Capstone final report

ECOM30002/90002 Econometrics 2, Semester 2, 2024

Quantifying the impact of water insecurity on mental illness in Ghana

Josh Copeland (SID: 1444772), Jocelyn Koswara (SID: 1473005), Ryan Luo (SID: 1395525)

Abstract:

This report seeks to explore how water insecurity causally affects mental health outcomes. We find that basic water access decreases mental health disorder likeliness. Robustness checks and extensions are performed to support the validity of our use of IV and 2SLS. The findings of this report are useful for exploring the effects of water security that may not be as apparent as effects on physical health.

Word Count: 2029

1. Introduction

The WHO defines health as the “state of complete physical, mental and social well-being” (World Health Organisation, 1948, p. 1). Despite this, much of the literature on the health impact of water insecurity focuses on physical factors (Achore & Bisung 2022; Fink, 2011; Apanga et al., 2021).

To develop a more holistic illustration of how critical water security is for overall health outcomes, including mental health, this report will seek to answer the question: does water insecurity negatively impact mental health outcomes? Our hypothesis is that it does, similar to the literature focusing on the relationship between water access and physical health.

We find evidence of this being the case, and in doing so hope to emphasise the importance of water access for overall public health outcomes, and compound the imperative for governments to invest in such services.

2. Data

To address our research question, we use the World Bank's Socioeconomic Panel Survey: 2009-10 to produce the variables provided in Table 1.

Table 1: Summary statistics

Variables	*Mean*	*S.D*	*Minimum*	*Maximum*
Has mental health disorder	0.31	0.46	0	1
Has access to basic drinking services	0.76	0.43	0	1
Female	0.55	0.5	0	1
Age	39.09	18.73	1	109
Part of religious minority	0.34	0.47	0	1
Lives in rural area	0.65	0.48	0	1
Distance from drinking water sources (minutes)	15.64	18.02	0	240

Note: number of observations is 9282.

Our dependent variable is a dummy variable indicating if an individual has a mental health disorder, where 31% of the individuals in our sample are identified as such. We identify an individual as having a mental health disorder where their Kessler 10 score is at least 20, consistent with public health literature (Victorian Department of Human Services, 2001; Worksafe Queensland, 2003).

Our key explanatory variable is a dummy variable indicating if an individual has access to basic drinking services, which covers 76% of our sample. UNICEF defines this as being 30 minutes from a drinking water source, coming from an improved source (2017, p. 12). Improved sources include: piped water, boreholes or tubewells, protected dug wells, protected springs, rainwater, and packaged or delivered water.

The remaining variables represent the control variables and instrument ("distance from drinking water source") we use to produce robust results.

3. Econometric models and methods

We propose to use an Instrumental Variable (IV) approach to estimate the causal impact of water security on mental health. However, we use other models to rationalise this approach, starting with the following multiple regression model (Model 1):

$$Mental\ Health\ Disorder_i = \beta_0 + \beta_1 Basic\ Water\ Access_i + \beta_2 Female + \beta_3 Age_i + \beta_4 Religious\ Minority_i + \beta_5 Rural_i + \epsilon_i$$

In this model, β_1 is our key causal parameter. Given the complexity of mental health determinants, the other parameters are introduced to control for inherent Omitted Variable Bias (OVB) (Kirkbride et al. 2024). Specifically, we have chosen being female, part of the religious minority (non-Christian), age and living in a rural area as control variables.

Given our limitation of using just four control variables (as outlined in the Capstone Project Proposal), it is not possible to address any further OVB. Regional climate is a possible remaining source of OVB, but a relevant variable (such as rainfall) is unobservable in our dataset.

To improve our model, an IV approach is required to deal with further OVB and reverse causality driving our model's remaining endogeneity. Concerns about reverse causality are logical, as having a chronic mental disorder may influence an individual's ability to maintain consistent water access. Given we expect water access improves mental health outcomes, it's likely our causal parameter suffers from attenuation bias as individuals with worse mental health are likely to have worse access to water.

We propose to use an individual's distance from their drinking water source in minutes as the instrument. We consider it relevant as the distance to drinking water access is likely strongly correlated with an individual's basic water access.

We also consider it exogenous to mental health status, as we don't think how far someone has to travel for water will have a direct impact on their mental health. By using this IV approach, we directly address our reverse causality concerns by estimating the effect of changes in water access which are unrelated to the individual's mental health status.

This approach is given by Model 2 below, which is a just-identified two equation system:

$$(1)\ Mental\ Health\ Disorder_i = \beta_0 + \beta_1 Basic\ Water\ Access_i + \beta_2 Sex_i + \beta_3 Age_i + \beta_4 Religious\ Minority_i + \beta_5 Rural_i + \epsilon_i$$

$$(2)\ WaterAccess_i = \pi_0 + \pi_1 Distance\ From\ Drinking\ Water + \pi_2 Sex + \pi_3 Sex + \pi_4 Age + \pi_5 Religious\ Minority + \pi_6 Rural + \epsilon_i$$

A possible complaint of this model is that it introduces new sources of OVB into the error term: the impact of the terrain (difficulty) and vehicle ownership on the travelling time to water.

On the impact of terrain on accessing water, we consider our rural area control variable is a reasonable proxy for this given time and data constraints. By observing the topography of Ghana (Bessah et al., 2022, p. 3) it's clear that all major cities (Accra, Kumasi & Tamale) are at relatively low elevations compared to the mountainous rural areas.

For vehicle ownership, while data for Ghana is difficult to observe (in our dataset and otherwise), the World Bank (2022, p. 7) estimates Africa's vehicle penetration rate at "31 light-duty vehicles per 1,000 persons". Therefore, we don't think vehicle ownership is a reasonable source of endogeneity above those control variables we've already identified.

4. Results

Table 2: Model estimate results

	Dependent variable:		
	Has mental health disorder	Has access to basic water services	Has mental health disorder
	(1)	(2a)	(2b)
Distance from drinking water source (mins)		-0.010*** (0.001)	
Access to basic drinking water services	-0.093*** (0.012)		-0.221*** (0.028)
Female	0.086*** (0.009)	0.007 (0.008)	0.086*** (0.009)
Age	0.004*** (0.0002)	-0.0002 (0.0002)	0.004*** (0.0003)
Part of religious minority	0.099*** (0.010)	-0.039*** (0.009)	0.083*** (0.011)
Lives in rural area	0.074*** (0.010)	-0.140*** (0.008)	0.044*** (0.012)
Constant	0.100*** (0.017)	1.031*** (0.011)	0.220*** (0.029)
Observations	9,282	9,282	9,282
F Statistic	130.714***	672.072***	

Note: *p<0.1; **p<0.05; ***p<0.01

In the OLS multiple regression Model 1, access to basic drinking water is associated with a 9.3 percentage point decline in the likelihood of having a mental health disorder. This result is highly significant even at the 1% level. Similarly, our choice of controls are all statistically significant at the 1% level of significance, indicating that these factors play a role in increasing the risk of mental disorders. This is interpreted as being female, an additional year of age, being in a religious minority and living in a rural area all

increase this likelihood of having a mental health disorder by 8.6, 0.04, 9.9 and 7.4 percentage points respectively.

In the first stage of our IV 2SLS, Model 2a, regressing basic water access on the control variables rejected the F-test null hypothesis at the 1% significance level, indicating that our instrument is relevant.

In our second stage, Model 2b, the regression of mental health disorder likeliness on basic water access is interpreted as basic water access decreasing the likelihood of mental disorders by 22.1 percentage points. This result holds significant at the 1% level, with the same applying to all the control variables. These controls are understood as being female, an additional year of age, being in a religious minority and living in a rural area all increase the likelihood of a mental disorder by 8.6, 0.4, 8.3 and 4.4 percentage points respectively. The increase in the estimate magnitude from 0.093 (OLS model) to 0.221 (IV 2SLS model) signifies that the initial model suffered from attenuation bias due to endogeneity, which was corrected by our choice of using an IV estimation approach.

5. Robustness checks and Extensions

5.1 Testing different water access definitions

As our choice of water access definition is ultimately arbitrary, it is important to show our results are robust to different definitions to ensure they do not reflect some data artefact. To do this, we have re-estimated Model 2 with different UNICEF definitions for water access shown in Table 3.

Table 3: UNICEF water access definitions

Service level	**Definition**
Safely Managed Access	Drinking water from an improved source readily available on premises, free from contamination
Basic Access	Drinking water from an improved source, collection time of less than 30 minutes per round trip.
No Access	Drinking water directly from a river, dam lake, pond, stream, canal
Note: improved sources include: piped water, boreholes or tubewells, protected dug wells, protected springs, rainwater, and packaged or delivered water.	

Source: UNICEF (2021, p. 12).

