DATA 621—Final Project

Critical Thinking Group 2

December 20, 2019

Contents

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 Bangladesh (Opar, et al., 2007). It marks whether a household switched wells after learning their routine well had unsafe levels of arsenic. After examining based on the data set we have realized that the majority of the population have switched their sources of drinking water motivated by the health concerns of the polluted water. Our final model uses Probit regression with interaction terms, boasting an accuracy of 0.63 and an F1 score of 0.73.

Keywords: Water contamination, arsenic poisoning, Bangladesh, developmental economics

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 prize winning design from the Gates competition remains too expensive to be practically implemented on a large scale.

Public education about the danger of poisoned water sources 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 drinking water that is clean, located on premises, and available regularly. Contamination is one of the most significant obstacle to raising this number, killing almost half a million people each year (World Health Organization 2019).

The largest instance 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 centimeters 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 Millennium 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 many of these new wells were contaminated with arsenic. It was estimated that up the 77 million people were affected—half the population of Bangladesh. Arsenic consumption results in cancer, painful skin lesions, and other disease. 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).

Although the World Health Organization considers water with concentration higher than 10 micrograms/liter as dangerous, the arsenic concentration used to define unsafe drinking water in the data set is based on the Bangladesh standard of 50 microgram per liter. All the households in the data set have original wells with arsenic levels above the Bangladesh standard of 50 microgram per liter. So, these are all affected households. The Bangladesh Arsenic Mitigation and water Supply Program (BAMWSP) coordinated a blanket survey of million tubewells. This survey generated nearly five million field-kit results of well-testing, which identified wells as safe or unsafe. Household response surveys in the area of Araihazar upazila (administrative region) indicate roughly half the affected households switched to safe wells. However, the survey also showed that a significant number of households did not stop drinking from unsafe wells after they had learned that it was unsafe (Van Geen, et al., 2006).

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). Other research investigated children's intellectual function after exposure to arsenic in Bangladesh. The study found that exposure to arsenic in drinking water was associated with reduced scores on measures of intellectual function, before and after adjusting for sociodemographic features known to contribute to intellectual function (Wassermanm et al., 2004).

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 that it is oftne possible to find a clean and safe well only a short distance from contaminated wells. Indeed, van Geen and his coauthors found about 90 percent of residents in the area under study 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 safer nearby water supply. 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 earlier 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,' thought to interact with physical variables (distance to nearest safe well, etc.). Social barriers could prevent residents from switching, 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.

Most of the cited research is primarily concerned with conducting surveys and 'simple' analysis of their results. The most statistically 'sophisticated' work in this literature is Gelman, et al. (2004), and Opar, et al. (2007).

Opar and his colleagues returned to Araihazar upazila several years after initial education efforts had been conducted; the data in this present paper is the result of their studies. They examined the effects of these efforts, which included public education, directly posting arsenic poisoning test results onto the wells themselves, and installing community wells. A Probit regression estimated the relationships between well switching and the following independent variables: water arsenic content, distance to nearest safest well, years of education, and 'easily observable proxies for income and wealth.' These variables were found to be significant, except the income and wealth-related variables. The Probit regression had a pseudo-R^2 of 0.29.

Although Gelman, et al. (2004) use an earlier version of this data, their primary concern seems to be using it as a demonstration of Bayesian decision analysis. The authors make a point to avoid parametric methods, including regression. Instead they rely on k-means clustering and *a priori* probability models to answer: How effective is encouraging villagers to switch to alternative, non-poisoned wells? Where should new (safe) wells be located to maximize their availability? How deep should new wells be drilled? They conclude that recommending new wells results reduces average arsenic exposure by 38 percent.

Methodology

The `Wells` data set is loaded from 'http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat'.

This data set has 3020 rows and only contains complete cases. The data set is split into train and test sets (70 and 30 percent, respectively).

This data set contains five variables. Variable `switch` is the binary response variable. The `arsenic` levels in all the observations are all **above** 0.5 (hundreds of micrograms per liter), which is above the Bangladesh safe standard of 50 micrograms per liter. `distance` ranges from less than 1 meter to approximately 340 meters (~ 0.2 miles) away from the closest known safe well. This range is within walking distance. The number of years in `education` of the head of household ranges from 0 (no education) to 16 (comparable to master's level) years.

Variable	Description
`switch`	Dependent binary variable. Describes whether family switched (1) or not (0).
`distance`	Distance (in meters) to the closest known safe well.
`education`	Number of years in education of the head of the household.
`association`	Describes whether or not any members of the household participated in any community organizations: no or yes.

Hypothesized theoretical effects of:

- `arsenic` on `switch` is positive. The higher the arsenic level is, the more likely households will switch.
- `distance` on `switch` is negative. The farther the safe well is, the less likely households will switch.
- `education` on `switch` is positive. The more education a household has, the more likely the household will switch.
- `association` on `switch` is positive. Households active in community organizations should be more likely to switch.

This project will investigate the relationship of each of the explanatory variables with the response variables and model the probability of households switching. The predicted probability of switching should help interested community organizations identify households that are high-risk for not switching. This should help community leaders focus on high-risk households or villages that may require more attention and resources.

