Improving Survey Data Quality Assurance with a User-Friendly Stata Package

NATHAN CUTLER, DANAE ROUMIS

FedCASIC: March 4, 2015

SOCIAL IMPACT

Introduction

↑ Attention to electronic data collection	Less attention to tools addressing some practical aspects of data quality monitoring
• Programming survey logic and constraints/validation • Eliminates time-intensive and error-prone data entry • Eliminates need for printing, transport, and storage of paper surveys • Frequent and rapid data upload can enable correction of discrepancies on a timely basis…	• Survey managers still need a way to efficiently inspect data • Communication between analysts and field staff needs to be systematic, organized, timely • Data checks conducted should be done consistently, documented in replicable way

Survey Data Quality

Even with electronic data collection…

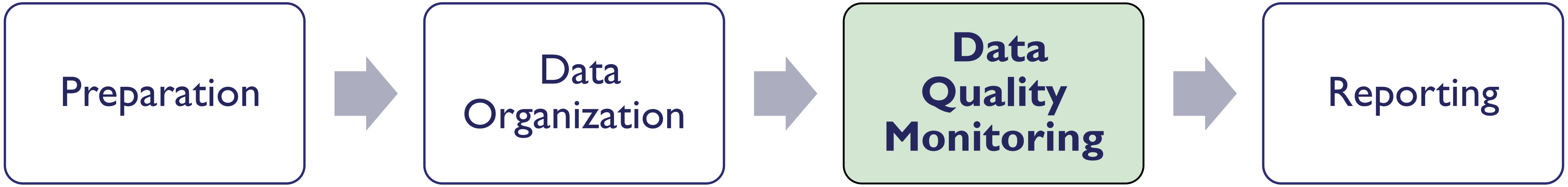
- Survey supervisors sometimes have less control over inspecting surveys with EDC as they are uploaded

- Lack of complete monitoring of data quality (checking protocols often involve sub-sample of questions or surveys; supervisors only responsible for own-interviewer questionnaires), making survey-wide issues more difficult to identify if relying on traditional approaches to data quality monitoring.
  - Enumerator tendencies or discrepancies between survey sites
  - Unusual patterns, even if valid (high rates of DK; surprising values)

- Imperfect programming (errors may persist even after detailed programming and piloting)

- More difficult to check on issues related to complex combinations of variables (i.e. indicators involving calculation of quantities using units that need to be standardized)