The outcome of this exercise is shown in Table 5. These results are consistent with our earlier results of water security reducing the prevalence of mental health disorders, supporting the integrity of our model. These is clear by observing the causal parameter, β_1, decreases as water security increases from no access to safely managed access and our initial parameter estimate for the impact of basic water access on is between these other two extreme definitions of access. This means as an individual's access to water improves, as does their mental health, regardless of which definition is used

Table 4: Estimates of different causal parameter definitions

	Dependent variable:		
	Has mental health disorder		
	(3)	(2)	(4)
Has “safely managed” water access			-0.537*** (0.073)
Has “basic” water access		-0.221*** (0.028)	
Has “no” water access	0.531*** (0.083)		

Note: The estimation of these models is identical to Model 2 in Table 2 besides different causal parameters. Estimates for control variables have not been included as they are not the focus of this exercise.

5.2 IV strength

The second robustness test entailed conducting Wu Hausman and Weak Instruments tests on the instrumental variable to prove that our use of IV in this report was necessary. The findings are shown in Table 6.

Table 5: Statistical inference tests

Test name	df1	df2	statistic	p-value
Weak Instruments	1	9276	354.29	< 2e-16 ***
Wu Hausman	1	9275	30.37	3.67e-08 ***
Sargan	0	NA	NA	NA
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

The Weak Instruments test is rejected even at the 1% significance level, indicating that our instrument is not weak and fulfils the requirement of relevance and exogeneity. This is important because it shows that the two stage least squares model is unbiased and thus reliable.

By similarly rejecting the Wu Hausman test that our IV model was exogenous, we know that the presence of endogeneity in this model would have created biased and inconsistent OLS estimators. Hence, we validate the use of 2SLS IV to address this endogeneity problem.

6. Conclusion

Our findings confirm the hypothesis that water insecurity negatively impacts mental health outcomes. Specifically, we find access to basic water services decreases the likelihood of having a mental health disorder by around 22 percentage points. To mitigate the impact of OVB we implemented the use of several control variables, and then opted for IV approach to estimation due to concerns about reverse causality between water access and mental health disorders.

This estimation approach was supported by the weak instruments and Wu-Hausman tests and is robust against different water definitions. However, as we detail in section 3, it is possible some endogeneity remains in our final model. We consider climatic conditions and vehicle ownership are two clear remaining sources of OVB due to data constraints.

However, even if we could control for vehicle ownership we don’t think this would be worthwhile given the very low rates of car ownership in Africa, implying this issue is a second-order concern given the limited number of control variables we’ve been able to use.

7. Appendix and References

Achore, M., & Bisung, E. (2022). Experiences of inequalities in access to safe water and psycho-emotional distress in Ghana. Social Science & Medicine, 114970, 301. https://doi.org/10.1016/j.socscimed.2022.114970

Apanga, P. A., Weber, A. M., Darrow, L. A., Riddle, M. S., Tung, W.-C., Liu, Y., & Garn, J. V. (2021). The interrelationship between water access, exclusive breastfeeding and diarrhea in children: A cross-sectional assessment across 19 African countries. Journal of Global Health, 11, 4001. https://doi.org/10.7189/jogh-11-04001

Bessah, E., Amponsah, W., Ansah, S. O., Afrifa, A., Yahaya, B., Wemegah, C. S., Tanu, M., Amekudzi, L. K., & Agyare, W. A. (2022). Climatic zoning of Ghana using selected meteorological variables for the period 1976–2018. Meteorological Applications, 29(1). https://doi.org/10.1002/met.2049

Fink, G., Günther, I., & Hill, K. (2011). The effect of water and sanitation on child health: Evidence from the demographic and health surveys 1986–2007. International Journal of Epidemiology, 2011,1–9. https://doi.org/10.1093/ije/dyr102

Kirkbride, J., Anglin, D., Colman, I., Dyxhoorn, K., Jones, P., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., Griffiths, S. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Pychiatry, 23(1). https://doi.org/10.1002/wps.21169

UNICEF. (2021). *Drinking water - UNICEF DATA*. UNICEF DATA. https://data.unicef.org/topic/water-and-sanitation/drinking-water/

Victorian Population Health Survey. Melbourne: Department of Human Services, Victoria; 2001. https://nla.gov.au/nla.obj-310121787/view

World Bank (2022). Motorization Management for Development: An Integrated Approach to Improving Vehicles for Sustainable Mobility. https://documents1.worldbank.org/curated/en/099645006172219760/pdf/P1731880a0ad070ff0a6cf01b424e24ed04.pdf

World Health Organisation. (1948). Constitution of the World Health Organisation. https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1

WorkSafe Queensland. (2003). *Kessler Psychological Distress Scale (K10)*. https://www.worksafe.qld.gov.au/__data/assets/pdf_file/0010/22240/kessler-psychological-distress-scale-k101.pdf

8. Code Output

Econometrics 2 capstone final report data code

Josh Copeland, Jocelyn Koswara and Ryan Luo

2024-10-1

Importing and cleaning data

Tables used for the progress report:

Psychology (S10AI)
Housing: water (S12AI)
Household background information (S1D)
Key household information (key_hhld_info)

In order to derive the following variables:

Binary variable indicating mental health status (1 = likely to have a mental health disorder) (S10AI) Binary variable indicating access to basic drinking water services (1 = has access) (S1D) Age (S10AI)
Binary variable indicating sex (1 = female) (S10AI)
Binary variable indicating religious minority (1 = not Christian) (S1D)
Binary variable indicating if the person lives in an urban or rural area (1 = in an urban area) (S1D) Analysis in this markdown document is separated by each data table imported. file:///C:/Users/joshc/OneDrive/Desktop/git/econometrics_2/3_final_report/ecom2_final.html 1/23

Importing the Pyschology table

```
#####################################################################
############### ################################## PSYCHOLOGY TABLE
###############################
#####################################################################
###############

s10ai <- read_csv("data/S10AI.csv") %>%
```

```
 select(hhno, hhmid, depression, sex = s1d_1, age = s1d_4i) %>%

 #Creating a new column as our depression_dummy. Kessler scores between 10-19
have a score o f one in the data (== "likely to be well"). Anyone with scored
higher than this has a score > 1, which classifies them as likely to have at
least a mild disorder.
 mutate(depression_dummy = case_when(

 depression > 1 ~ 1, # Depressed
 TRUE ~ 0 # Not depressed

 )) %>%

 # Turning sex into a dummy variable (1 == female)

 mutate(sex = case_when(

 sex == 1 ~ 0,
 sex == 2 ~ 1

 ))


######################### EXTRACTING JUST THE RELEVANT VARIABLES ################

s10ai <- s10ai %>%
 select(hhno, hhmid, depression_dummy, sex_dummy = sex, age)
```

Importing the housing table

We are importing this table to create a dummy variable for access to basic drinking services.

UNICEF defines a household's access to water as "basic" if it satisfies the following conditions:

> It's delivered from one of the following sources: piped water, boreholes, tubewells, protected dug well, protected springs, rainwater and packaged of delivered water.
>
> A round trip to collect water does not exceed 30 minutes.

UNICEF actually has different definitions for water access, where we want to capture dummies for them as one of our extensions. They are defined as:

> Safely managed: Drinking water from an improved water souce which is located on

premises, available when needed and free of faecal and priority contamination.

Basic: as defined above.

No services: drinking water collected directly fro ma river, dam, lake, pond, stream, canal or irrigation channel.

```
#####################################################################
###############  ################################  HOUSING  TABLES
#################################
#####################################################################
###############
```