Probit and logit regression models are going to be built and evaluated. Model selection is going to be based on which model has the minimum AIC.

In probit regression, the cumulative standard normal distribution function is used to model the regression function when the dependent variable is binary. In logit regression, the cumulative distribution function is used. Both types of regression are appropriate for modeling binary response variables.

Experimentation and Results

Data Exploration

Appendix A contains a high-level summary of the dataset. As you can see, all the observations in the data set are complete cases.

Plots of the density/distribution of each of the variable are also available in Appendix A. There are more households that switched. The majority of the original household wells had arsenic levels less than 2.5 (hundreds of micrograms per liter). Most of households are less than 100 meters away from the nearest safe well. The households can be roughly grouped into those with less than 2.5 years of education, those with 2.5 to 7.5 years of education, and those with more than 7.5 years of education. There are less households with associations to the community.

We also examined the associations between each variable (see Appendix B for plotting). Overall, households with higher arsenic levels tend to switch more. Households that are closer to nearest safe wells tend to switch more. Households with higher education tend to switch more. Interestingly, households with no associations to the community have a greater proportion of switching at 59 percent compared to 56 percent for those with associations to the community. The observed relationships support the theoretical relationships discussed above except for the household's association with the community. We expected household's with associations to the community to have a greater proportion of switching. However, the difference between 56 percent and 59 percent is not that big, and this difference may not be significant.

```
## [1] "Percent of families with association to community that switched: 0.56"
## [1] "Percent of families without association to community that switched: 0.59"
```

We examined how the distribution of `switch` varied with the independent variables (density plots in Appendix B).

Below is a density plot of each explanatory variable grouped by `switch`. The arsenic density plot shows that there is an arsenic threshold where more households tend to switch. The density plot for distance shows that households that are closer to safe wells have a higher proportion of switching. There is a threshold in distance where we start to see less households switching. This is roughly around 75 meters. The density plot for education shows that there is an education threshold (around 6 years) where we see a higher proportion of switching.

Correlation provides a single point estimate of the relationship betwen the variables. Correlation table and plot are available in Appendix C. They show a weak, positive correlation between arsenic and switch. None of the explanatory variables appear to be strongly correlated to each other. There is a negative correlation between switch and distance. There is an overall negative correlation between switch and association, and a positive correlation between education and switch.

‘

Modeling

Models	Pseudo R^2	AIC
M1: Logit Full Model	R^2: 0.0864052	AIC: 2753.7
M2: StepAIC Logit Model with Interactions	R^2: 0.09112204	AIC: 2747.8
M3: Probit Full Model	R^2: 0.0855046	AIC; 2755.2
M4: StepAIC Probit Model with Interactions	R^2: 0.09036308	AIC; 2749.1

M_1: Logit Full Model

This is a binary regression that uses the *logit* function as the link function. All four explanatory variables are included in this model. The pseudo R^2 is 0.0864052. The AIC is 2753.7.

```
summary(m_1)
```

```
##
## Call:
## glm(formula = switch ~ arsenic + dist + educ + assoc, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5622  -1.1958   0.7635   1.0549   1.6608
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.244136   0.118976  -2.052   0.0402 *
## arsenic      0.465440   0.050715   9.178  < 2e-16 ***
## dist        -0.008123   0.001209  -6.717 1.85e-11 ***
## educ         0.052717   0.011485   4.590 4.43e-06 ***
## assoc1      -0.093370   0.091797  -1.017   0.3091
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2884.3  on 2114  degrees of freedom
## Residual deviance: 2743.7  on 2110  degrees of freedom
## AIC: 2753.7
##
## Number of Fisher Scoring iterations: 4
```

```
PseudoR2(m_1, which = 'Nagelkerke')
```

```
## Nagelkerke
##  0.0864052
```

M_2: StepAIC Logit Model with Interactions

This is a logit model generated through stepAIC that includes interactions among all the variables. The pseudo R^2 is 0.09112204. The AIC is 2747.8. Interactions *dist:educ* and *arsenic:educ* are included in the model.

```
m_2 <- stepAIC(m_1, trace=0, scope=list(upper = ~ arsenic * dist * educ * assoc, lower=~1))
```

```
summary(m_2)
```

```
##
## Call:
## glm(formula = switch ~ arsenic + dist + educ + dist:educ + arsenic:educ,
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4145  -1.2036   0.7542   1.0628   1.8217
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.0086505  0.1473773  -0.059   0.9532
## arsenic       0.3931632  0.0750992   5.235 1.65e-07 ***
## dist         -0.0115178  0.0020185  -5.706 1.16e-08 ***
## educ         -0.0062144  0.0239367  -0.260   0.7952
## dist:educ     0.0006449  0.0003063   2.105   0.0353 *
## arsenic:educ  0.0189268  0.0133402   1.419   0.1560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2884.3  on 2114  degrees of freedom
## Residual deviance: 2735.8  on 2109  degrees of freedom
## AIC: 2747.8
##
## Number of Fisher Scoring iterations: 4
```

```
PseudoR2(m_2, which = 'Nagelkerke')
```

```
## Nagelkerke
## 0.09112204
```

