Data and Replication Files for "Water Quality Awareness and Breastfeeding: Evidence of Health Behavior Change in Bangladesh" Pinar Keskin, Gauri Kartini Shastry & Helen Willis
Review of Economics and Statistics
June 2016

This document describes the do-files and data-files necessary to replicate the results in the paper, "Water Quality Awareness and Breastfeeding: Evidence of Health Behavior Change in Bangladesh." We describe how to construct the final datasets and where to find the relevant raw data.

Software, file structure and raw data

The data formation and analysis was done using Stata13, run on Unix.

The do-file "0_doeverything.do" runs all subsequent dofiles: it generates the final datasets using the raw data from the British Geological Survey and Bangladesh Demographic Health Survey and generates the tables in the paper. The dofiles assume that all dofiles are in the same folder. In that folder, the raw data must be placed in a folder called "rawdta" with separate subdirectories for each type of data, with the following names "BGS Survey", "DHS1999", "DHS2004", and "DHS2007". The created .dta files will be saved in a subfolder "dtafiles" (which the dofiles will create if necessary) and the .csv files with the tables will be saved in a subfolder "csvfiles" (which the dofiles will create if necessary).

British Geological Survey (BGS) data

The British Geological Survey data (DPHE/BGS National Hydrochemical Survey) was downloaded from this website,

http://www.bgs.ac.uk/research/groundwater/health/arsenic/Bangladesh/data.html

on October 6, 2011. We made some minor modifications to the excel file (removing header rows, etc.). The file is available on the RESTAT Dataverse. Please save the file as "rawdata/BGS Survey/BGS Survey Cleaned"

Bangladesh Demographic Health Surveys (1999-00, 2004, 2007)

The Bangladesh Demographic Health Survey (BDHS) data are available from the DHS website (http://dhsprogram.com/Data/) and, more specifically, the links below. The data are free to download and use, but users are required to register for access and explain their intent to use the data. Thus, the data has not been made available on the RESTAT Dataverse. Data on the GPS locations for the clusters in each round requires additional approval from the DHS program. Once downloaded from the DHS program, please place the Stata files in the following sub-folders.

From http://dhsprogram.com/data/dataset/Bangladesh_Standard-DHS_2000.cfm?flag=0

DHS 1999-00 files	Subdirectory and filename
GPS data	rawdta/DHS1999/BangladeshGPS_1999.dta
Child-level data	rawdta/DHS1999/BDKR41DT/BDKR41FL.DTA
Household-level data	rawdta/DHS1999/BDHR41DT/BDHR41FL.DTA
Women's survey	rawdta/DHS1999/BDIR41DT/BDIR41FL.DTA

From http://dhsprogram.com/data/dataset/Bangladesh_Standard-DHS_2004.cfm?flag=0

DHS 2004 files	Subdirectory and filename
GPS data	rawdta/DHS2004/BangladeshGPS_2004.dta
Child-level data	rawdta/DHS2004/BDKR4JDT/BDKR4JFL.DTA
Household-level data	rawdta/DHS2004/BDHR4JDT/BDHR4JFL.DTA
Community leader survey	rawdta/DHS2004/bdsq4idt/bdSQ4IFL.dta
Women's survey	rawdta/DHS2004/BDIR4JDT/BDIR4JFL.DTA

From http://dhsprogram.com/data/dataset/Bangladesh_Standard-DHS_2007.cfm?flag=0

DHS 2007 files	Subdirectory and filename
GPS data	rawdta/DHS2007/BangladeshGPS_2007.dta
Child-level data	rawdta/DHS2007/BDKR51DT/BDKR51FL.DTA
Community leader survey	rawdta/DHS2007/BDSQ51DT/bdCQ51FL.DTA
Women's survey	rawdta/DHS2007/BDIR51DT/BDIR51FL.DTA

Dofiles

Dofile 0_doeverything.do calls all of the dofiles listed below in the right order to generate the final datasets from the raw data and replicate all the tables in the paper.