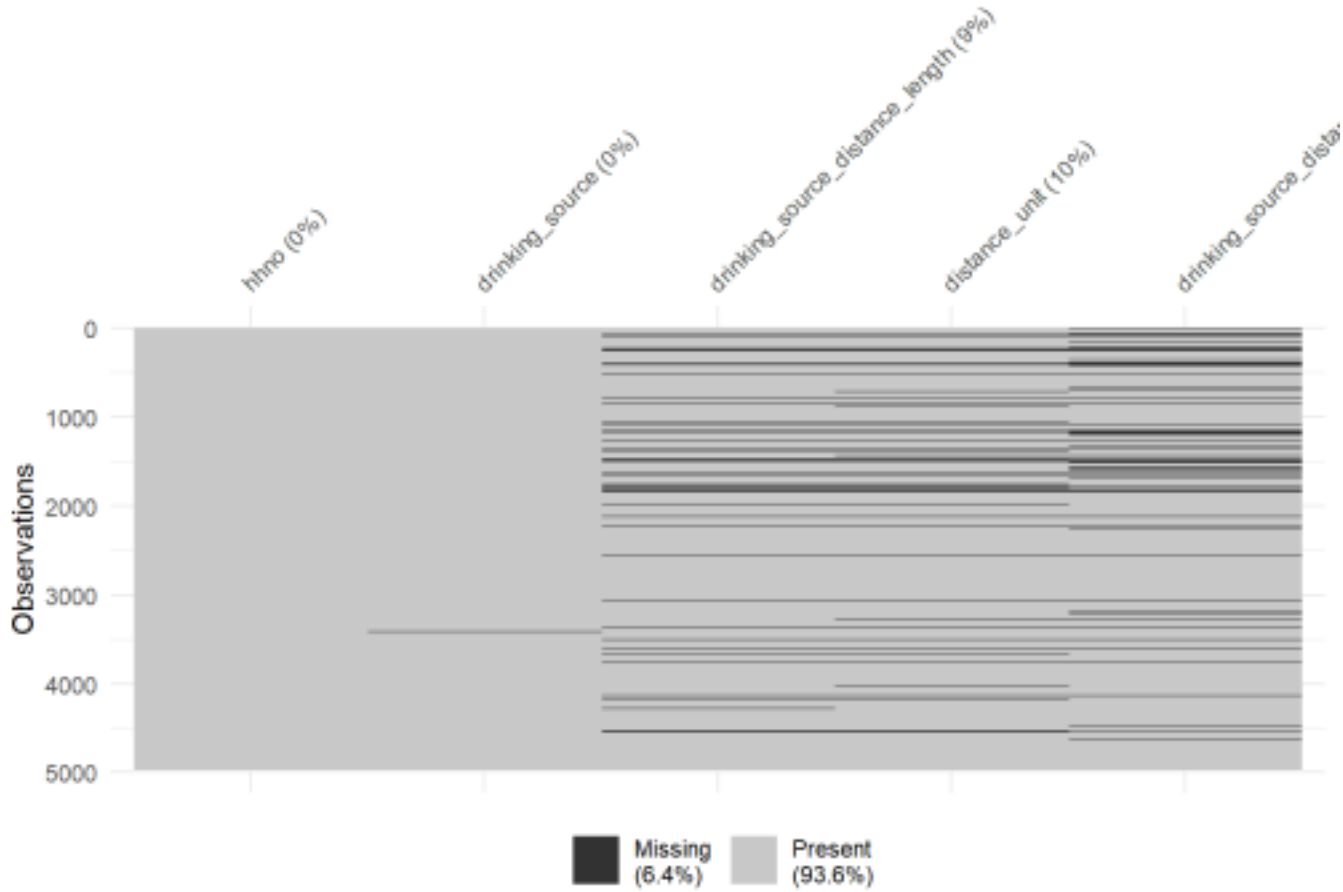
```
################################ WATER TABLE ####################################
```

```
s12ai <- read_csv("data/S12AI.csv") %>%
 select(hhno,drinking_source = s12a_9i, drinking_source_distance_length =
s12a_10ai, distanc e_unit = s12a_10aii, drinking_source_distance_mins =
s12a_11) %>%

 #Editing the drinking_source_distance cells to make them all the same
scale: kilometres.
 mutate(drinking_source_distance_length = case_when(

 distance_unit == 0 ~ 0, # In house
 distance_unit == 1 ~ as.numeric(drinking_source_distance_length) * 0.0009144,
# Yards to kilometers
 distance_unit == 2 ~ as.numeric(drinking_source_distance_length) / 1000, #
Meters to kil ometers
 distance_unit == 3 ~ as.numeric(drinking_source_distance_length), # Already
in kilometer s
 distance_unit == 4 ~ as.numeric(drinking_source_distance_length) * 1.609344,
# Miles to kilometers
 TRUE ~ drinking_source_distance_length

 ))
```

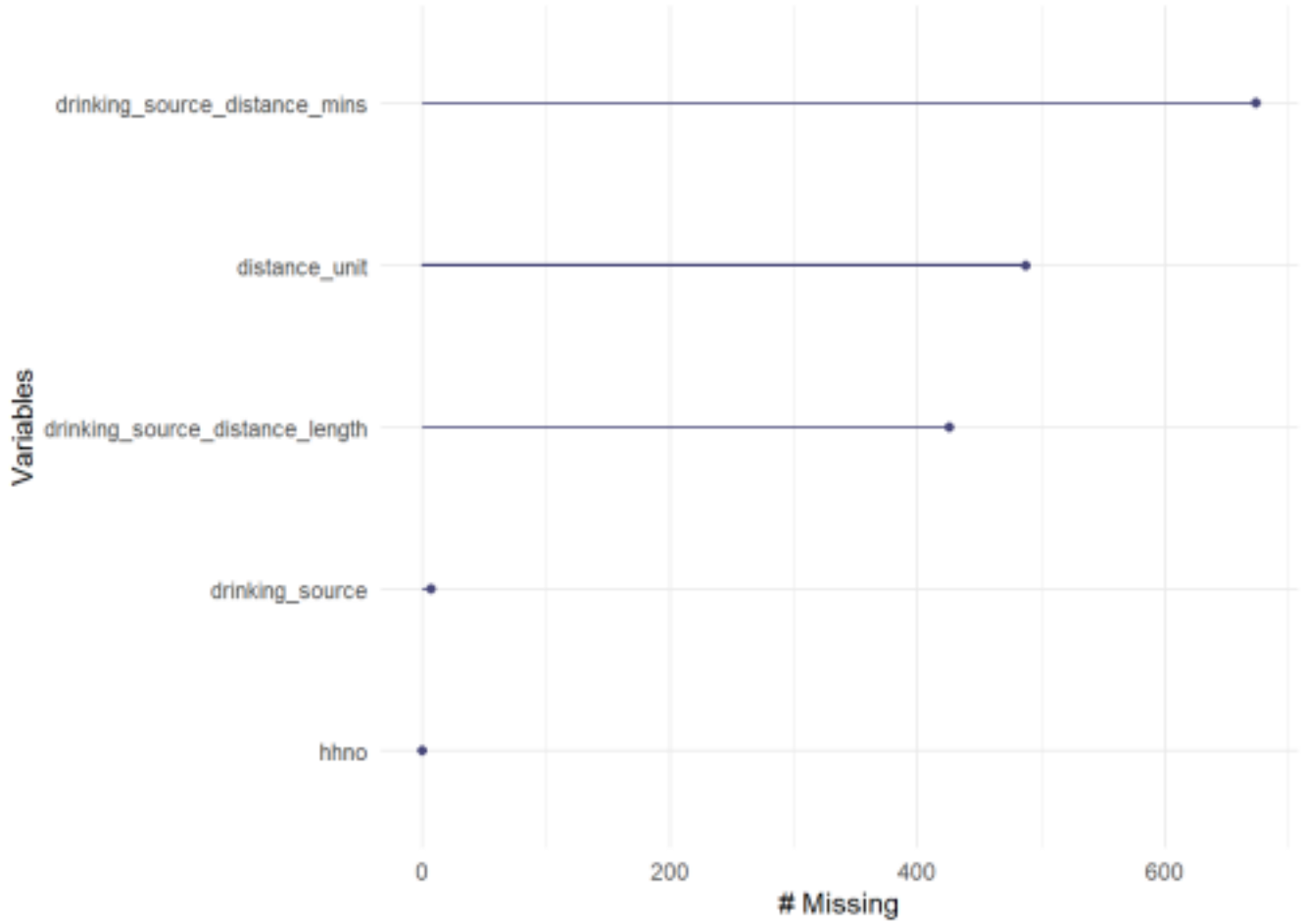
```
## Warning: One or more parsing issues, call `problems()` on your data
frame for details, ## e.g.:
```

```
## dat <- vroom(...)
## problems(dat)


## Rows: 4972 Columns: 72
## ── Column specification
```

```
─── ## Delimiter: ","
## chr (2): s12a_15, s12a_15i
## dbl (67): id1, id3, id4, id2, s12a_1, s12a_2i, s12a_2ii,
s12a_2iii, s12a_3, ... ## lgl (3): s12a_4i, s12a_4ii, s12a_4iii
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet


this message. vis_miss(s12ai)
```




```
gg_miss_var(s12ai)
```



The charts above shows us that there is a lot of missing values for the distance variables in both length and mins. This likely have something todo with the drinking source of each household. I need to collect all the NA data together in order to diagnose the problem.

The charts below show us that:

Most of the problem is in 1 and 2, which correspond to plumbing in the house. We can change their distances to zero.