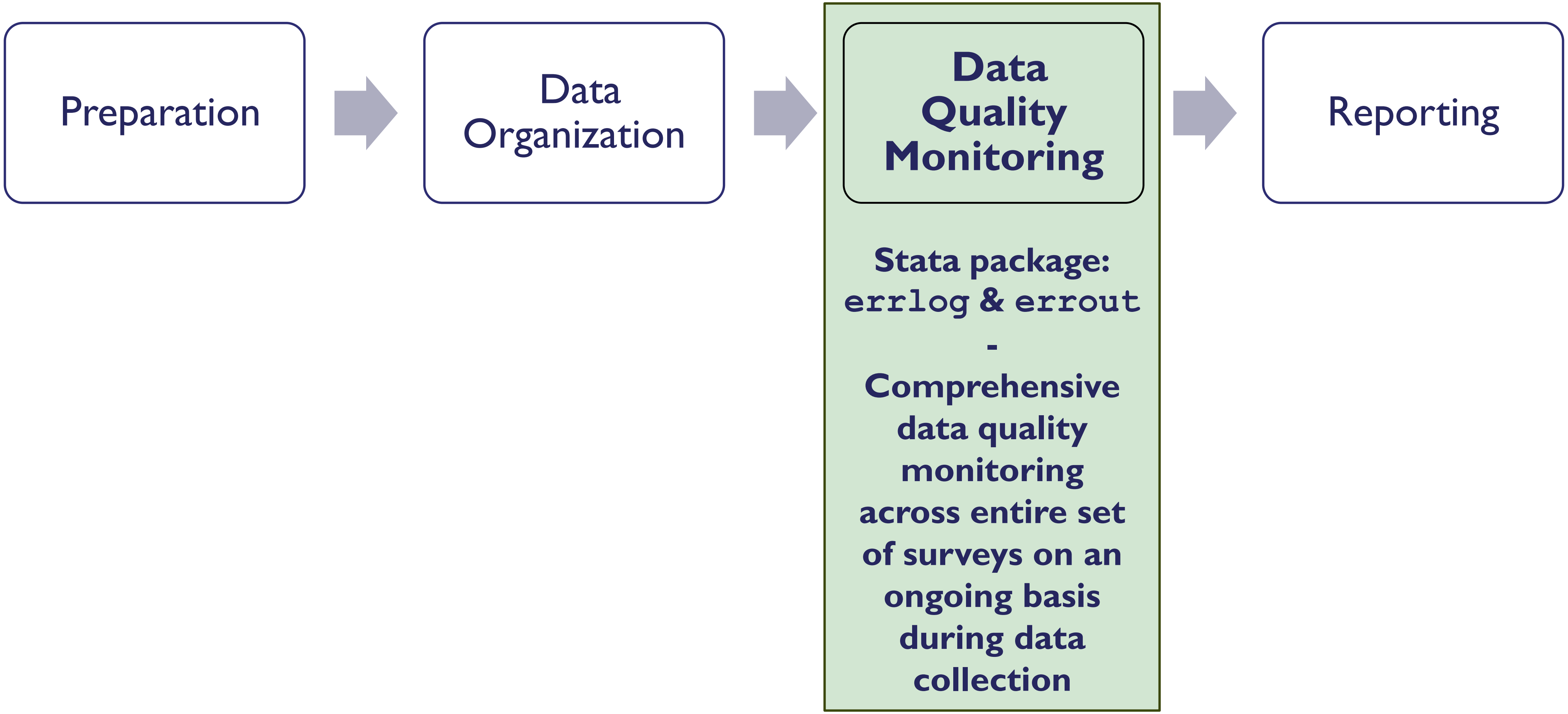
Data Quality Assurance Toolkit



System governed by a comprehensive toolkit containing a set of checklists, templates, and guidelines.

Tools are customized to the needs of each project, and cover all activities in the evaluation life cycle, including preparation and start-up, fieldwork and data collection, data management and data quality monitoring, and analysis and reporting.

Data Quality Assurance Toolkit



Preparation

Data Organization

Data Quality Monitoring

Stata package: `errlog` & `errout`

-

Comprehensive data quality monitoring across entire set of surveys on an ongoing basis during data collection