Data generating dofiles:

Filename	Brief description
1_expandgps_allyears.do	Expands cluster-level BDHS data for merging with BGS wells
2_mergegps_allyears.do	Generates simple measures of arsenic contamination and .dta files necessary for calculating additional measures
3_mergehh2004.do	Merge child-level BDHS data with arsenic measures for 2004
4_mergehh2007.do	Merge child-level BDHS data with arsenic measures for 2007
5_mergehh1999.do	Merge child-level BDHS data with arsenic measures for 1999
6_simulatearsenicmeasures.do	Calculate preferred measures of arsenic contamination
7_generatingchildleveldata.do	Cleans demographic and outcome variables
8_matchnearest2004.do	Matches 1999 and 2007 clusters to the nearest 2004 cluster
9_arsenicdataprep.do	Final data preparation for household level 2004 data
10_mergeinmoredata.do	Final data preparation for child level data

Table generating dofiles:

Filename	Brief description
tables1and2.do	Estimates regressions from tables 1 and 2 and saves the results in the csvfiles subdirectory
figures2and3.do	Estimates Kernel-weighted local polynomials and cluster-bootstrapped confidence intervals for figures 2 and A3 and saves the results in the csvfiles subdirectory
maintables.do	Estimates all summary statistics and regressions from tables 3-A5 and saves the results in the csvfiles subdirectory

Data Dictionary

The two final datasets analyzed are mergedallrecoded.dta (child-level data) and HouseholdArsenic_2004.dta (household-level data).

Data dictionary for mergedallrecoded.dta:

Variable name	Brief description
Source: BDHS	
age	mother's current age
ageatdeath	age at death in months
anyeduc	any formal education
anyindivclubs	participates in any clubs
byear1996	dummy for birth year 1996
byear1997	dummy for birth year 1997
byear1998	dummy for birth year 1998
byear1999	dummy for birth year 1999
byear2000	dummy for birth year 2000
byear2001	dummy for birth year 2001
byear2002	dummy for birth year 2002
byear2003	dummy for birth year 2003
byear2004	dummy for birth year 2004
byear2005	dummy for birth year 2005
byear2006	dummy for birth year 2006
byear2007	dummy for birth year 2007
caseid	case id
childage	age of child in months
childagewithdied	age of child in months with age at death if child died
childbyear	year of birth
childdied	dummy variable for child died
childpotentialage	age of child in months with age child would have been if child died
contaminatedorsurface	household's water is contaminated with arsenic or from

	surface source
dhsid04	cluster number
dhsid04matched	DHS cluster id in 2004, and closest 2004 cluster in 1999 and 2007
dhsid2004_year	group(dhsidyear childbyear)
dhsidyear	unique year-cluster identifier
diarrhea	had diarrhea recently
district	district
electricity	has electricity
exclusivebf	child is exclusively breastfeeding
exclusivebf12	child is exlcusively breastfed for those over 12 months of age
exclusivebf6	child is exlcusively breastfed for those under 6 months of age
exclusivebf614	child is exlcusively breastfed for those between 6 and 14 months of age
fathereducyrs	husbands education-single yrs
heardofarsenic	whether heard of arsenic contamination problem
heardofarsenic_2002	interaction of post-campaign and heard of arsenic
htasdreal	z-score height for age
malechild	dummy for child is male
momwthtsd	Mother's weight for height z-score
monthsbf12	child was breastfed for at least 12 months, conditional on surviving
monthsbf18	child was breastfed for at least 18 months, conditional on surviving
monthsbf24	child was breastfed for at least 24 months, conditional on surviving
monthsbf36	child was breastfed for at least 36 months, conditional on surviving
monthsbf6	child was breastfed for at least 6 months, conditional on surviving
monthsbreast	months of breastfeeding
moreautonomy	higher than average decision-making power in household
mothereducyrs	mother's education in years
motherid	unique id to match with mother-level data
motherworks	mother currently working outside the home
muslim	dummy for respondent is muslim
pipedwater	household water source is piped water
plainwater	child was given plain water yesterday
post2002	dummy for born in 2002 or after
stillbreastfeeding	child is currently breastfeeding
surfacewater	dummy for household's primary water source is surface water
tubewell	dummy for household's primary water source is tubewell
urban	dummy variable for in an urban area
urbanrural	cluster in urban or rural area

