DATA 621—Final Project

Critical Thinking Group 2

November 20, 2019

Contents

Executive Overview

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 ... advising households can be an effective and cheap solution, provided alternatives are available ...

Introduction and Goals

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 developed 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.

Information, however, 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?

Researchers collected data to test this hypothesis (Madajewicz, Pfaff, van Geen, et al. 2007). They studied the rural area of Arahazar, Bangladesh, which relies on wells for its water source. Studies in the early 1990s revealed abnormally high levels of naturally occuring arsenic in Bangladeshi wells. Further studies estimate that chronic arsenic exposure is responsible for up to 5 percent of mortality in these regions (van Green 2018). Part of the government's response, with the aid of universities and NGOs, includes mass well testing. Thus reliable information at a granular level of individual wells is available. It is a good test case for the impact of informing households of their water quality.

Researchers returned after several years to mark whether households had indeed switched wells. They found that even when individual households and villages are advised which wells are unsafe to drink, only about half switch their consumption to safer wells (van Green, et al., 2006, 283). The data set contains scientific factors like the level of arsenic contamination in the usual well, as well as social factors like level of education and participation in the community. Learning more about what factors prevent households from switching could improve this technique even further—perhaps by better allocation of resources—making it an even cheaper solution to this grave crisis.

The rest of this paper is organized as follows. After exploring the data and noting any issues it has, we develop a number of statistical models to try to explain why some households fail to switch to safer wells. To validate the models, we have split the data set into a training and test set (70/30). Each trained modeled is applied to the hold out set. Inferences are made off of the best model, and conclusions are drawn.

Data Exploration

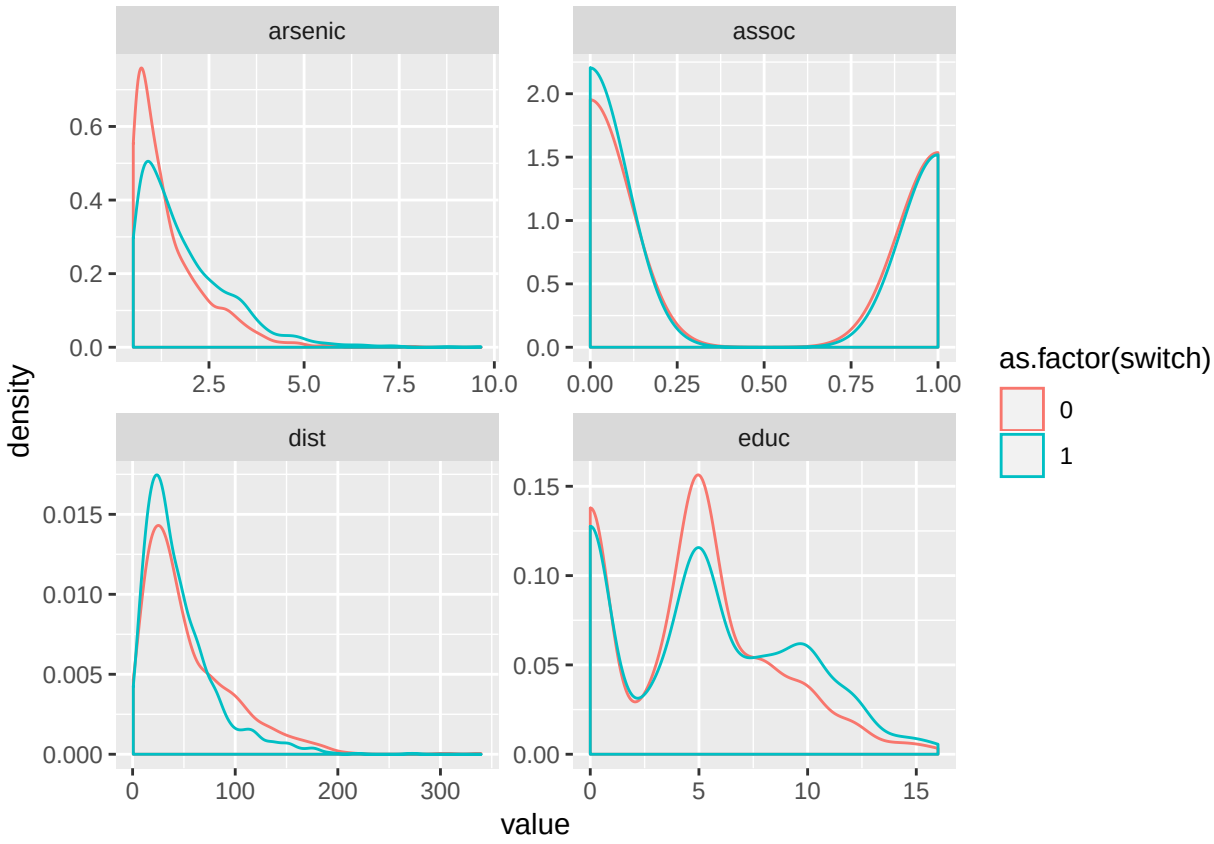
Load data and split into test and train:

```
## [1] "Rows in training data set: 2114"
```

```
## [1] "Rows in test data set: 906"
```