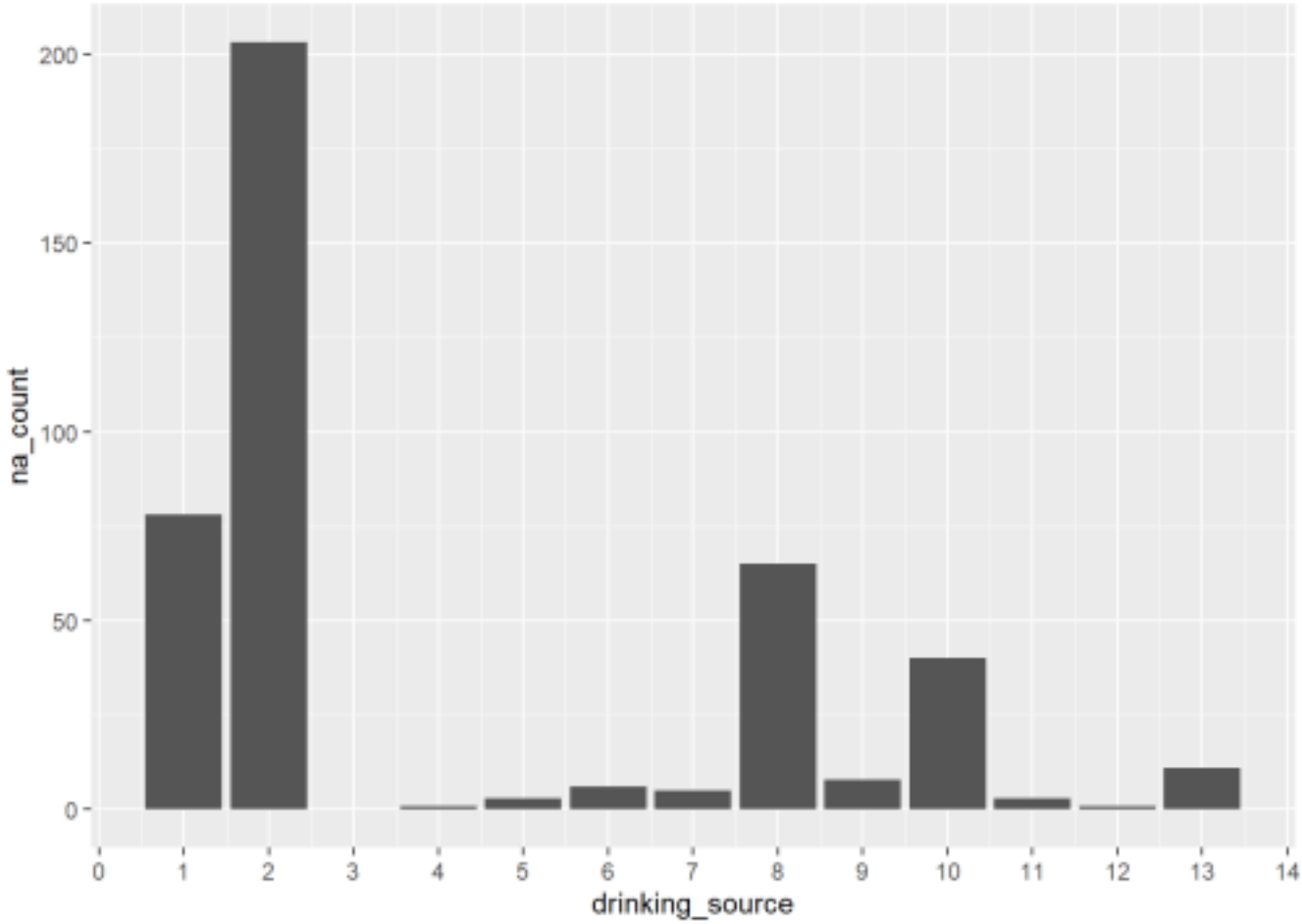
8 is also a clear problem, which is bottled water. We think its reasonable to assume this botteld water is available at the house, so can change this distance to zero as well.

9 and 10 are protected wells and boreholes. Without more information about how far away they are (unavailable) we need to leave these as NAs.

```
# Extracting and charting NA data

na_data <- s12ai %>%
 filter(is.na(drinking_source_distance_length)) %>%
 group_by(drinking_source) %>%
```

```
 summarise(na_count = n())

ggplot(na_data, aes(x = drinking_source, y = na_count)) +
 geom_bar(stat = "identity") +
 scale_x_continuous(breaks = scales::pretty_breaks(n = 14))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



 # Now I have diagnosed the problem, I need to make the necessary changes to the dataframe suc h that dirnking_sources with values 1 and 2 have a distance of zero in both length and minute s. All other NAs remain given data limitations.

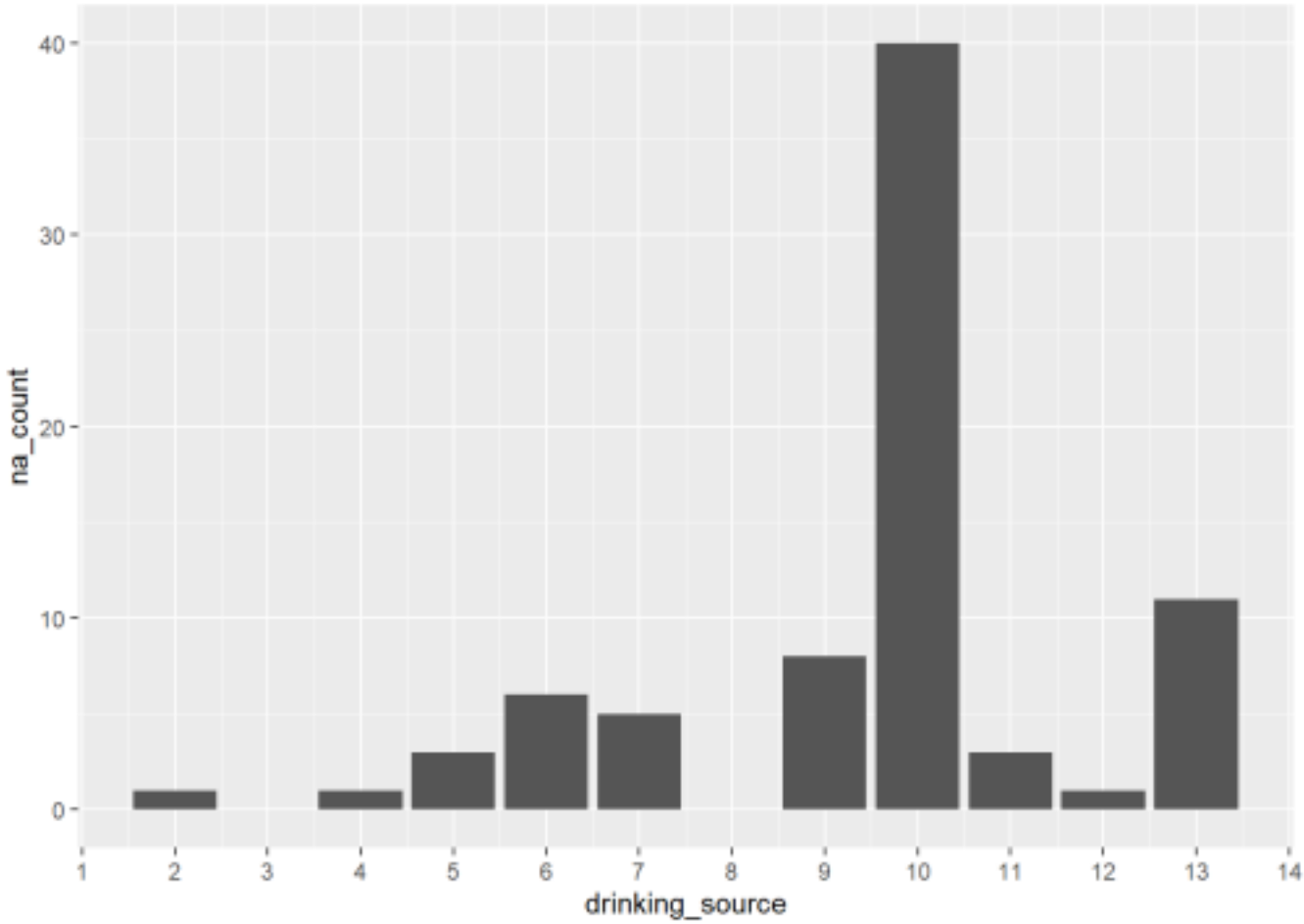
```
s12ai <- s12ai %>%
 mutate(drinking_source_distance_length = case_when(