v002	household number
v003	respondent's line number
villagepipedwater	village leader lists piped water as primary water source in cluster
wthtsdreal	z-score weight for height
year	survey year
Source: BGS matched with BDHS cluster	
as_walkable5_g	average arsenic level of wells in 5 mi
as_walkable5_g_2002	interaction of post-campaign and average arsenic level of wells in 5 mi
fraction_mindistanceC_1m_G	probability of living within 1 mi of contaminated Well
fraction_mindistanceC_1m_G_2002	interaction of post-campaign and unweighted measure of contamination
fraction_mindistanceU_1m_G	Fraction of Points within 1 m of Uncontaminated Well
fractioncont5	fraction of wells contaminated in 5 mi
fractioncont5_2002	interaction of post-campaign and fraction of wells contaminated in 5 mi
mean_mindistanceU_G	Average Distance to Closest Uncontaminated Well
numcontaminated5	number of contaminated wells in 5 mi
numcontaminated5_2002	interaction of post-campaign and number of contaminated wells in 5 mi
wfraction_mindistanceC_1m_G	weighted probability of living within 1 mi of contaminated Well
wfraction_mindistanceC_1m_G_2002	interaction of post-campaign and weighted measure of contamination
wfraction_mindistanceC_1m_G_b00	interaction of born in 2000 and weighted measure
wfraction_mindistanceC_1m_G_b01	interaction of born in 2001 and weighted measure
wfraction_mindistanceC_1m_G_b02	interaction of born in 2002 and weighted measure
wfraction_mindistanceC_1m_G_b03	interaction of born in 2003 and weighted measure
wfraction_mindistanceC_1m_G_b04	interaction of born in 2004 and weighted measure
wfraction_mindistanceC_1m_G_b05	interaction of born in 2005 and weighted measure
wfraction_mindistanceC_1m_G_b06	interaction of born in 2006 and weighted measure
wfraction_mindistanceC_1m_G_b07	interaction of born in 2007 and weighted measure
wfraction_mindistanceC_1m_G_b96	interaction of born in 1996 and weighted measure
wfraction_mindistanceC_1m_G_b97	interaction of born in 1997 and weighted measure
wfraction_mindistanceC_1m_G_b98	interaction of born in 1998 and weighted measure
wfraction_mindistanceC_1m_G_b99	interaction of born in 1999 and weighted measure
wfraction_mindistanceU_1m_G	Weighted Fraction of Points within 1 m of Uncontaminated Well
wmean_mindistanceU_G	Weighted Average Distance to Closest Uncontaminated Well

Data dictionary for HouseholdArsenic_2004.dta:

Variable name	Brief description
Source: BDHS	
contaminatedorsurface	household's water is contaminated with arsenic or from surface source
contaminatedwell	household's water is contaminated with arsenic
dhsid	id number of household
dhsid04	cluster number
dishwashingsource	water source used for dishwashing
heardofarsenic	whether heard of arsenic contamination problem
hv003	respondent's line number
hv005	sample weight
numarsenictest	mean of range of 2004 DHS household arsenic test
paintedwell	household's water source is a well that is painted any color
redwell	household's water source is a well painted red
shdistr	district
sourcetestedwater	type of water source household says uses for drinking
surfacewaterD	whether uses surface water for drinking
surfacewaterDW	whether uses surface water for dishwashing
v002	household number
wellmarking	marking on household's tubewell
year	year
Source: BGS matched with BDHS cluster	
as_walkable5_g	average arsenic level of wells in 5 mi
fraction_mindistanceC_1m_G	probability of living within 1 mi of contaminated Well
fraction_mindistanceU_1m_G	Fraction of Points within 1 m of Uncontaminated Well
fractioncont5	fraction of wells contaminated in 5 mi
mean_mindistanceU_G	Average Distance to Closest Uncontaminated Well
numcontaminated5	number of contaminated wells in 5 mi
numwalkable5	number of wells in 5 mi
wfraction_mindistanceC_1m_G	weighted probability of living within 1 mi of contaminated Well
wfraction_mindistanceU_1m_G	Weighted Fraction of Points within 1 m of Uncontaminated Well
wmean_mindistanceU_G	Weighted Average Distance to Closest Uncontaminated Well