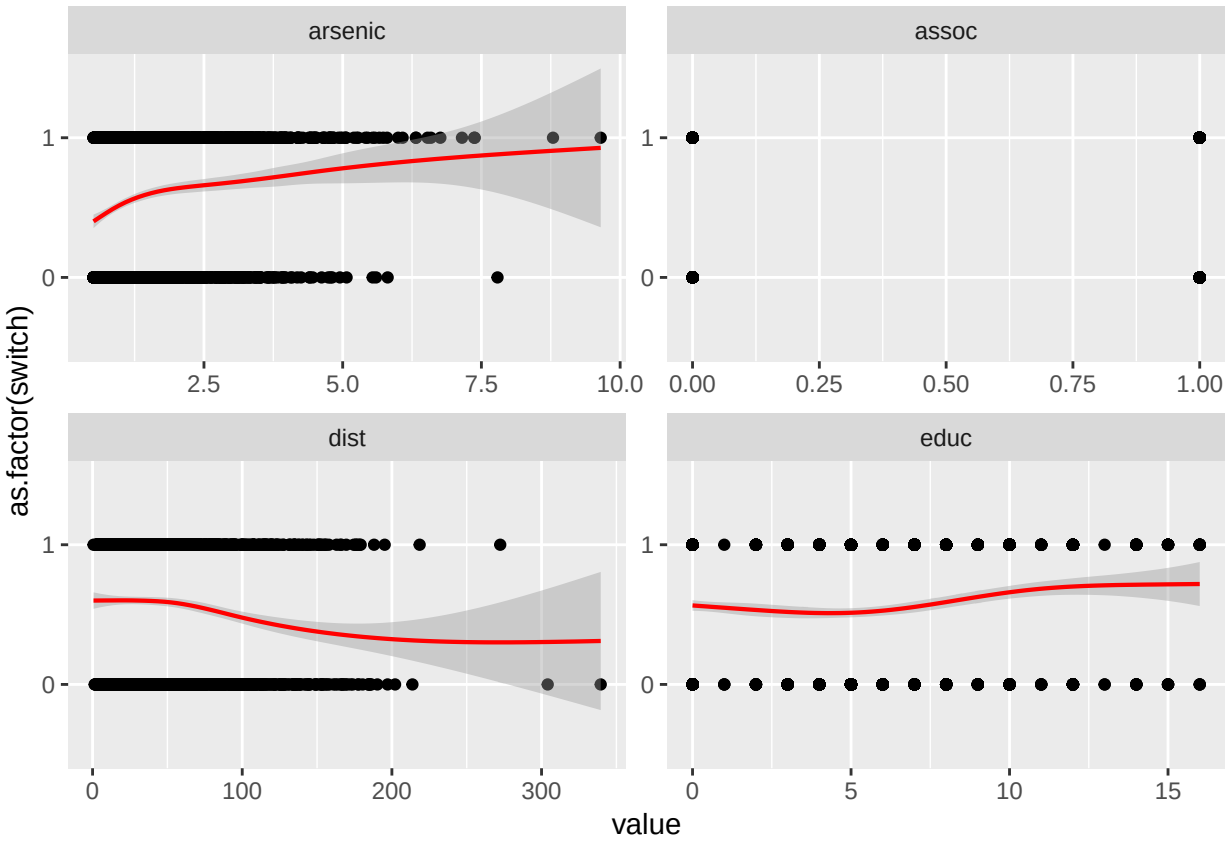
TODO describe this more density plots by `switch`

```
train %>%
  gather(-switch, key='variable', value='value') %>%
  ggplot(aes(x=value, group=as.factor(switch), color=as.factor(switch))) +
    facet_wrap(~ variable, scales='free') +
    geom_density()
```



** TODO describe this more** relationship between vars

```
train %>%
  gather(-switch, key='variable', value='value') %>%
  ggplot(aes(x=value, y=as.factor(switch), group=1)) +
    geom_point() +
    geom_smooth(color='red', size=.75) +
    facet_wrap(~ variable, scales='free')
```



Data Preparation

Modeling

blah balh

M_0: Dummy model

Dummy model predicting class proportion

lm(y ~ 1, df)

M_1: Full model

Use all variables, no interactions or polynomials or anything interesting

M_2:

Most interesting models

M_i: Other models

whatever makes sense

Evaluating the Models on the Test Set

Evaluate each model based on its $F1$ score on the test set, although we include other metrics as a convenianceS:

Table of results (fill in):

	Description	F1	Accuracy	Sensitivity	Specificity
M_0	dummy	0.6966	0.53448	1.0000	0.0000
M_1	full	0.9076	0.8965	**0.9516**	0.8333
M_2	all + interactions	**0.9344**	**0.9310**	0.9193	**0.9444**
M_3	parred down + interactions	0.8615	0.8448	0.9032	0.7777
M_4	PCA	0.8062	0.7844	0.8387	0.7222

Conclusion

concluding remarks and evaluate best model for learnings

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/.