M_3: Probit Full Model

This is a binary regression that uses the *probit* function as the link function. All explanatory variables are included in this model. The pseudo R^2 is 0.0855046. The AIC is 2755.2.

```
m_3 <- glm(switch ~ arsenic + dist + educ + assoc, family = binomial(link="probit"), data=train)
```

```
summary(m_3)
```

```
##
## Call:
## glm(formula = switch ~ arsenic + dist + educ + assoc, family = binomial(link = "probit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6821  -1.1977   0.7702   1.0557   1.6515
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.1386937  0.0732087  -1.894   0.0582 .
## arsenic      0.2750260  0.0297432   9.247  < 2e-16 ***
## dist        -0.0049326  0.0007348  -6.713 1.91e-11 ***
## educ         0.0329979  0.0070351   4.690 2.73e-06 ***
## assoc1      -0.0601532  0.0565802  -1.063   0.2877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2884.3  on 2114  degrees of freedom
## Residual deviance: 2745.2  on 2110  degrees of freedom
## AIC: 2755.2
##
## Number of Fisher Scoring iterations: 4
```

```
PseudoR2(m_3, which = 'Nagelkerke')
```

```
## Nagelkerke
##  0.0855046
```

M_4: StepAIC Probit Model with Interactions

This is a probit model generated through stepAIC that includes interactions among all the variables. The pseudo R^2 is 0.09036308. The AIC is 2749.1. Interactions *dist:educ* and *arsenic:educ* are included in the model.

```
m_4 <- stepAIC(m_3, trace=0, scope=list(upper = ~ arsenic * dist * educ * assoc, lower=~1))
```

```
summary(m_4)
```

```
##
## Call:
## glm(formula = switch ~ arsenic + dist + educ + dist:educ + arsenic:educ,
##     family = binomial(link = "probit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4573  -1.2059   0.7621   1.0648   1.8171
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0091886  0.0908730   0.101   0.9195
## arsenic       0.2264689  0.0445455   5.084 3.70e-07 ***
## dist         -0.0069238  0.0012097  -5.724 1.04e-08 ***
## educ         -0.0042956  0.0146704  -0.293   0.7697
## dist:educ     0.0003775  0.0001849   2.041   0.0412 *
## arsenic:educ  0.0126022  0.0078157   1.612   0.1069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2884.3  on 2114  degrees of freedom
## Residual deviance: 2737.1  on 2109  degrees of freedom
## AIC: 2749.1
##
## Number of Fisher Scoring iterations: 5
```

```
PseudoR2(m_4, which = 'Nagelkerke')
```

```
## Nagelkerke
## 0.09036308
```

Evaluation

Model M_4 (StepAIC Probit with Interactions) has the highest F1 score of 0.728466.

	Description	F1	Accuracy	Sensitivity	Specificity
M_1	Logit Full Model	0.7100592	0.621	0.8061	0.3698
M_2	StepAIC Logit + Interactions	0.727572	0.6343	0.8484	0.3438
M_3	Probit Full Model	0.7139241	0.6254	0.8119	0.3724
M_4	StepAIC Probit + Interactions	0.728466	0.6343	0.8522	0.3385

Predict `test` set based on each model.

```
predict1 <- factor(round(predict(m_1, test, type='response')), levels=c('0', '1'))
predict2 <- factor(round(predict(m_2, test, type='response')), levels=c('0', '1'))
predict3 <- factor(round(predict(m_3, test, type='response')), levels=c('0', '1'))
predict4 <- factor(round(predict(m_4, test, type='response')), levels=c('0', '1'))
```

Model 1:

The F1 score for model 1 is 0.7100592.

```
(cm1 <- caret::confusionMatrix(data=predict1,
               reference = factor(test$switch, levels=c('0', '1')), positive='1'))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 142 101
##          1 242 420
##
##                Accuracy : 0.621
##                  95% CI : (0.5885, 0.6527)
##     No Information Rate : 0.5757
##     P-Value [Acc > NIR] : 0.003104
##
##                   Kappa : 0.1849
##
##  Mcnemar's Test P-Value : 4.053e-14
##
##             Sensitivity : 0.8061
##             Specificity : 0.3698
##          Pos Pred Value : 0.6344
##          Neg Pred Value : 0.5844
##              Prevalence : 0.5757
##          Detection Rate : 0.4641
##    Detection Prevalence : 0.7315
##       Balanced Accuracy : 0.5880
##
##        'Positive' Class : 1
##
```

```
cm1$byClass['F1']
```

```
##        F1
## 0.7100592
```

Model 2:

The F1 score for model 2 is 0.727572.

```
(cm2 <- caret::confusionMatrix(data=predict2,
               reference = factor(test$switch, levels=c('0', '1')), positive='1'))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 132  79
##          1 252 442
##
##                Accuracy : 0.6343
##                  95% CI : (0.6019, 0.6657)
##     No Information Rate : 0.5757
##     P-Value [Acc > NIR] : 0.0001892
##
```

```
##                   Kappa : 0.2042
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8484
##             Specificity : 0.3438
##          Pos Pred Value : 0.6369
##          Neg Pred Value : 0.6256
##              Prevalence : 0.5757
##          Detection Rate : 0.4884
##    Detection Prevalence : 0.7669
##       Balanced Accuracy : 0.5961
##
##        'Positive' Class : 1
##
```

```
cm2$byClass['F1']
```

```
##       F1
## 0.727572
```

Model 3:

The F1 score for model 3 is 0.7139241.

```
(cm3 <- caret::confusionMatrix(data=predict3,
               reference = factor(test$switch, levels=c('0', '1')), positive='1'))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 143  98
##          1 241 423
##
##                Accuracy : 0.6254
##                  95% CI : (0.593, 0.657)
##     No Information Rate : 0.5757
##     P-Value [Acc > NIR] : 0.001311
##
##                   Kappa : 0.1938
##
##  Mcnemar's Test P-Value : 1.235e-14
##
##             Sensitivity : 0.8119
##             Specificity : 0.3724
##          Pos Pred Value : 0.6370
##          Neg Pred Value : 0.5934
##              Prevalence : 0.5757
##          Detection Rate : 0.4674
##    Detection Prevalence : 0.7337
##       Balanced Accuracy : 0.5921
##
##        'Positive' Class : 1
##
```

```
cm3$byClass['F1']
```

```
##        F1 
## 0.7139241 
```

Model 4:

The F1 score for model 4 is 0.728466.

```
(cm4 <- caret::confusionMatrix(data=predict4,
               reference = factor(test$switch, levels=c('0', '1')), positive='1'))
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 130  77
##          1 254 444
## 
##                Accuracy : 0.6343          
##                  95% CI : (0.6019, 0.6657)
##     No Information Rate : 0.5757          
##     P-Value [Acc > NIR] : 0.0001892       
## 
##                   Kappa : 0.2031          
## 
##  Mcnemar's Test P-Value : < 2.2e-16       
## 
##             Sensitivity : 0.8522          
##             Specificity : 0.3385          
##          Pos Pred Value : 0.6361          
##          Neg Pred Value : 0.6280          
##              Prevalence : 0.5757          
##          Detection Rate : 0.4906          
##    Detection Prevalence : 0.7713          
##       Balanced Accuracy : 0.5954          
## 
##        'Positive' Class : 1               
## 
```

```
cm4$byClass['F1']
```

```
##       F1 
## 0.728466 
```

Discussion and Conclusion

Using the winning model M_4, we can predict the probability of switching for each household in the `test` set. Given a list of households from several communities, we can use model M_4 to identify potential households that are high risk for not switching.

For the `test` set, we can categorize households into "Likely to switch", "Somewhat Likely to switch", and "Unlikely to Switch". In this example, households with probability of switching that's greater than 0.55 are categorized as likely to switch. Probability less than 0.45 are categorized as "Unlikely to Switch". The rest are categorized as “Somewhat Likely to Switch”. Perhaps local community leaders can use this prediction to help identify households predicted to be high-risk for not switching.

```
test$prob_switch <- predict(m_4, test, type='response')
test$category[test$prob_switch < .45] = 'Unlikely to Switch'
test$category[test$prob_switch >= .45 & test$prob_switch <= .55] = 'Somewhat Likely to Switch'
test$category[test$prob_switch > .55] = 'Likely to Switch'
```

Below shows 5 households that are unlikely to switch based on predictions made by model M_4.

	switch	arsenic	dist	assoc	educ	prob_switch	category
52	0	0.51	99.013	0	0	0.2874467	Unlikely to Switch
53	1	0.64	115.913	1	0	0.2583532	Unlikely to Switch
114	0	0.78	90.921	0	5	0.4034535	Unlikely to Switch
159	0	0.61	74.999	0	5	0.4154924	Unlikely to Switch
163	0	1.02	103.061	1	4	0.3883791	Unlikely to Switch

Based on the predicted probabilities, most of the households are likely to switch (see also Appendix D).

The models should be evaluated and validated on more data. Different models should be investigated as well to see if useful predictions can be with fewer explanatory variables. A simpler model is much more cost effective as it requies less data to use.