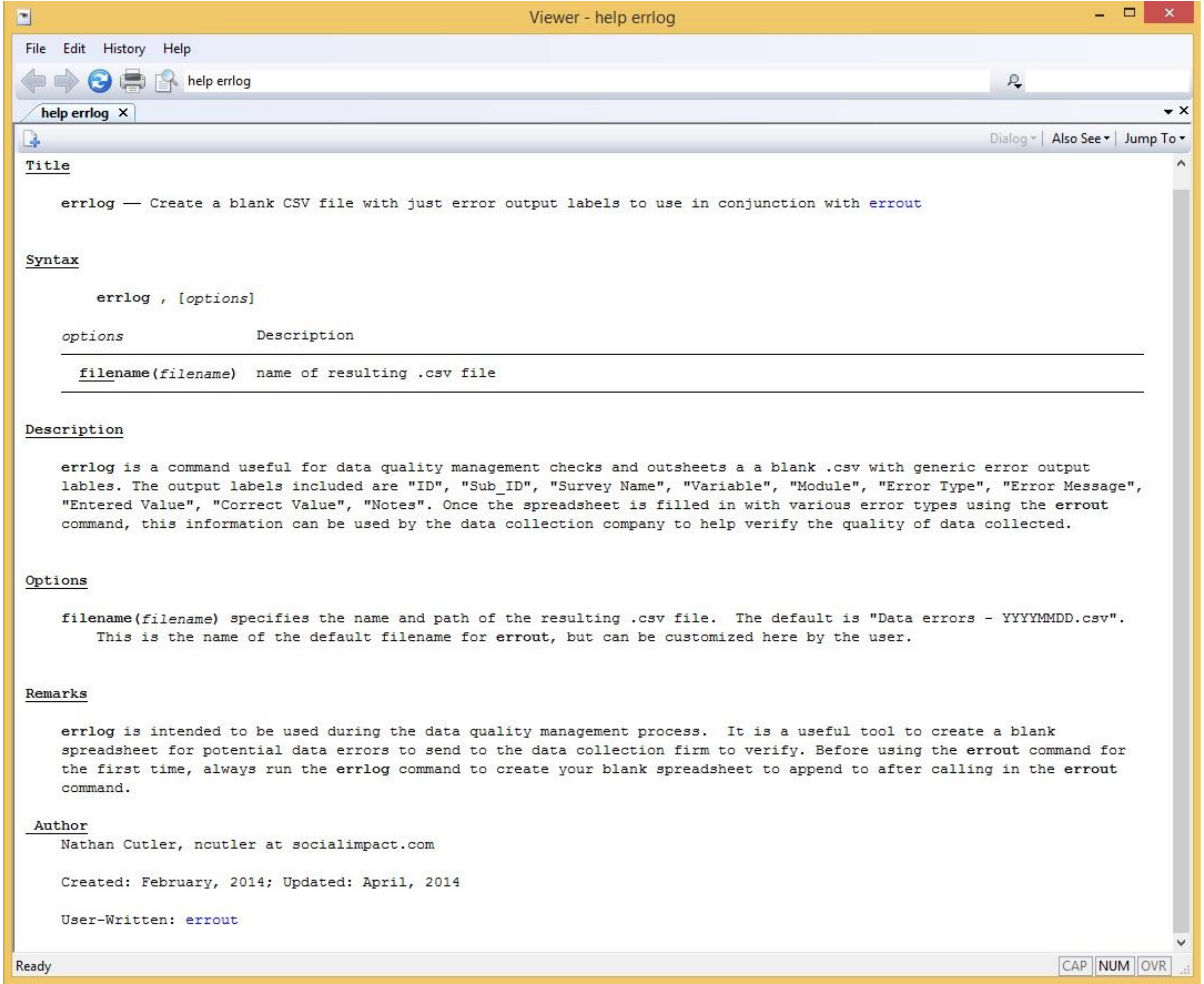
Reporting

-errlog-

- Creates blank CSV file with error output labels
- Only needs to be run once for a specific CSV file
- Spreadsheet can be sent to data collection firms for review once output is created
- Must be used prior to **-errout-**
- Syntax:

```
errlog , [options]

options                 Description
----------------------------------------------------------------------
filename(filename)      name of resulting .csv file
----------------------------------------------------------------------
```



Title

errlog — Create a blank CSV file with just error output labels to use in conjunction with errout

Syntax

errlog , [*options*]

options	Description
filename(*filename*)	name of resulting .csv file

Description

errlog is a command useful for data quality management checks and outsheets a a blank .csv with generic error output lables. The output labels included are "ID", "Sub_ID", "Survey Name", "Variable", "Module", "Error Type", "Error Message", "Entered Value", "Correct Value", "Notes". Once the spreadsheet is filled in with various error types using the **errout** command, this information can be used by the data collection company to help verify the quality of data collected.

Options

filename(*filename*) specifies the name and path of the resulting .csv file. The default is "Data errors - YYYYMMDD.csv". This is the name of the default filename for **errout**, but can be customized here by the user.

Remarks

errlog is intended to be used during the data quality management process. It is a useful tool to create a blank spreadsheet for potential data errors to send to the data collection firm to verify. Before using the **errout** command for the first time, always run the **errlog** command to create your blank spreadsheet to append to after calling in the **errout** command.

