Subject: Reading data files into R Studio
Posted by DHS user on Thu, 02 Feb 2017 11:28:28 GMT
View Forum Message <> Reply to Message

I am conducting research using the 1990 & 2010 Pakistan DHS. I am interested in utilizing the household and woman files to understand ways in which factors such as education, income, religion, age at marriage etc. impact a woman's family planning uptake / contraceptive use.

I have dowloaded the FLAT files for the 1990 and 2012 PDHS and there are a few issues that are preventing me from reading data files into RStudio. As an example, for file "PKBR21FL.DAT" I examined the first two rows of this data file.

1. The rows do not appear to have the same number of elements. I presume there are missing entries which might be causing this. I checked in the "Coding Standards" section of DHS VI Individual recode manual (pdf), which mentions that a value of BLANK means "Variable is not applicable for this respondent either because the question was not asked in a particular country or because the question was not asked of this respondent due to the flow or skip pattern of the questionnaire."

Without knowing how to parse the rows of the file, a statistical software program such as R/RStudio cannot read this into a data matrix of fixed size, with rows corresponding to the individual records in the flat file, and columns corresponding to all possible individual parameters/ factors. Since the data files are not comma delimited, I do not know how to proceed with parsing missing data for individuals that have been replaced with a BLANK. Does the latter mean " "?

2. It is very unclear as to what order the columns in "PKBR21FL.DAT" are in. I cannot find in the PDF manual a place where it provides a mapping between the ordering of the columns, and their definitions as they appear on page 9 (Section H00). As an example, the Country variable appears in position 6, however this appears as variable #2 (HV000) on page 9.

Your clarification of these two questions would be greatly appreciated.