 is.na(distance_unit) & drinking_source %in% c(1, 2, 8) ~ 0,
```

```
 TRUE ~ drinking_source_distance_length
 )) %>%

 mutate(drinking_source_distance_mins = case_when(
 is.na(distance_unit) & drinking_source %in% c(1, 2, 8) ~ 0,
 TRUE ~ drinking_source_distance_mins
 ))
```

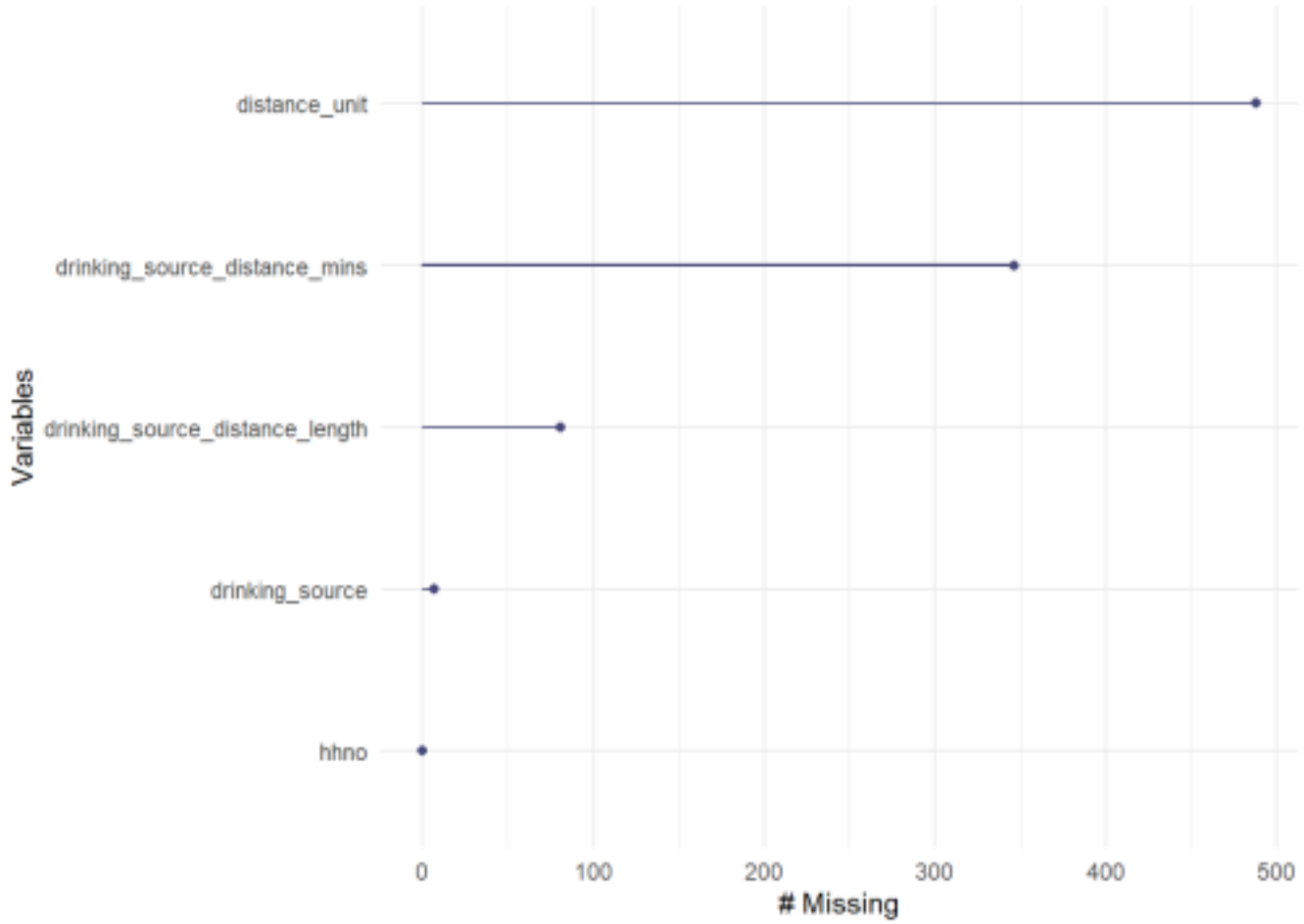
 # Repeating the NA value analysis/chart below, the scale are now sufficiently small to contin ue/we don't have any other information that could help reduce the incidence of NAs.

```
na_data <- s12ai %>%
 filter(is.na(drinking_source_distance_length)) %>%
 group_by(drinking_source) %>%
 summarise(na_count = n())

ggplot(na_data, aes(x = drinking_source, y = na_count)) +
 geom_bar(stat = "identity") +
 scale_x_continuous(breaks = scales::pretty_breaks(n = 14))
```

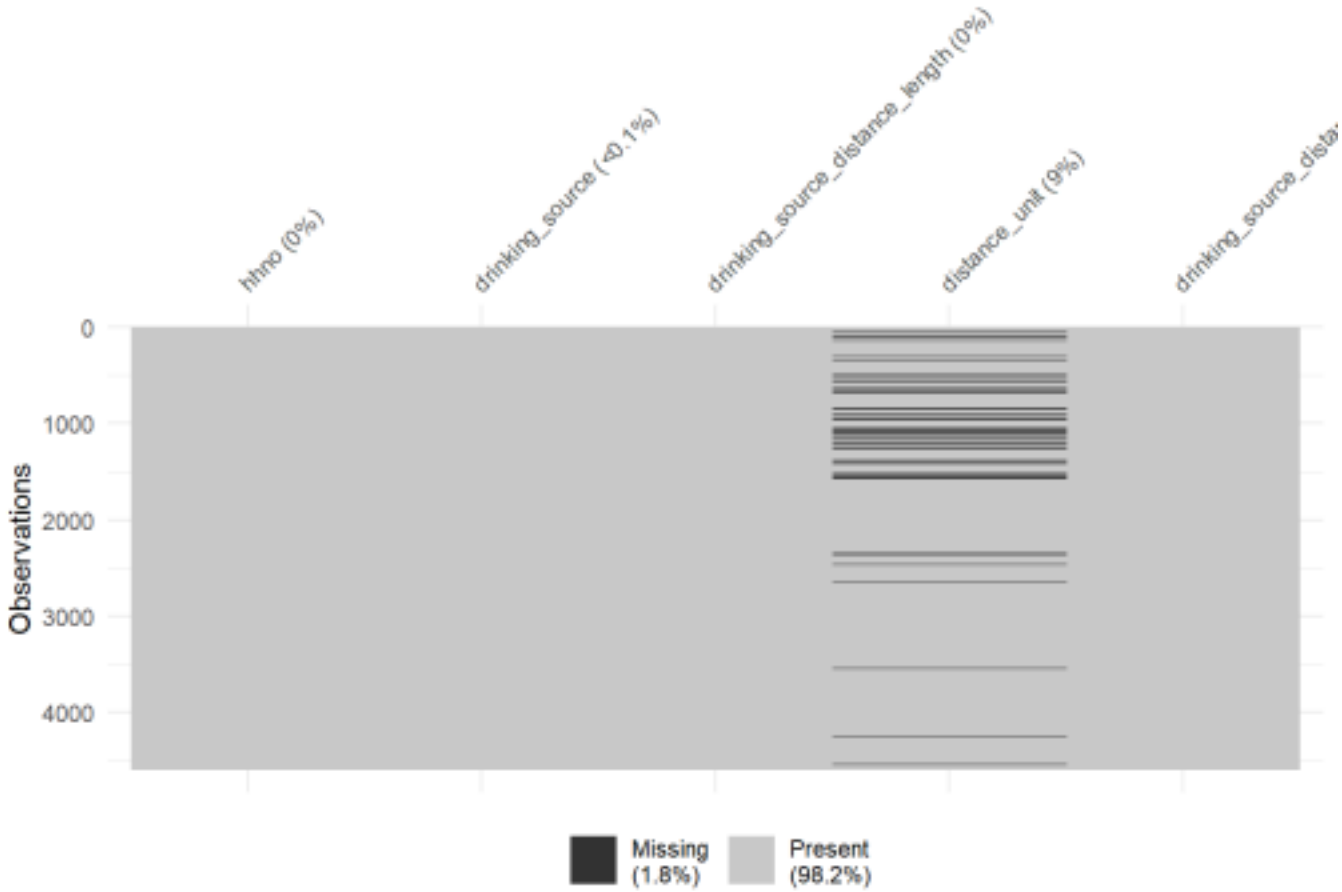
```
## Warning: Removed 1 row containing missing values or values outside the scale range ##
(`geom_bar()`).
```




```
gg_miss_var(s12ai)
```




```
# Because we can't deal with the remaining NAs, we exclude them from our analysis. However, w e only exclude where NAs appear in the drinking_source_distance_length and drinking_source_di stance_mins variables.

s12ai <- s12ai %>%
 filter(!is.na(drinking_source_distance_length)) %>%
 filter(!is.na(drinking_source_distance_mins))

vis_miss(s12ai)
```



Now we can actually produce our dummy variable for access to “basic drinking services”.