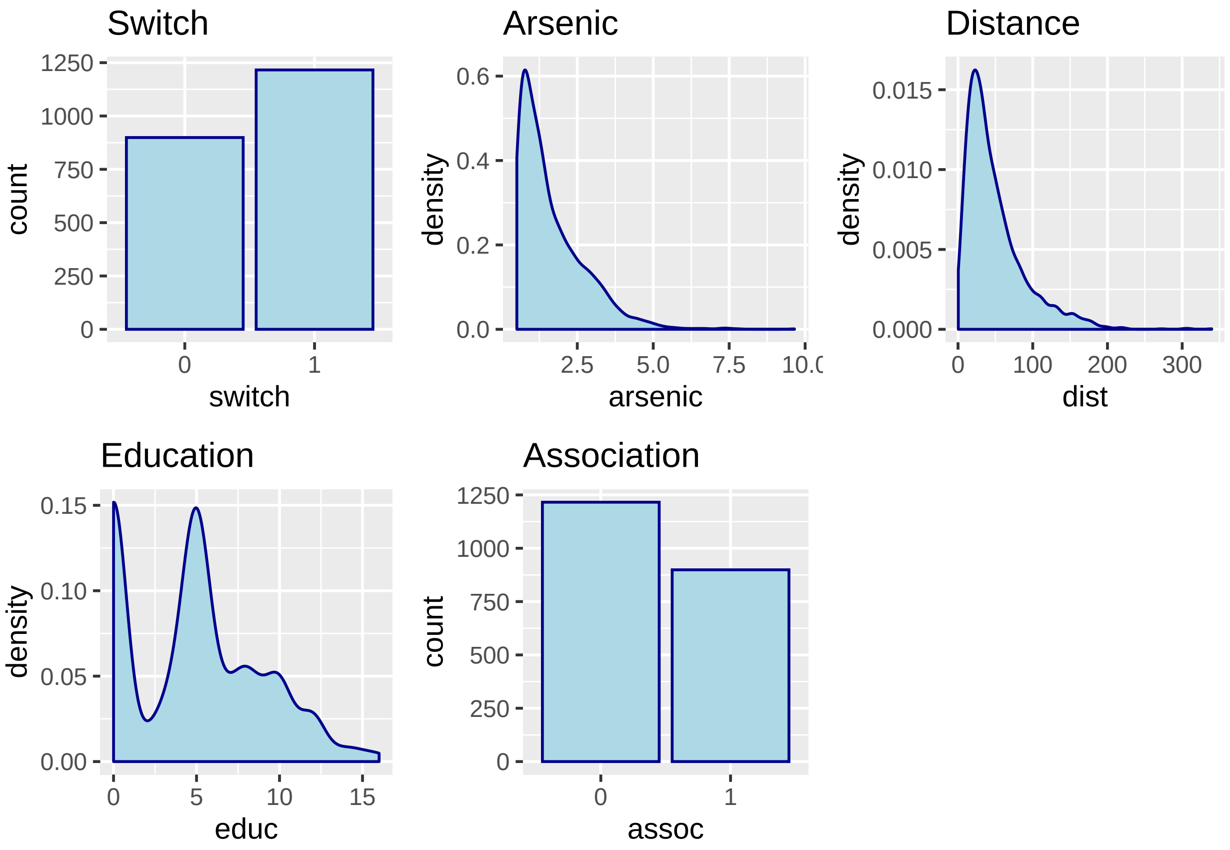
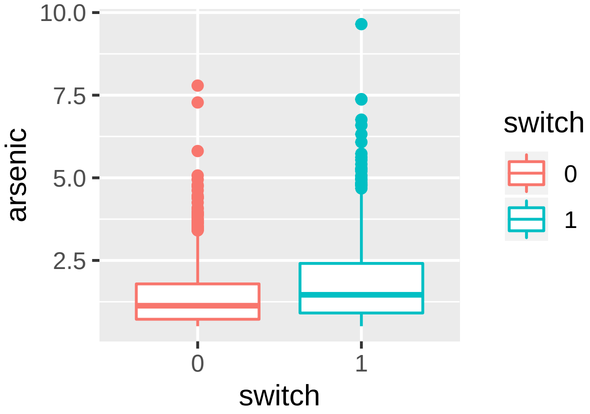


