

Health and Education Monitoring and Evaluation Capacity Building Project  
Using Stata Software in Data Analysis

# Final Report

Prepared by Anna Krivelyova, ICF Macro

Tashkent, Uzbekistan

January 2010



## Table of Contents

1. Training Overview
2. Needs Assessment
3. Performance of Participants
4. Conclusions and Recommendations
5. Appendices
  - Appendix A. Training Program

## Training Overview

In December 2009, as part of the Health and Education Monitoring and Evaluation Capacity Building Project, ICF Macro in collaboration with the Center for Economic Research (CER) delivered the second module of the training workshop, “*Using Stata Software in Data Analysis.*”<sup>1</sup> STATA is a full-featured statistical programming language for Windows, Macintosh, UNIX, and Linux. It is a complete, integrated statistical package that provides tools for data analysis, data management, and graphics. STATA is distributed in more than 150 countries and is used by professionals in many fields of research.

The long-term goal of this training is to build capacity of local specialists in data analysis. The specific objectives of the second mission were

- 1) to review the materials introduced during the first module of training,
- 2) to introduce participants to the basics of statistical data analysis,
- 3) to review Stata command and syntax to perform data analysis,
- 4) to make specific recommendations for future trainings.

The second module of the training was delivered during December 21–24, 2009 period in Tashkent, Uzbekistan. Prior to the training, a detailed training program was developed by ICF Macro International in collaboration with CER and Tahlil (see Appendix A). Tahlil had also forwarded to ICF Macro several databases containing survey data collected as part of the Health and Education Monitoring and Evaluation Capacity Building Project, survey instruments, and data summary tables. ICF Macro International prepared additional chapters of the training manual, which contains detailed information on multivariate regression analysis, panel data analysis and overview of Stata commands necessary to execute the analyses. First four chapters of the manual were developed for the first training module and contained a general description of Stata functionality; a review of major data management topics; a description of general Stata command syntax; instructions on using various Stata commands to manage data; and a brief introduction to the concepts of correlation, statistical testing, and Stata commands (full manual can be found in see Appendix B). Complete manual has been translated into Russian and by CER and provided to all training participants. Uzbekistan Regional Panel Survey (URPS) data were used for most of the examples in the training manual. CER translated the manual into Russian and distributed a Russian version of the full manual to the training participants. In addition, training participants were provided with access to electronic files with sample Stata data sets. The sample datasets contained select variables from the household survey dataset and were customized depending on the topics covered during the training. The training was conducted in Russian.

In general, the training curriculum entailed a combination of classroom teaching and participatory experiences providing participants with hands-on, practical experiences throughout the training. Since some of the training participants were new and also a rather long break between the two training modules, the training started with the brief review of Stata basics covered during the first training module. This review included information on the Stata interface (e.g., Stata windows, toolbars, and menus), work documentation (e.g., using

---

<sup>1</sup> I would like to express my deepest gratitude to the CER, and in particular to the deputy director Ildus Kamilov, for their invaluable advice in designing the training, thoughtful selection of training participants, flawless logistics, and warm hospitality during my stay in Uzbekistan.

*log* and *do-files*). basic data management (e.g., loading datasets, describing datasets, producing basic descriptive statistics and frequency tables), and variable creation and transformation.

Next elementary statistical concepts were reviewed (e.g., correlation, chi-square tests, *t* tests). After sufficient practice with simple statistical testing, participants were introduced to the concept of multivariate analysis. Ordinary least squares (OLS) regression was covered in much detail to provide a solid foundation for additional statistical methods covered later in the course. The training covered theoretical concepts behind the multiple regression analysis and gave numerous examples of applications of this statistical methodology. The trainer also reviewed theoretical assumptions of the OLS regression and discusses problems that arise if these assumptions are violated (e.g., non-linearity, multicollinearity, heteroskedasticity). The class was presented with the examples of Stata commands that can be used to test potential violations of OLS assumptions and methods of correcting some of the violations (e.g., robust standard errors, logarithmic transformations). Other methods that can be used in dealing with violations of OLS assumptions (e.g., instrumental variables method) were not presented due to time limitations. Additional topics included in the training were logistic regression and multinomial logistic regression.

