pass). The column is left blank if there is a need for manual review (for example, if the user requests an exception, or if this is an output such as a graph that can only have a manual review).

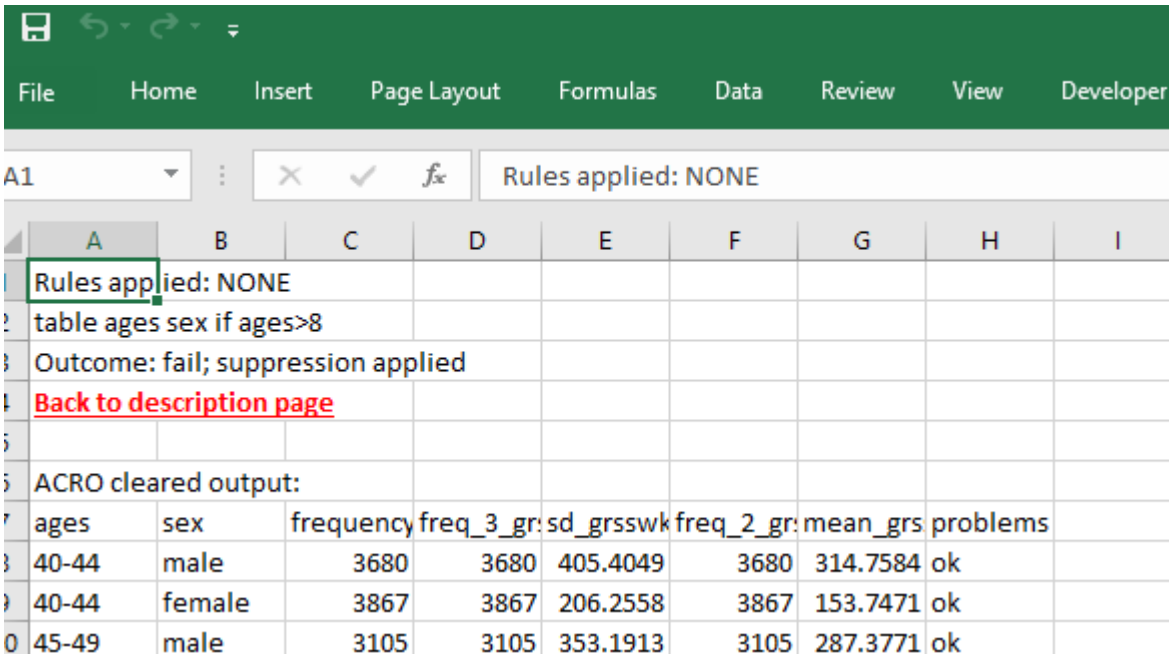
The output checker's role is to determine whether the outputs marked as 'for review' should be released or not. To do this, the checker needs to visit each spreadsheet marked as 'review' and

decide whether it can be released or not – for example, a requested exception may be justified, or a graph may have no problematic outliers or boundary cases. There are hyperlinks on each page to aid navigation (note however, that the hyperlinks are created by the user at the time of generating the statistics, and may not work if the file is copied to a location with a different folder structure):



	A	B	C	D	E	
1	Sheet	Automatic check	Final decision	Description safe/unsafe	Reason for automatic decision	Exception req
2	activity	ok	ok	unsafe statistic: table	fail; suppression app	n/a
3	max_act	ok	ok	unsafe statistic: table	fail; suppression app	n/a
4	output_1	ok	ok	unsafe statistic: tabulate	pass	n/a
5	output_2_A	ok	ok	unsafe statistic: tabulate	pass	n/a
6	output_2_B	ok	ok	unsafe statistic: tabulate	fail; suppression app	n/a
7	output_2_C	ok	ok	unsafe statistic: tabulate	fail; suppression app	n/a
8	output_2_D	ok	ok	unsafe statistic: tabulate	pass	n/a
9	output_3	ok	ok	unsafe statistic: table	fail; suppression app	n/a
10	output_4	review		unsafe statistic: table	fail; exception reque	trust me, I'm a
11	small_act_A	review		unsafe statistic: table	fail; exception reque	It's not feaible
12	small_act_B	review		unsafe statistic: table	fail; exception reque	It's not feaible
13	small_act_C	review		unsafe statistic: table	fail; exception reque	It's not feaible
14	small_act_D	review		unsafe statistic: table	fail; exception reque	It's not feaible
15						