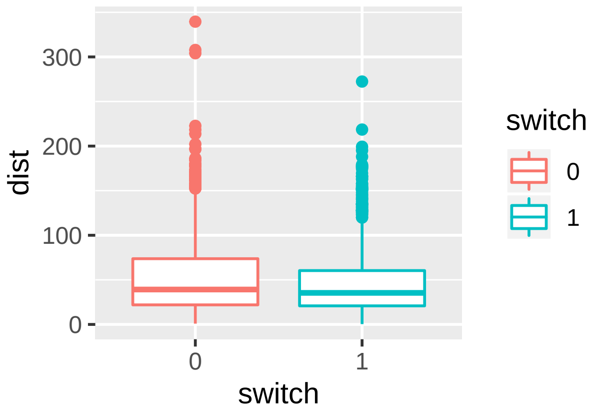
Figure 1: Density plots of the data set

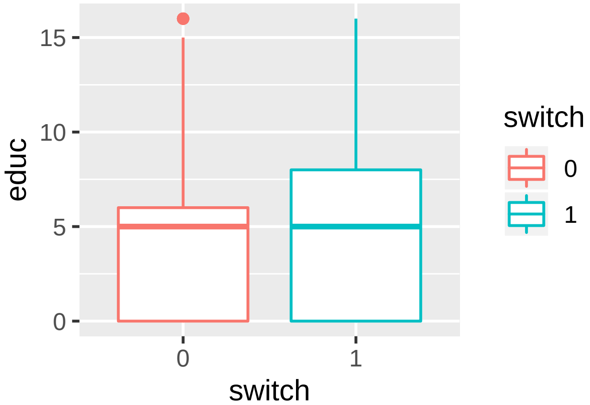
Appendix A: Summarizing of data

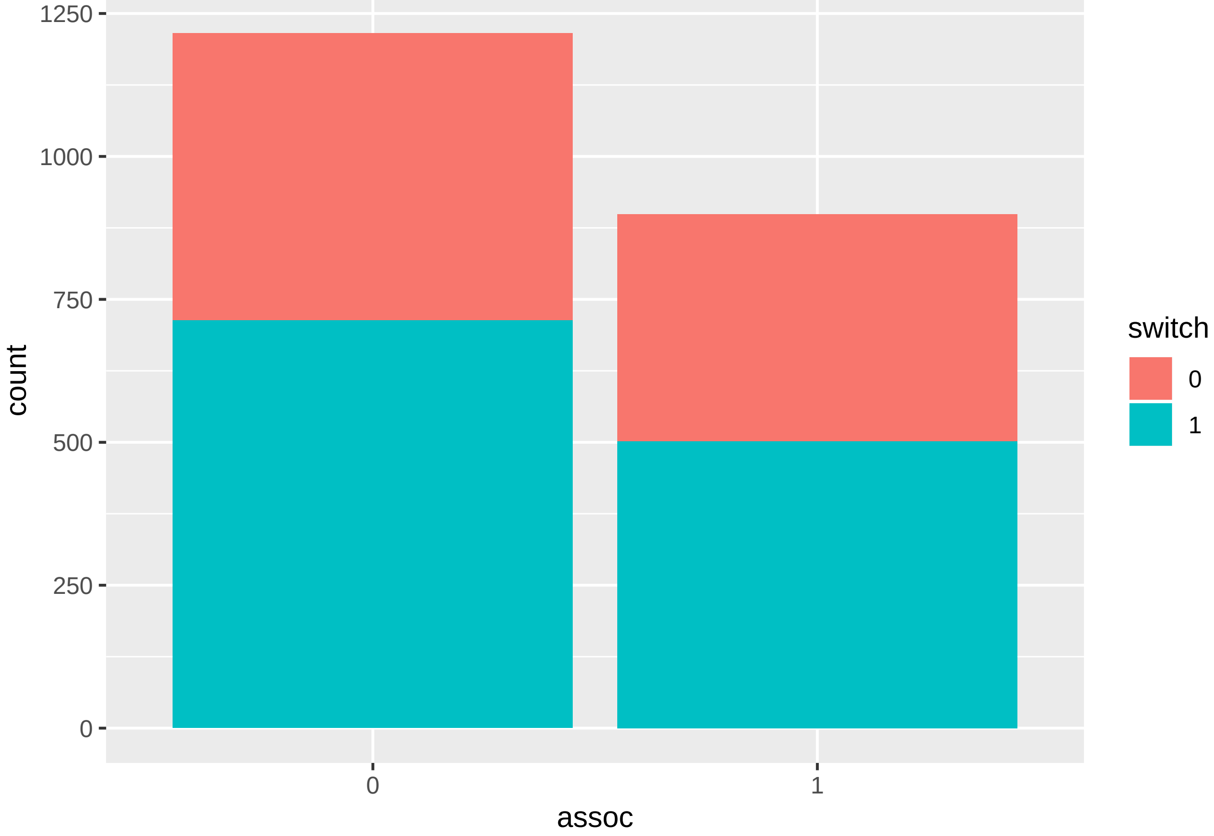
```
##  switch      arsenic           dist          assoc         educ
##  0: 899   Min.   :0.510   Min.   :  0.387   0:1216   Min.   : 0.000
##  1:1216   1st Qu.:0.820   1st Qu.: 21.318   1: 899   1st Qu.: 0.000
##            Median :1.290   Median : 36.875            Median : 5.000
##            Mean   :1.636   Mean   : 49.112            Mean   : 4.822
##            3rd Qu.:2.185   3rd Qu.: 64.057            3rd Qu.: 8.000
##            Max.   :9.650   Max.   :339.531            Max.   :16.000
```