Throughout the training, participants practiced completing various tasks in Stata. When each new command was introduced, participants repeated the actions of the trainer as projected on the screen. After sufficient practice with each new command, participants were given exercises to complete independently. Participants were encouraged to ask questions and ask for help if needed. Help was provided by the instructor or by other participants who already had mastered the particular skills. At the end of each training day, about one hour was allocated to individual consultations, which were provided to participants who needed extra help understanding the material and to more advanced students who wanted to learn additional topics and commands.

On the last day of training, a time was allocated to provide an overview of more advanced statistical methods, such as fixed and random effects models and time-series analysis. Since survey data collected as part of the Health and Education Monitoring and Evaluation Capacity Building Project, was not suitable to be used as an example for time-series analysis, instructor has used daily and monthly stock market volatility index data available for United States markets (VIX index).

## Needs Assessment

While no formal needs assessment was conducted during the training, participants were assessed informally in terms of their experience working with statistical software packages, their understanding of statistical concepts, and their experience with data analysis. The majority of the participants had little or no prior experience working with integrated statistical software packages. Many were limited to using Excel for most of the data management and analysis tasks, or using data systems specific to their agencies (e.g., Medstat). Others were experienced users of SPSS and/or Stata. The level of understanding of statistical concepts and methods among participants was highly variable. While some were familiar with sampling and regression methodologies, others needed to be introduced to basic concepts such as distributions and descriptive statistics.

## Conclusions and Recommendations

Based on feedback from the participants and observed performance, overall, the training objectives were met. While there was variation in both initial skill and experience levels and in performance throughout the training, all of the participants who attended the training had gained at least basic skills in using Stata for data analysis. The CER team did an excellent job in selecting training participants. While representing different professional backgrounds and experiences, all participants were driven to learn new concepts and gain new skills. Though, the level of statistical background differed greatly among participants, all attendees actively participated in practice exercises and discussion and interpretation of results of statistical analysis. All of the planned topics for the first module were covered, including the introduction to basics of data analysis, such as correlations and simple statistical testing. The recommendations outlined below are based on the discussion with the training participants, needs assessment and observed performance. The recommendations below are divided into two groups. First set of recommendations are aimed to improve future trainings in statistical software. Second set consists of more general recommendations on future trainings aimed at improving skills in statistical analysis.

### **Future trainings in software for data management and analysis**

*The participants should be provided with basic training in statistics prior to training in software for data analysis.* The lack of statistical skills among many of the training participants makes it challenging to limit the instruction to cover only topics related to Stata or any other statistical software package. In consultation with the CER, this particular training has been structured to combine reviewing of theoretical statistical concepts and practical application in Stata. However, we suggest that future trainings in software should be preceded by extensive training in statistics. The topics covered in the first module of the training did not require extensive knowledge of statistical data analysis and in the future can be offered to a wider audience. However, in order to cover full analytical capabilities of statistical software packages, such as Stata or SPSS, the audience should have a more extensive background in statistics. Variation in initial skills and how quickly the participants absorbed new material during the training, sometimes resulted in situations where more advanced students felt that they were held back. Such situations can be difficult to avoid in educational settings, but should be minimized.

*Participants should be provided with access to Stata software.* The skills gained by participants during the training may quickly deteriorate if participants do not continue to practice using Stata. Only a few of the participants had access to Stata software outside of the classroom. While the costs of purchasing Stata for each of the training participants may be prohibitive, it is recommended that at least limited access to Stata software should be provided to the relevant agencies.

### **Future trainings in statistical data analysis**

*Longer and more in-depth trainings are needed.* Many of the training participants expressed concern about general lack of staff with solid statistical training in their respective agencies. Unlike trainings on the use of statistical software packages, which can be conducted in a relatively short periods of time, in-depth training in statistics requires much longer time.

While some of the trainings should be done in person, it will also be possible to have part of the training to be delivered as a self-learning or online course. Given that the potential time requirement, a possible multi-module structure of in depth course in statistics, and the fact that training participants are likely to be busy professionals, distance learning may be an optimal option for such course. ICF Macro has extensive qualifications in designing distance-learning programs and will be happy to provide CER and the World Bank with more in-depth recommendations if so desired.