Figure 1 The description sheet, including hyperlinks (in red)



	A	B	C	D	E	F	G	H	I
	Rules applied: NONE								
	table ages sex if ages>8								
	Outcome: fail; suppression applied								
	Back to description page								
	ACRO cleared output:								
	ages	sex	frequency	freq_3_gr	sd_grsswk	freq_2_gr	mean_grs	problems	
	40-44	male	3680	3680	405.4049	3680	314.7584	ok	
	40-44	female	3867	3867	206.2558	3867	153.7471	ok	
	45-49	male	3105	3105	353.1913	3105	287.3771	ok	

Figure 2 Results sheet, with hyperlink back to front page (in red)

If this manual review approves the output, then the checker returns to the 'description sheet and enters 'ok' in the 'final decision' column. If the checker decides that the output does not pass a

manual review, then she enters 'fail' (not case-sensitive) in the 'final decision' column. The 'final decision' column has been automatically prefilled to some extent, but this can be partially overruled by the output checker (see 4.3 below).

See the example below. On the left is the original output produced by ACRO. On the right is the output after the checker has carried out her review. In this example 'output_4' is rejected but 'small_act' has been approved:

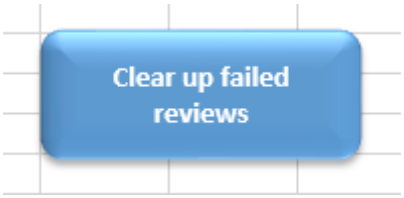
A	B	C	
Sheet	**Automatic check**	**Final decision**	**Descript**
activity	ok	ok	unsafe s
max_act	ok	ok	unsafe s
output_1	ok	ok	unsafe s
output_2	ok	ok	unsafe s
output_3	fail	fail	unsafe s
output_4	review		unsafe s
small_act	review		unsafe s

Figure 3 Original outcome

A	B	C	
Sheet	**Automatic check**	**Final decision**	**Descripti**
activity	ok	ok	unsafe s
max_act	ok	ok	unsafe s
output_1	ok	ok	unsafe s
output_2	ok	ok	unsafe s
output_3	fail	fail	unsafe s
output_4	review	fail	unsafe s
small_act	review	ok	unsafe s

Figure 4 final decisions after manual review

The checker then clicks on the button "clear up failed reviews":



This goes through the 'final classification' column and deletes the information in any sheets that were marked for review, but given a 'fail' by the output checker. Only the original command is retained in the sheet. In the 'additional notes' column, the sheet is marked as "Not cleared after review; output deleted" if it failed the review.

If the button is pressed several times, this has no effect if it finds no sheets to clear. Hence, the checker can review and clear results sequentially if preferred (pressing the clear button after each sheet is checked). The clearing is final and cannot be undone (the output checker could of course load up the file again), and so for good practice this file should now be saved under a new name.

The checker may want to add a comment in the 'notes' column, explain why the review was rejected. This should help the user to generate more appropriate results in future.

4.3 Checker role: overriding automatic checking

There may be a reason why the output checker wants to override the automatic check. For example, in version 1 of ACRO there is no mechanism to automatically check class/group disclosure. In this case, the checker can easily force the sheets to be deleted by over-writing the final decision column to read "fail":

	A	B	C	
	Sheet	**Automatic check**	**Final decision**	**Description**
	activity	ok	ok	unsafe stati
	max_act	ok	ok	unsafe stati
	output_1	review	fail	unsafe stati
	output_2	ok	ok	unsafe stati
	output_3	fail	fail	unsafe stati
	output_4	review	fail	unsafe stati
	small_act	review	ok	unsafe stati

This will delete the problematic sheet. However, this will show an inconsistency between the automatic check (approval) and the outcome (rejection). We therefore recommend that, in this case, the 'notes' column be used to explain the outcome.

Note that is not possible to over-ride a fail generated by ACRO, as this will not place the results in the sheet for review.

4.4 Checker role: returning the output to the user

Having completed the manual checks, the checker should return the workbook to the user outside the secure environment in accordance with the normal operating procedures.

The final workbook can be saved as an xlsx file – the macros are no longer needed.

This is the end of the checking process.

Note on file re-naming

The hyperlinks within the file are written by ACRO, based on the name of the file to be created. If the file is subsequently renamed, the hyperlinks will not work, making navigation harder. The macro is unaffected by this as it only needs to know the sheet names.