```
s12ai <- s12ai %>%
```

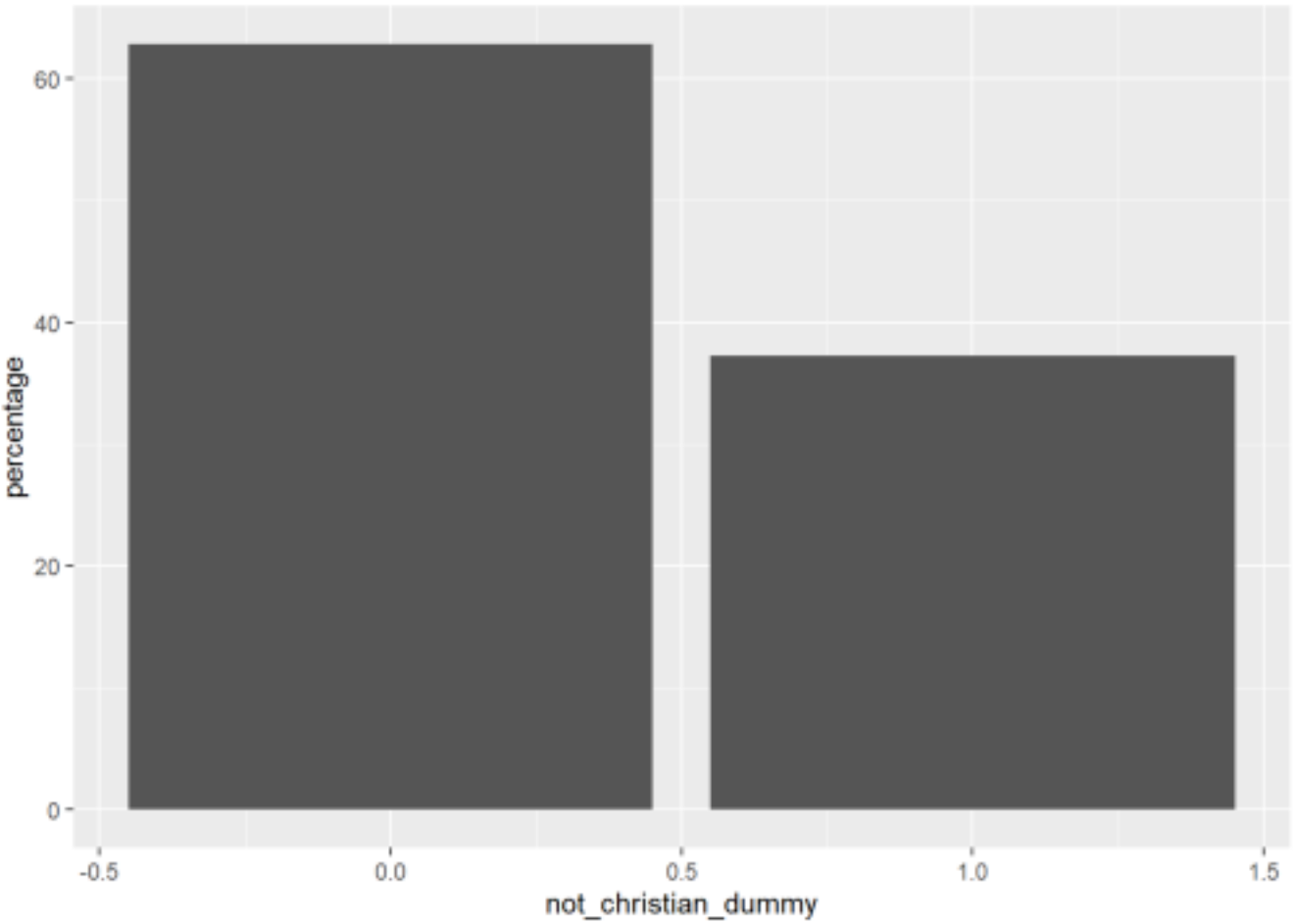
```
# Basic access

mutate(

basic_access_dummy = case_when(

drinking_source_distance_mins <= 30 &
drinking_source %in% c(1, # Indoor plumbing
2, # Inside standpipe
                              5, # Pipe in niehgbouring household
                               6, # Private outside standpipe/tap
                             7, # Public standpipe
                             8, # Sachet/bottled water
                             9, # Borehole
                             10) # Protected well
~ 1,
TRUE ~ 0
)) %>%

# Safely managed

mutate(

safely_managed_dummy = case_when(

drinking_source_distance_mins <= 2 &
drinking_source %in% c(1, # Indoor plumbing
2, # Inside standpipe
                              5, # Pipe in niehgbouring household
                               6, # Private outside standpipe/tap
                             7, # Public standpipe
                             8, # Sachet/bottled water
                             9, # Borehole
                             10) # Protected well
~ 1,
TRUE ~ 0
)) %>%

# No service

mutate(no_service_dummy = case_when(

drinking_source %in% c(12, # River/Stream
14) #Dugout/ponk/lake/dam

~ 1,
```

```
TRUE ~ 0
))
```


Importing the hosehold background information table

```
########################## RELIGIOUS MINORITY DUMMY ##############################


s1d <- read_csv("data/S1D.csv") %>%
 select(hhno, hhmid, religion = s1d_13, ethnicity = s1d_16) %>%
 mutate(not_christian_dummy = 0) %>%
 mutate(not_christian_dummy = case_when(

 # The following values of religion correspond with Christianity: 1,2,3,4,5 and 7.
 religion %in% c(1,2,3,4,5,7) ~ 0,
 TRUE ~ 1

 ))


## Warning: One or more parsing issues, call `problems()` on your data frame for details, ##
e.g.:
## dat <- vroom(...)
## problems(dat)


## Rows: 18889 Columns: 48
## ── Column specification
──────────────────────────────────────────────────────────────────────── ## Delimiter:
","
## dbl (46): id1, id2, id3, id4, hhmid, s1d_1, s1d_2, sid_3i, s1d_3ii, s1d_3iii... ##
lgl (2): s1d_28, s1d_33
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Is it reasonable to think of Christian as the relgious majority? The chart below suggest th ey
account for ~ 60% of the population. Therefore, it's reasonable to account for non-Christi ans
are part of the relgious minority in Ghana.

religion_dummy_frequency <- s1d %>%
 group_by(not_christian_dummy) %>%
 summarise(count = n()) %>%
 mutate(percentage = (count / sum(count)) * 100)



ggplot(religion_dummy_frequency, aes(not_christian_dummy, percentage)) + geom_bar(stat = "ide
ntity")
```




```
######################### EXTRACTING JUST THE RELEVANT VARIABLES ################
s1d <- s1d %>%
 select(hhno, hhmid, not_christian_dummy)
```

Importing key household information

```
key_hhld_info <- read_csv("data/key_hhld_info.csv") %>%
 select(hhno, rural_dummy = urbrur) %>%
 mutate(rural_dummy = case_when(

 rural_dummy == "1" ~ 0,
 TRUE ~ 1

 ))


## Rows: 5009 Columns: 9
## ── Column specification
──────────────────────────────────────────────────────────────────────── ## Delimiter:
```

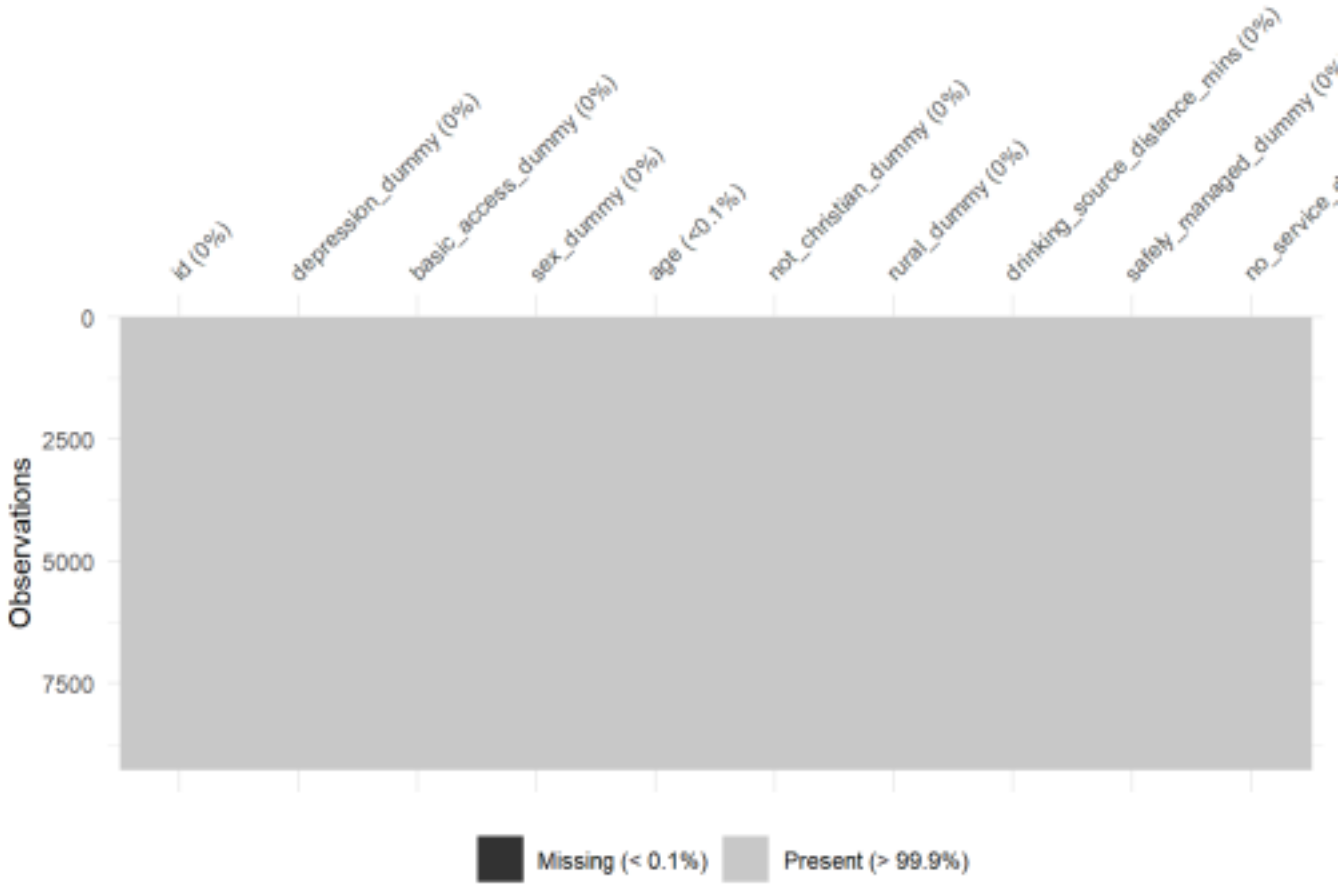
```
","
## dbl (9): id1, id2, id3, id4, hhno, urbrur, loc7, hhweight3, ppweight3
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Joining data

Household data is not provided at the individual level. Therefore, we need to append it to our psychological data.