***The content of the training in statistical data analysis should include thematic modules.***

There was a great degree of variation among training participants in terms of the types of data their respective agencies collect and as a result the types of analytic techniques that are specific to different data types. For example, representatives of one of the agencies commented that most of the data that their agency collects and analyzes is relatively high frequency time series data. As a result, it will be beneficial for the statistical analysts from this agency to learn more detailed information about time series methods, rather than learn statistical issues specific to the analysis of panel data. The trainer did cover some basic information regarding time series analysis to the course participants, but it is unlikely that after this overview, the participants will be able to perform this type of analysis independently. We suggest that future trainings should have a general introductory statistics course, followed by various optional modules (e.g., time-series module, panel data analysis module, survey methodology etc). This way the participants can chose what is more relevant to the analytical and data issues that are specific to their agencies.

# Appendix A

## Training Program

# Health & Education Monitoring and Evaluation Capacity Building Project

## Using Stata Software in Data Analysis

Training Module 2: Introduction to Stata

Tashkent, Uzbekistan

December 21-24, 2009

Instructor: Anna Krivelyova, ICF Macro

Prepared by ICF Macro and CER

### Preliminary Program

DECEMBER 21, 2009	
Session	Time
<i>Introductions</i>	10.00 - 10.30
<i>Review Stata basics</i>	10.00 - 11.30
<i>Short Break</i>	11.30 - 11.45
<i>Review variable types</i>	11.45 - 13.00
<i>Break</i>	13.00 - 14.00
<i>Describing basic relationship between the variables</i>	14.00 - 15.00
<i>Short Break</i>	15.00 - 15.15
<i>Simple statistical testing</i>	15.15 - 16.15
<i>Closure</i>	16.15 - 16.30
<i>Office Hours</i>	16.30 - 18.00

  

DECEMBER 22, 2009	
Session	Time
<i>Review/preview</i>	10.00 - 10.30  - 10.00
<i>Introduction to multivariate analysis</i>	10.00 - 11.00
<i>Short Break</i>	11.30 - 11.45
<i>Ordinary Least Squares (OLS) Regression</i>	11.45 - 13.00

<b>Break</b>	13.00 - 14.00
<b>OLS Regression, cont.</b>	14.00 - 15.00
<b>Short Break</b>	15.00 - 15.15
<b>OLS Regression assumptions</b>	15.15 - 16.15
<b>Closure</b>	16.15 - 16.30
<b>Office Hours</b>	16.30 - 18.00
<b>DECEMBER 23, 2009</b>	
<b>Session</b>	<b>Time</b>
<b>Review/preview</b>	10.00 - 10.30
<b>Quiz</b>	10.00 - 10.30
<b>OLS Regression assumptions, cont.</b>	10.30 - 11.00
<b>Short Break</b>	11.30 - 11.45
<b>Violations of OLS Regression assumptions</b>	11.45 - 13.00
<b>Break</b>	13.00 - 14.00
<b>Violations of OLS Regression assumptions, cont.</b>	14.00 - 15.00
<b>Short Break</b>	15.00 - 15.15
<b>Violations of OLS Regression assumptions, cont.</b>	15.15 - 16.15
<b>Closure</b>	16.15 - 16.30
<b>Office Hours</b>	16.30 - 18.00
<b>DECEMBER 24, 2009</b>	
<b>Session</b>	<b>Time</b>
<b>Review/preview</b>	10.00 - 10.30
<b>Multivariate models for analysis of categorical variables</b>	10.30 - 11.30
<b>Short Break</b>	11.30 - 11.45
<b>Multivariate models for analysis of categorical variables, cont.</b>	11.45 - 13.00
<b>Break</b>	13.00 - 14.00
<b>Overview of other model types</b>	14.00 - 15.00
<b>Short Break</b>	15.00 - 15.15
<b>Overview of other model types, cont.</b>	15.15 - 16.15
<b>Closure</b>	16.15 - 16.30
<b>Office Hours</b>	16.30 - 18.00