Author

Nathan Cutler, ncutler at socialimpact.com

Created: February, 2014; Updated: April, 2014

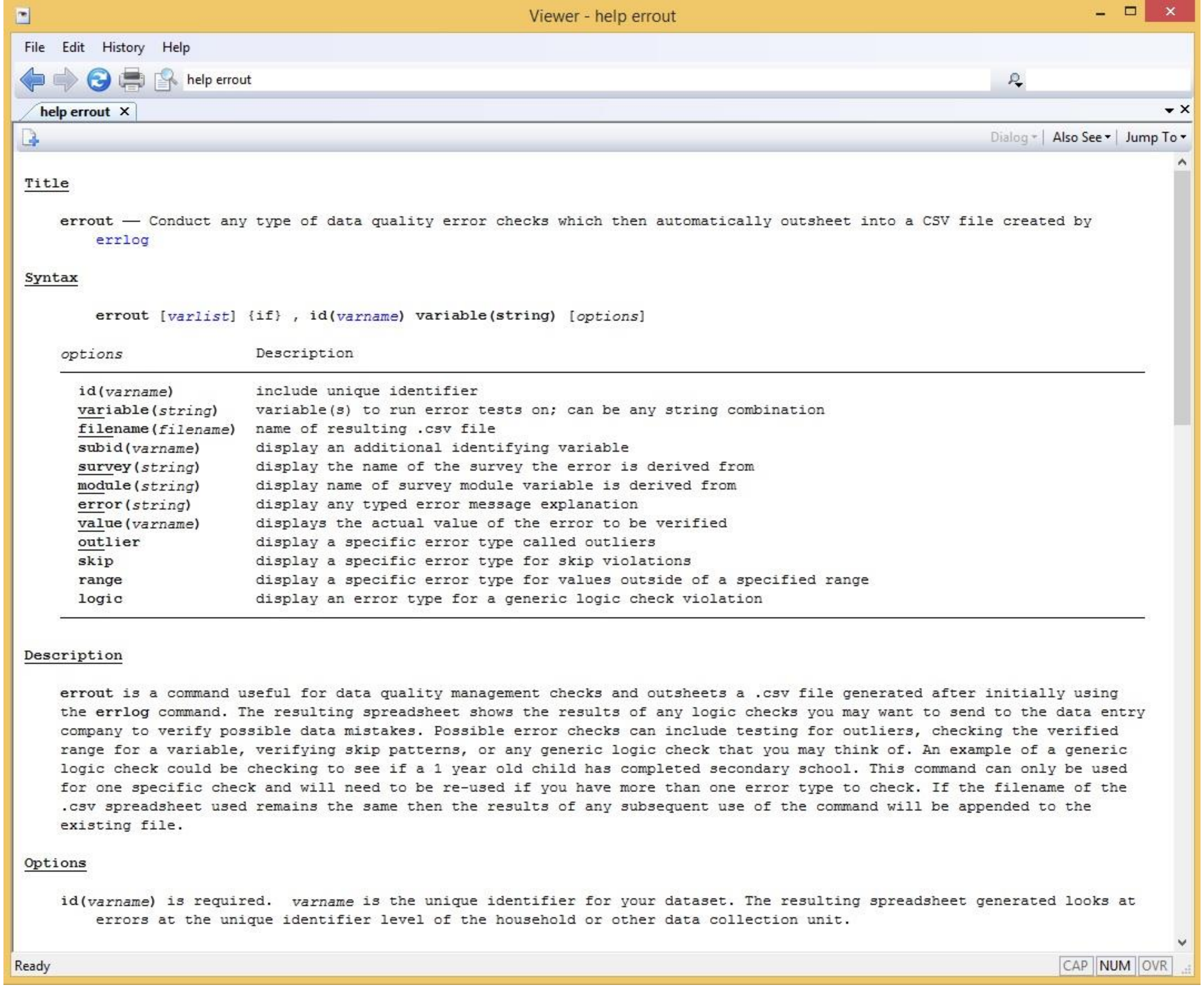
User-Written: errout

-errout-

- Outputs data quality errors into the CSV file created by `-errlog-`
- Output any errors or logic check that the user can come up with (skip violations, outliers, illogical, strange patterns, calculated values, etc.)
- Syntax:

```
errout [varlist] {if} , id(varname) [options]
```