Doing a quick NA visualisation I can see that there are a few columns with NA values. Given how small they are as proportions, I omit the NA values for depression and drinking_source_distance. I don't both with distance_unit (its only use was to help us clean the data earlier.)

```
data <- s10ai %>%
 inner_join(s12ai, by = "hhno") %>%
 inner_join(key_hhld_info, by = "hhno") %>%
 inner_join(s1d, by = c("hhno", "hhmid")) %>% # This data is collected on the individual, t
herefore we need to join at the sub-household level.

 mutate(id = hhno + hhmid) %>% # Creating a single hh identifier column

 select(id, depression_dummy, basic_access_dummy, sex_dummy, age, not_christian_dummy, rural
_dummy, drinking_source_distance_mins, safely_managed_dummy, no_service_dummy) #getting data
columns into a helpful order

vis_miss(data)
```



#

```
Omitting the very few remaining NA values
data <- data %>%
 na.omit()
```


Creating summary statistics

```
vars <- colnames(data)[!colnames(data) %in% c("id")]

# Create summary statistics
summary_stats <- data %>%
 summarise(across(all_of(vars),
 list(
 mean = ~ mean(.x, na.rm = TRUE),
 sd = ~ sd(.x, na.rm = TRUE),
 min = ~ min(.x, na.rm = TRUE),
```

```
 max = ~ max(.x, na.rm = TRUE)
 ),
 .names = "{.col}_{.fn}"))

# Reshape to long format
summary_stats <- summary_stats %>%
 pivot_longer(cols = everything(),
 names_to = c("variable", "statistic"),
 names_pattern = "(.*)_(.*)") %>% # Match everything before the last undersco re
 mutate(value = round(value,2))


summary_stats <- summary_stats %>%
 pivot_wider(names_from = statistic, values_from = value)

summary_stats$max <- format(summary_stats$max, scientific = FALSE)


print(summary_stats)


## # A tibble: 9 × 5
## variable mean sd min max
## <chr> <dbl> <dbl> <dbl> <chr>
## 1 depression_dummy 0.31 0.46 0 " 1"
## 2 basic_access_dummy 0.76 0.43 0 " 1"
## 3 sex_dummy 0.55 0.5 0 " 1"
## 4 age 39.1 18.7 1 "109"
## 5 not_christian_dummy 0.34 0.47 0 " 1"
## 6 rural_dummy 0.65 0.48 0 " 1"
## 7 drinking_source_distance_mins 15.6 18.0 0 "240"
## 8 safely_managed_dummy 0.12 0.33 0 " 1"
## 9 no_service_dummy 0.15 0.35 0 " 1"


##################################### SAVING OFF DATA ############################

write_csv(summary_stats, "summary_stats.csv")

write_csv(data, "model_data.csv")
```

Producing model outputs

```
data <- read_csv("model_data.csv")
```

```
## Rows: 9282 Columns: 10
## ── Column specification
──────────────────────────────────────────────────────────────────────── ## Delimiter:
","
## dbl (10): id, depression_dummy, basic_access_dummy, sex_dummy, age, not_chri... ##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Key models/output

```
################################### MODELS ######################################

# linear regression model
m1 <- lm(depression_dummy ~ basic_access_dummy, data = data)


# Multiple linear regresion model
m2 <- lm(depression_dummy ~ basic_access_dummy + sex_dummy + age + not_christian_dummy + rura
l_dummy, data = data)


# First stage 2SLS model
m3 <- lm(basic_access_dummy ~ drinking_source_distance_mins + sex_dummy + age + not_christian
_dummy + rural_dummy,
 data = data)

# Second stage 2SLS model
m4 <- ivreg(depression_dummy ~ basic_access_dummy + sex_dummy + age + not_christian_dummy + r
ural_dummy |
 sex_dummy + age + not_christian_dummy + rural_dummy + drinking_source_distance_mins, da ta =
data)

summary(m4, vcov=vcovHC, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = depression_dummy ~ basic_access_dummy + sex_dummy +
## age + not_christian_dummy + rural_dummy | sex_dummy + age +
## not_christian_dummy + rural_dummy + drinking_source_distance_mins,
## data = data)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.7651 -0.3145 -0.1899 0.4967 0.9777
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2200668 0.0294799 7.465 9.08e-14 ***
## basic_access_dummy -0.2208962 0.0280809 -7.866 4.06e-15 ***
## sex_dummy 0.0861642 0.0093199 9.245 < 2e-16 ***
## age 0.0038505 0.0002521 15.273 < 2e-16 ***
## not_christian_dummy 0.0833770 0.0108167 7.708 1.41e-14 ***
## rural_dummy 0.0443765 0.0115649 3.837 0.000125 ***
##
## Diagnostic tests:
## df1 df2 statistic p-value
## Weak instruments 1 9276 354.29 < 2e-16 ***
## Wu-Hausman 1 9275 30.37 3.67e-08 ***
## Sargan 0 NA NA NA
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4494 on 9276 degrees of freedom
## Multiple R-Squared: 0.05307, Adjusted R-squared: 0.05256
## Wald test: 133.4 on 5 and 9276 DF, p-value: < 2.2e-16
```

Formatting output

```
##################################### OUTPUT TABLES ################################
```

```
key_results <- stargazer(m1,
 m2,
                         m3,
                         m4,
                         type = "html",
```

```
                          title = "Table 2: Model estimates",
     se = list(
     sqrt(diag(vcovHC(m1))),
                           sqrt(diag(vcovHC(m2))),
                           sqrt(diag(vcovHC(m3))),
                           sqrt(diag(vcovHC(m4)))
                           ),
     out = "Key_Results.html"

     )
```



```
##
## <table style="text-align:center"><caption><strong>Table 2: Model estimates</strong></capti
on>
## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left"></td><td colspan="4"><em>Dependent variable:</em></td></tr>
## <tr><td></td><td colspan="4" style="border-bottom: 1px solid black"></td></tr> ## <tr><td
style="text-align:left"></td><td colspan="2">depression_dummy</td><td>basic_access
_dummy</td><td>depression_dummy</td></tr>
## <tr><td style="text-align:left"></td><td colspan="2"><em>OLS</em></td><td><em>OLS</em></td
><td><em>instrumental</em></td></tr>
## <tr><td style="text-align:left"></td><td colspan="2"><em></em></td><td><em></em></td><td><
em>variable</em></td></tr>
## <tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td><td>(3)</td><td>(4)</td></tr> ##
<tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left">basic_access_dummy</td><td>-0.134<sup>***</sup></td><td>-0.093<sup>***</sup></td><
td></td><td>-0.221<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.012)</td><td>(0.012)</td><td></td><td>(0.028)
</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left">drinking_source_distance_mins</td><td></td><td></td><td>-
0.010<sup>***</sup></td><td></td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.001)</td><td></td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left">sex_dummy</td><td></td><td>0.086<sup>***</sup></td><td>0.0
07</td><td>0.086<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td></td><td>(0.009)</td><td>(0.008)</td><td>(0.009)
</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left">age</td><td></td><td>0.004<sup>***</sup></td><td>-0.0002</
td><td>0.004<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td></td><td>(0.0002)</td><td>(0.0002)</td><td>(0.000
3)</td></tr>
```

```
## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left">not_christian_dummy</td><td></td><td>0.099<sup>***</sup></
td><td>-0.039<sup>***</sup></td><td>0.083<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td></td><td>(0.010)</td><td>(0.009)</td><td>(0.011)
</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left">rural_dummy</td><td></td><td>0.074<sup>***</sup></td><td>-
0.140<sup>***</sup></td><td>0.044<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td></td><td>(0.010)</td><td>(0.008)</td><td>(0.012)
</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left">Constant</td><td>0.409<sup>***</sup></td><td>0.100<sup>***
</sup></td><td>1.031<sup>***</sup></td><td>0.220<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.010)</td><td>(0.017)</td><td>(0.011)</td><td>
(0.029)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td></tr> ## <tr><td
colspan="5" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left">Observations</td><td>9,282</td><td>9,282</td><td>9,282</td><td>9,282</td></tr> ##
<tr><td style="text-align:left">R<sup>2</sup></td><td>0.015</td><td>0.066</td><td>0.266</t
d><td>0.053</td></tr>
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.015</td><td>0.065</td><td
>0.266</td><td>0.053</td></tr>
 ## <tr><td style="text-align:left">Residual Std. Error</td><td>0.458 (df = 9280)</td><td>0.44
```



