DATA 621—Final Project

Critical Thinking Group 2

December 8, 2019

Abstract

Nearly a billion people lack access to clean drinking water (World Health Organization 2019). There are many well-known solutions to this problem, but most of them are too expensive to work in the areas hardest-hit. Providing affected areas better information about their water is cheap—but how effective is it?

To answer this question, we examine a dataset collected in rural Bangledash. It marks whether a household switched wells after learning their routine well had unsafe levels of arsenic.

After fitting several statistical models, we find that . . . [inferences]

Introduction

Perhaps the greatest public health crisis in the world remains access to clean drinking water and proper sanitation. Billionaire and philanthropist Bill Gates regards it as so serious, he spent millions of dollars holding a 'Reinvent the Toilet' challenge (Bill and Melinda Gates Foundation 2012).

The central hurdle, however, is not scientific, so much as *economic*. The developing nations that suffer the most from lack of clean water often have the least resources to deal with it. In many cases, solutions imported from developed nations—e.g., industrial water treatment plants—are simply too expensive. Even the winning solutions from the Gates Foundation's reinvented toilets remain too expensive to be practically implemented on a large scale.

Transmitting information is far less expensive than other proposed solutions. But can providing affected households information about their unsafe drinking water really help mitigate the water crisis? Are households able to change water supply, even when it comes with costs?

Literature Review

As of 2017, 29 percent of the world lacked accessed to safe and managed drinking water that is clean, located on premises, and available regularly. Contamination is a massive obstacle to raising this number. Through diarrhoea, drinking contaminated water kills almost half a million people each year (World Health Organization 2019).

The largest case of ground water contamination was discovered in Bangladesh in the early 1990s. Throughout the second half of the twentieth century, the government, humanitarian NGOs, and the private sector attempted to solve the country's water supply issues by mass installing *tube wells* throughout the country. Typically five centimenters in diameter, these tubes are inserted into the ground to depths less than 200 meters. Water is brought to the surface via a hand pump. In 1997, UNICEF announced it had surpassed its Millenium goal to provide 80 percent of Bangladesh with 'safe' drinking water thanks to these tube wells (van Geen, et al., 2002).

Tragically, research in the 1990s slowly uncovered that up to 77 million were drinking from tube wells contaminated with arsenic—half the population of Bangladesh. Arsenic consumption results in cancer, painful skin lesions, and other awful illnesses. The World Health Organization (WHO) considers water with a

concentration higher than 10 micrograms/liter as dangerous. Studies estimate that 10 percent of people that consume water with 500 micrograms/liter of arsenic will likely die from its effects (van Geen, et al., 2002).

Several studies have documented the extent of arsenic poisoning in Bangladesh. A survey conducted in the mid-1990s examined 1630 residents of affected regions. They found that 57.5 percent suffered from skin lesions associated with toxic levels of arsenic (Dhar, et al., 1997). Another study examined 7264 patients, finding that a full one-third suffered from the same kind of skin lesion (Biswas, et al., 1999).

It is not an overstatement to say this is a crisis that dwarfs the Chernobyl incident, or really any other nuclear accident in history.

There is one bright side, however. A study in the Araihazar upazila district found that the distribution of arsenic in groundwater is 'spatially highly variable.' This means it is not the case that excessive arsenic is concentrated in large regions. Instead, it is often that case that a contaminated well will be very near a safe well. Indeed, van Geen and his coauthors found about 90 percent of residents lived within 100 meters of a safe well (van Geen, et al., 2002).

This fact suggests a quick solution to Bangladesh's water problem: Find the poisoned wells and get residents to switch to a safe water supply that is likely nearby. Poisoned wells can be readily identified with cheap field kits. van Geen, et al., consider the 'real problem' to be convincing residents to switch to the safer wells. In their paper, they conclude 'social barriers to well-switching need to be better understood and, if possible, overcome.'

Researchers set about doing just that. Schoenfeld (2005) likewise confirmed that well switching was influenced by 'less predictable factors,' that interacted with physical variables (distance to nearest safe well, etc.). Social barriers could influence residents to not switch, even after being informed of the health risk of arsenic poisoning. On the other hand, a village 'arsenic activist' could persuade even those far from a safe well to switch.

Another study (Opar, et al., 2007) of the effect of information on well switching determined that 'the response to information is large and rapid,' provided residents were given *well-specific* information. Mass media campaigns were found to be mostly ineffective (and too expensive), while door-to-door campaigns had a positive effect.

Methodology

Our dataset is derived from Madajewicz, et al. (2007; also avaiable for R in the `carData` package as `Wells`). We propose to investigate how social factors and distance to nearest well affect the likelihood of a household to switch from a poisoned to safe well. A sample of its contents should familiarize the reader with its structure:

```
##   switch arsenic    dist assoc educ
## 1      1    2.36 16.826     0    0
## 2      1    0.71 47.322     0    0
## 3      0    2.07 20.967     0   10
## 4      1    1.15 21.486     0   12
## 5      1    1.10 40.874     1   14
## 6      1    3.90 69.518     1    9
```

It contains 3020 observations.

Our dependent variable is `switch`: coded as 0 if the family does not switch their water source after being informed that it is poisoned, and as 1 if they move to a different well. We hope to predict propensity to switch using these independent variables:

- `arsenic`: Hundreds of micrograms per liter of arsenic detected in a household's original well. Above 0.5 is considered unsafe.