Appendix B: Univariate associations









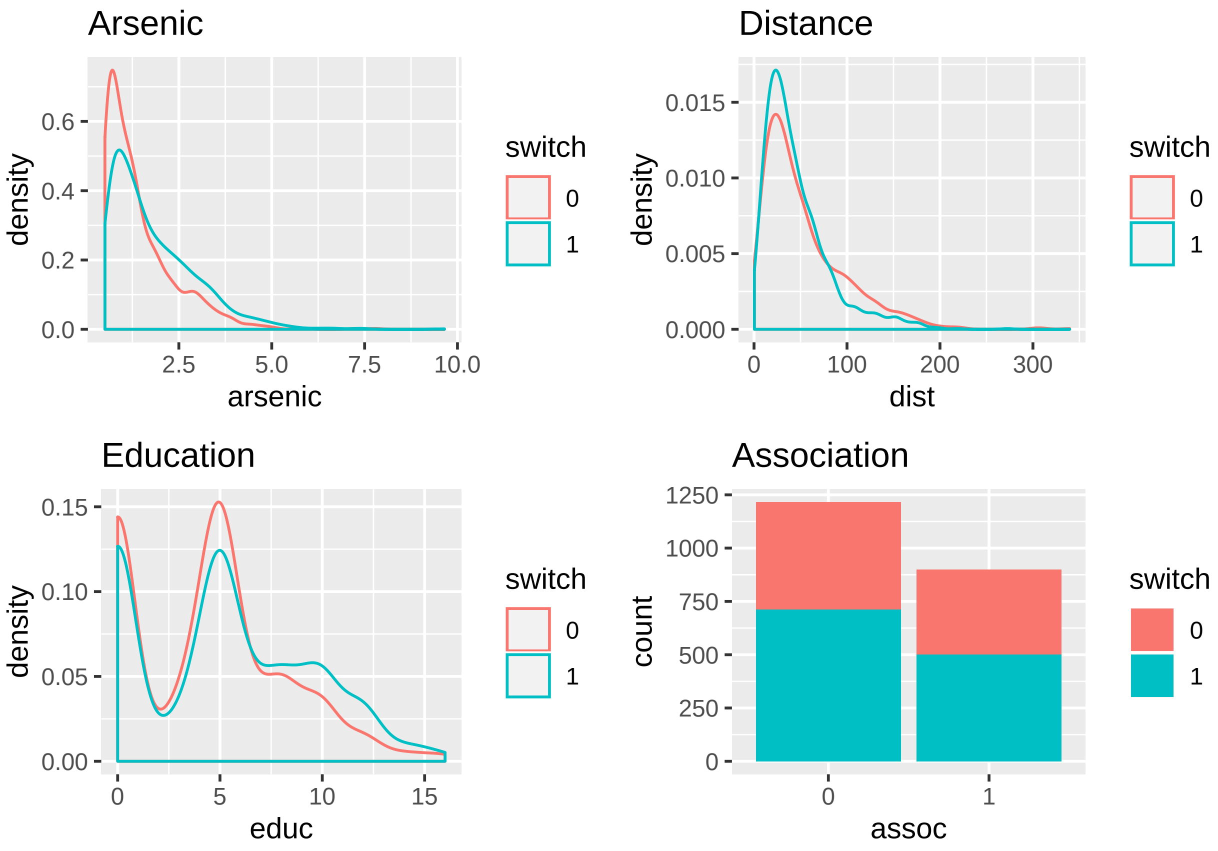


Figure 2: Distributions of switch with the independent variables.

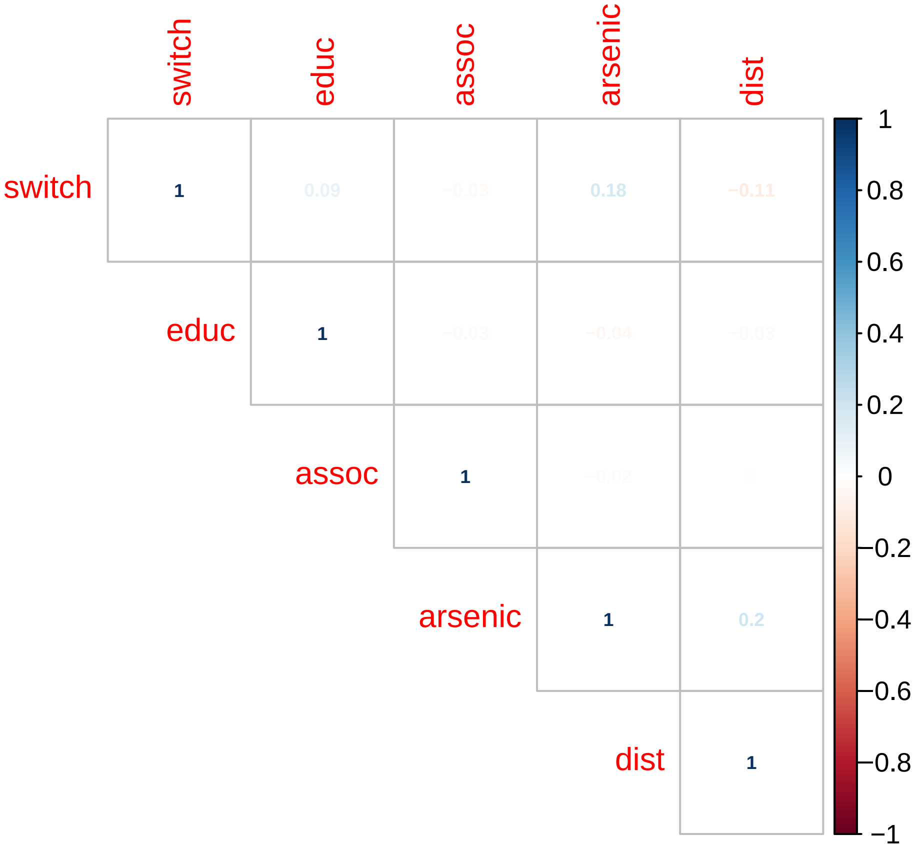


Figure 3: Plot: Variable correlations.