  - The variety of options specify which columns (`-errlog-` labels) are filled in the spreadsheet
  - User has control over whether entered values and corrected values are entered into spreadsheet
  - Options include the filename, any sub-id, survey name, type of error, and the actual value entered
  - Extremely flexible command – any calculation can be performed prior to running this line of code



Title

errout — Conduct any type of data quality error checks which then automatically outsheet into a CSV file created by errlog

Syntax

```
errout [varlist] {if} , id(varname) variable(string) [options]
```

options	Description
id(*varname*)	include unique identifier
variable(*string*)	variable(s) to run error tests on; can be any string combination
filename(*filename*)	name of resulting .csv file
subid(*varname*)	display an additional identifying variable
survey(*string*)	display the name of the survey the error is derived from
module(*string*)	display name of survey module variable is derived from
error(*string*)	display any typed error message explanation
value(*varname*)	displays the actual value of the error to be verified
outlier	display a specific error type called outliers
skip	display a specific error type for skip violations
range	display a specific error type for values outside of a specified range
logic	display an error type for a generic logic check violation

Description

errout is a command useful for data quality management checks and outsheets a .csv file generated after initially using the **errlog** command. The resulting spreadsheet shows the results of any logic checks you may want to send to the data entry company to verify possible data mistakes. Possible error checks can include testing for outliers, checking the verified range for a variable, verifying skip patterns, or any generic logic check that you may think of. An example of a generic logic check could be checking to see if a 1 year old child has completed secondary school. This command can only be used for one specific check and will need to be re-used if you have more than one error type to check. If the filename of the .csv spreadsheet used remains the same then the results of any subsequent use of the command will be appended to the existing file.

Options

id(*varname*) is required. *varname* is the unique identifier for your dataset. The resulting spreadsheet generated looks at errors at the unique identifier level of the household or other data collection unit.

Example

```
errlog, filename(Enterprise Errors)

errout if Q701_Y<1985, id(EST_NO) var(Q701_Y) file(Enterprise Errors)
surv(Enterprise) mod(Module 7) err(Value cannot be after 1985)
val(Q701_Y) range
```

EST_NO	Variable	Survey Name	Module	Error Type	Error Message	Entered Value	Correct Value	Notes
6	Q701_Y	Enterprise	Module 7	Value outside of range	Value cannot be after 1985	1978	1995	First sewer connection network in Zarqa was in 1985
45	Q701_Y	Enterprise	Module 7	Value outside of range	Value cannot be after 1985	1979	1995	First sewer connection network in Zarqa was in 1985

Case Study

Millennium Challenge Corporation (MCC)

Jordan Compact

Water Sector Impact Evaluation

Overview

- Household survey of 3000+ respondents across Zarqa and Amman Governorates
- 50+ Enumerators
- Survey administered on Android tablets using SurveyCTO/ODK software

Data Quality Monitoring Constraints

- Constant and thorough DQM can be a challenge even with EDC
- Large number of surveys uploaded on a daily basis to the servers, and therefore supervisors face high volume of quality control checks

Solution to Constraints

- SI wrote Stata code with –errout– and –errlog– ahead of time to run data quality checks on frequent basis during data collection
- Send output sorted by household to data collection firm to verify observations
- Data collection firm used information to re-train enumerators
- Output used to verify data and make any corrections needed during fieldwork

Lessons Learned

- Communication with the data collection firm is paramount; detecting more potential issues may add workload to data collection firm; set reasonable and mutual expectations about the process
- Spreadsheet may include output that do not reflect real issues in need of correction, data checks should not be overly stringent, and survey managers should review output before sending to data collection firm
- The tool is an effective platform for collaboration on data quality monitoring, eliminates duplication of checks, and can be updated as new ideas and challenges are discovered by different team members

Next steps & release

Package to be released March 2015

Will be released on SI Website: http://www.socialimpact.com/services/impact-evaluation.html and uploaded to SSC directory for easy download in Stata

Next Steps

- Possibility of integrating the two commands & adding additional options
- Possibility of creating additional option to output/export meta information about errors

Questions and feedback welcomed

Contact: Nate Cutler, Social Impact: ncutler@socialimpact.com