- `distance`: Meters to the nearest safe well.

- `education`: Years of education of the head of household.

- `association`: Dichotomous variable, marking whether any of the members of the household engage in community or civic organizations.

We hypothesize that, theoretically speaking,

- `arsenic` has a *positive* relationship with `switch`. The more poisoned a well is, the more likely a family is to seek alternatives.

- `distance` is *negatively* related to `switch`. If using an alternative well is too inconveniant, households are less likely to make a change.

- Higher `education` education *increases* the propensity for families to switch.

- Higher `association` *increases* households' probability of switching to safer wells.

Statistical modeling is the chief activity of this paper. We seek to develop a robust model that elucidates the relationship between these independent variables and `switch`.

Logistic regression is the appropriate modeling strategy, as the dependent variable `switch` takes either `0` or `1` as its value. We strongly suspect some of these variables have interaction effects.

To ensure that our model does not overfit the data, we use cross validation. Models are trained on a majority of the dataset, but a smaller portion is held back. This test set will not be examined in data exploration, or be exposed to the models at all. This allows us to compare the models' predictions for the test set with reality, providing an unbiased estimate of model performance.

Of course, performance on the test set needs to be quantified. We propose using the F1 score, frequently used in classification for its ability to balance precision and recall

Even though our winning model will be decided based on its F1 score on the test set, we still report and concern ourselves with the other measures of performance, on both train and test sets. These will include Nagelkerke's R^2, deviance based psuedo-R^2, and precision/recall.

During the modeling process, we take care to conduct a thorough analysis of the errors, or *residuals*. Residuals can be tricky with logistic regression, so we propose three alternative methods of diagnostics:

1. *Hosmer-Lemeshow test*: Available in the `ResourceSelection` package (the `hoslem.test` function), this test bins the sample into g groups, and compares the expected and observed proportion of successes in each bin. For a well-fit model, the expected and observed proportions of success will be about equal for each bin.

2. *Binned residuals*: Similar the the HL test, this procedure (via `performance::binned_residuals`) is based on binning residuals. From there, the idea is the same as normal regression: There should be no pattern in the residuals.

3. *Quantile residuals*: Via the `statmod::qresid` package, this is an alternative to deviance and Pearson residuals specifically designed for generalized linear models (GLMs). A model's quantile residuals are statistically guaranteed to have an approximately normal shape if the model is well-fit. (It is unclear to us how useful they are with logistic regression, but they will be explored.)

Outliers and leverage will also be checked to ensure a good fit.

Finally, once the winning model has been ascertained, inferences and conclusions will be drawn.

References

- Bill and Melinda Gates Foundation. 2012. 'Bill Gates Names Winners of the Reinvent the Toilet Challenge.' Press release. https://www.gatesfoundation.org/media-center/press-releases/2012/08/bill-gates-names-winners-of-the-reinvent-the-toilet-challenge/.

- Biswas, B.K., U.K. Chowdhury, R.K. Dhar, B., et al. 1999. 'Groundwater arsenic contamination and sufferings of people in Bangladesh, a report up to January 1999.' In *International Conference, Arsenic in Bangladesh Ground Water: World's Greatest Arsenic Calamity*, Staten Island, New York.

- Dhar, Ratan Kr, Bhajan Kr Biswas, Gautam Samanta, et al. 1997. 'Groundwater arsenic calamity in Bangladesh.' *Current Science* vol. 73, no. 1: 48–59.

- Madajewicz, Malgosia, Alexander Pfaff, Alexander van Geen, et al. 2007. 'Can information alone both improve awareness and change behavior? Arsenic contamination of groundwater in Bangladesh.' *Journal of Development Economics* vol. 84, no. 2: 731–54. Draft available from https://www.ldeo.columbia.edu/~avangeen/publications/documents/Madajewicz_JDE_inpress.pdf.

- Opar, Alisa, Alex Pfaff, A.A. Seddique, et al. 2007. 'Responses of 6500 households to arsenic mitigation in Araihazar, Bangladesh.' *Health & Place* vol. 13, no. 1: 164–72.

- Schoenfeld, Amy. 2005. 'Area, village, and household response to arsenic testing and labeling of tubewells in Araihazar, Bangladesh.' New York City: Columbia University. Available at https://www.ldeo.columbia.edu/~avangeen/arsenic/documents/Schoenfeld_MS_05.pdf.

- van Geen, Alexander, Habibul Ahsan, Allan H. Horneman, et al. 2002. 'Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh.' *Bulletin of the World Health Organization* no. 80: 732-737.

- van Geen, Alexander, M. Trevisani, J. Immel, et al. 2006. 'Targeting Low-arsenic Groundwater with Mobile-phone Technology in Araihazar, Bangladesh.' *Journal of Health, Population, and Nutrition* vol. 24, no. 3: 282–97. Available at https://www.ldeo.columbia.edu/~avangeen/publications/documents/vanGeen_JHPN_06_000.pdf.

- van Geen, Alexander. 2018. 'Q&A With Lex Van Geen on Arsenic Contamination.' Interview by Peter Debaere. *UVA Darden Global Water Blog.* March 1. https://blogs.darden.virginia.edu/globalwater/2018/03/01/qa-with-lex-van-geen/.

- World Health Organization. 2019. 'Drinking water fact sheet.' June 14. https://www.who.int/news-room/fact-sheets/detail/drinking-water/.