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

CLEAN WATER GENERALLY

BANGLADEHS WATER

EFFECT OF BANGLESH WATER

EFFORTS TO COUNTER: government's mass testig (cell phone); also "Similar to other regions of Bangladesh and West Bengal, India, the distribution of arsenic in Araihazar is spatially highly variable (range: 5*860 mg/l) and therefore difficult to predict. Because of this variability, however, close to 90% of the inhabitants live within 100 m of a safe well. " (from first article)

TWO PREVIOUS VAN NUEM ARTICLES

** First article: https://www.scielosp.org/article/bwho/2002.v80n9/732-737/ **

** Second article: https://www.ldeo.columbia.edu/~avangeen/publications/documents/Madajewicz_JDE_inpress.pdf **

https://www.who.int/bulletin/archives/78%289%291093.pdf – nature 1998 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3754452/

https://www.nature.com/articles/26387

https://www.scielosp.org/scielo.php?pid=S0042-96862003000900005&script=sci_arttext&tlng=en

https://www.scielosp.org/scielo.php?pid=S0042-96862012001100013&script=sci_arttext&tlng=pt

https://www.bmj.com/content/342/bmj.d2431.full

Methodology

We propose to investigate this hypothesis using the dataset from van Green, et al. 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 zero if the family does not switch their water source after being informed that it is poisoned, and as one 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 explored in this paper. We seek to develop a robust model that elucidates the relationship between these independent variables and `switch`.

The dependent variable `switch` takes either `0` or `1` as its value. Thus logistic regression is the appropriate model. We strongly suspect some of these variables have interaction effects, so we will test them here.

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 AIC.

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 the same, 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 if 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/.
- 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.
- van Green, A., 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 Green, A. 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/.