Appendix C: Correlation

	switch	arsenic	dist	assoc	educ
switch	1.0000	0.1763	-0.1058	-0.0288	0.0931
arsenic	0.1763	1.0000	0.1982	-0.0189	-0.0396
dist	-0.1058	0.1982	1.0000	-0.0008	-0.0261
assoc	-0.0288	-0.0189	-0.0008	1.0000	-0.0279
educ	0.0931	-0.0396	-0.0261	-0.0279	1.0000

Appendix D: Model predictions

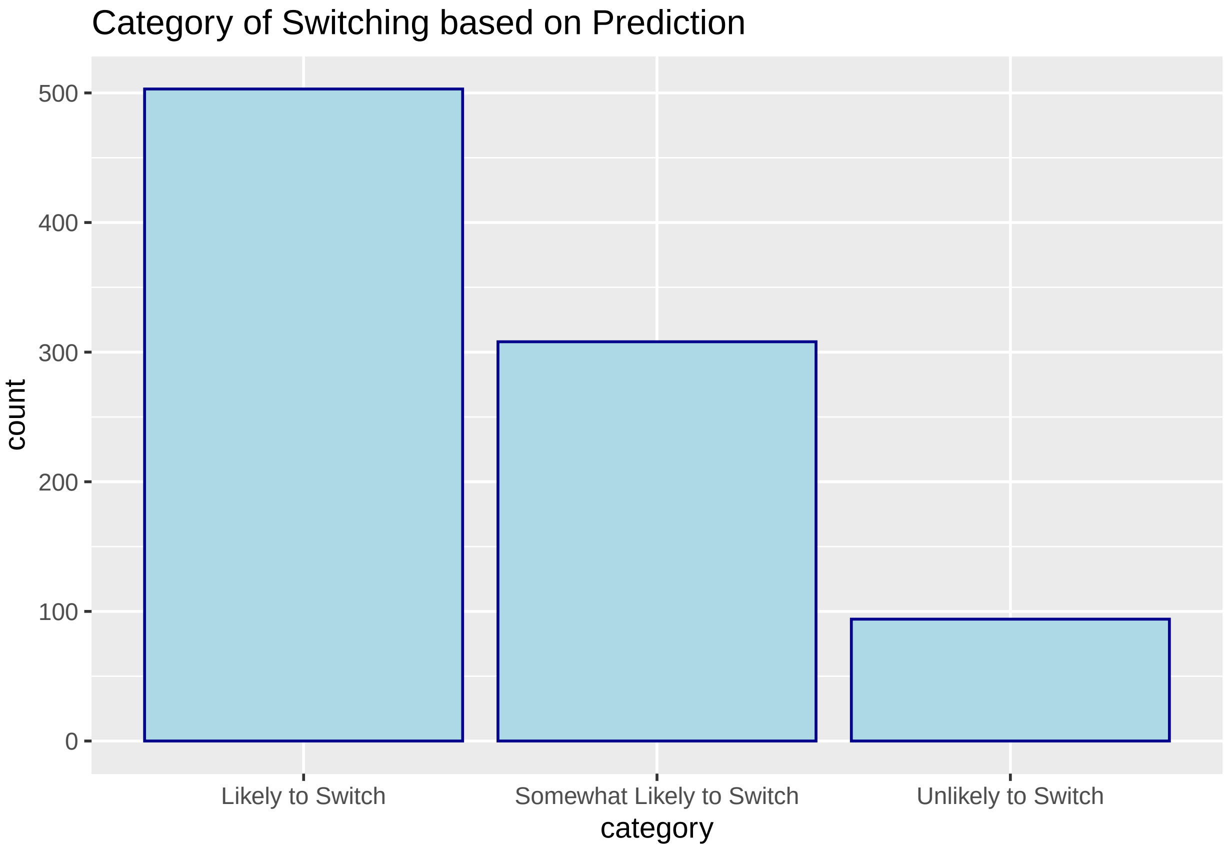


Figure 4: Model predictions.

References

- Bill and Melinda Gates Foundation. 2012. 'Bill Gates Names Winners of the Reinvent the Toilet Challenge.' Press release, August. 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.

- Gelman, Andrew, Matilde Trevisani, Hao Lu, and Alexander van Geen. 2004. 'Direct data manipulation for local decision analysis as applied to the problem of arsenic in drinking water from tube wells in Bangladesh.' *Risk Analysis* vol. 24, no. 6: 1597–1612. Available at https://www.ldeo.columbia.edu/~avangeen/pdf/Gelman_RiskAnal04.pdf.

- 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. Availab at http://www.academia.edu/download/45587532/Responses_of_6500_households_to_arsenic_20160512-23903-h8dy3v.pdf.

- 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. Available at https://www.ldeo.columbia.edu/~avangeen/pdf/vanGeen_WHO_02.pdf.

- 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/.

- Wasserman, Gail A., et al. "Water Arsenic Exposure and Children's Intellectual Function in Araihazar, Bangladesh." Environmental Health Perspectives, vol. 112, Sept. 2004, https://ehp.niehs.nih.gov/doi/full/10.1289/ehp.6964.