```
6 (df = 9276)</td><td>0.367 (df = 9276)</td><td>0.449 (df = 9276)</td></tr>
## <tr><td style="text-align:left">F Statistic</td><td>145.179<sup>***</sup> (df = 1; 9280)</
td><td>130.714<sup>***</sup> (df = 5; 9276)</td><td>672.072<sup>***</sup> (df = 5; 9276)</td>
<td></td></tr>
## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left"><em>Note:</em></td><td colspan="4" style="text-align:right"><sup>*</sup>p<0.1; <su
p>**</sup>p<0.05; <sup>***</sup>p<0.01</td></tr>
## </table>


key_results <- stargazer(m1,
 m2,
                         m3,
                         m4,
                         type = "text",
                         title = "Table 2: Model estimates",
 se = list(
 sqrt(diag(vcovHC(m1))),
                          sqrt(diag(vcovHC(m2))),
                          sqrt(diag(vcovHC(m3))),
```

```
                    sqrt(diag(vcovHC(m4)))
                    ) )
```



```
##
## Table 2: Model estimates
## ================================================================================================
==================================
## Dependent variable:  ## ------------------------------------------------------------
---------------------------------
## depression_dummy basic _access_dummy depression_dummy
```

```
## OLS
OLS instrumental
##
variable
## (1) (2)
(3) (4)
## ------------------------------------------------------------------------------------------------
----------------------------------
## basic_access_dummy -0.134*** -0.093***  -0.221***
## (0.012) (0.012)
(0.028)
##
## drinking_source_distance_mins
-0.010***
##
(0.001)
##
## sex_dummy 0.086***
0.007 0.086***
## (0.009)
(0.008) (0.009)
##
## age 0.004***
-0.0002 0.004***
## (0.0002)
(0.0002) (0.0003)
##
## not_christian_dummy 0.099***
-0.039*** 0.083***
## (0.010)
(0.009) (0.011)
##
## rural_dummy 0.074***
-0.140*** 0.044***
## (0.010)
(0.008) (0.012)
##
## Constant 0.409*** 0.100***
1.031*** 0.220***
## (0.010) (0.017)
(0.011) (0.029)
##
## ------------------------------------------------------------------------------------------------
----------------------------------
## Observations 9,282 9,282
```

```
9,282 9,282
## R2 0.015 0.066
0.266 0.053
## Adjusted R2 0.015 0.065
0.266 0.053
## Residual Std. Error 0.458 (df = 9280) 0.446 (df = 9276) 0.36 7 (df = 9276) 0.449 (df = 9276)
## F Statistic 145.179*** (df = 1; 9280) 130.714*** (df = 5; 9276) 672.072* ** (df = 5; 9276)
## ================================================================================================
==================================
## Note:
*p<0.1; **p<0.05; ***p<0.01


##################### AUTOMATING OUTPUTTING TABLES TO XLSX FILE ################

key_results_table <- read_html("Key_Results.html") %>%
 html_table(fill = TRUE) %>%
 pluck(1) %>%
 filter(if_any(c(X1, X2, X3), ~ . != ""))

write_xlsx(key_results_table, "key_results_table.xlsx")
```

Robustness checks

Changing definitions of water access

Ultimately, our chosen definition of water access has been chosen arbitrarily from UNICEF's taxonomy of access. If our results are robust, the size of our estimators should decrease as we move from the lowest ("no service") to highest ("safely managed") level of water access. The better the access, the greater an impact it should have on reducing the incidence of mental health issues.

```
################################## CREATING MODELS ##############################


#No service
m5 <- ivreg(depression_dummy ~ no_service_dummy + sex_dummy + age + not_christian_dummy + rur
al_dummy |
 sex_dummy + age + not_christian_dummy + rural_dummy + drinking_source_distance_mins, da ta =
data)


# Basic access
m6 <- ivreg(depression_dummy ~ basic_access_dummy + sex_dummy + age + not_christian_dummy + r
ural_dummy |
 sex_dummy + age + not_christian_dummy + rural_dummy + drinking_source_distance_mins, da ta =
data)


# Safely managed
m7 <- ivreg(depression_dummy ~ safely_managed_dummy + sex_dummy + age + not_christian_dummy +
rural_dummy |
 sex_dummy + age + not_christian_dummy + rural_dummy + drinking_source_distance_mins, da ta =
data)



################################## COMPARING OUTPUT ##############################

rc_water_def_results <- stargazer(
 m5,
 m6,
 m7,
 type = "html",
 title = "Table X: Changing water definition estimates",
 se = list(
 sqrt(diag(vcovHC(m5))),
 sqrt(diag(vcovHC(m6))),
 sqrt(diag(vcovHC(m7)))
 ),
 out = "RC_Water_Def_Results.html"
 )
```

```
##
## <table style="text-align:center"><caption><strong>Table X: Changing water definition estim
ates</strong></caption>
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left"></td><td colspan="3"><em>Dependent variable:</em></td></tr>
## <tr><td></td><td colspan="3" style="border-bottom: 1px solid black"></td></tr> ## <tr><td
style="text-align:left"></td><td colspan="3">depression_dummy</td></tr> ## <tr><td
style="text-align:left"></td><td>(1)</td><td>(2)</td><td>(3)</td></tr> ## <tr><td colspan="4"
style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left">no_service_dummy</td><td>0.531<sup>***</sup></td><td></td><td></td></tr> ## <tr><td
style="text-align:left"></td><td>(0.083)</td><td></td><td></td><td></td></tr> ## <tr><td
style="text-align:left"></td><td></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">basic_access_dummy</td><td></td><td>-0.221<sup>***</sup></
td><td></td></tr>
## <tr><td style="text-align:left"></td><td></td><td>(0.028)</td><td></td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">safely_managed_dummy</td><td></td><td></td><td>-0.537<sup>
***</sup></td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.073)</td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">sex_dummy</td><td>0.090<sup>***</sup></td><td>0.086<sup>**
*</sup></td><td>0.088<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.010)</td><td>(0.009)</td><td>(0.010)</td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">age</td><td>0.004<sup>***</sup></td><td>0.004<sup>***</sup
></td><td>0.004<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.0003)</td><td>(0.0003)</td><td>(0.0003)</td></
tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">not_christian_dummy</td><td>0.077<sup>***</sup></td><td>0.
083<sup>***</sup></td><td>0.099<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.012)</td><td>(0.011)</td><td>(0.011)</td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">rural_dummy</td><td>-0.003</td><td>0.044<sup>***</sup></td
><td>-0.041<sup>*</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.018)</td><td>(0.012)</td><td>(0.021)</td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">Constant</td><td>0.001</td><td>0.220<sup>***</sup></td><td
>0.166<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.013)</td><td>(0.029)</td><td>(0.025)</td></tr> ##
<tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left">Observations</td><td>9,282</td><td>9,282</td><td>9,282</td></tr>
## <tr><td style="text-align:left">R<sup>2</sup></td><td>-0.068</td><td>0.053</td><td>-0.015
```

```
</td></tr>
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>-0.069</td><td>0.053</td><td
d>-0.016</td></tr>
## <tr><td style="text-align:left">Residual Std. Error (df = 9276)</td><td>0.477</td><td>0.44
9</td><td>0.465</td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-a
lign:left"><em>Note:</em></td><td colspan="3" style="text-align:right"><sup>*</sup>p<0.1; <su
p>**</sup>p<0.05; <sup>***</sup>p<0.01</td></tr>
## </table>
```



```
rc_water_def_results <- stargazer(
 m5,
 m6,
 m7,
 type = "text",
 title = "Table X: Changing water definition estimates",
 se = list(
 sqrt(diag(vcovHC(m5))),
 sqrt(diag(vcovHC(m6))),
 sqrt(diag(vcovHC(m7)))
 ) )
```

```
##
## Table X: Changing water definition estimates
## ==============================================================
## Dependent variable:
## ----------------------------
## depression_dummy
## (1) (2) (3)
## --------------------------------------------------------------
## no_service_dummy 0.531***
## (0.083)
##
## basic_access_dummy -0.221***
## (0.028)
##
## safely_managed_dummy -0.537***
## (0.073)
##
## sex_dummy 0.090*** 0.086*** 0.088***
## (0.010) (0.009) (0.010)
```

```
##
## age 0.004*** 0.004*** 0.004***
## (0.0003) (0.0003) (0.0003)
##
## not_christian_dummy 0.077*** 0.083*** 0.099***
## (0.012) (0.011) (0.011)
##
## rural_dummy -0.003 0.044*** -0.041*
## (0.018) (0.012) (0.021)
##
## Constant 0.001 0.220*** 0.166***
## (0.013) (0.029) (0.025)
##
## ---------------------------------------------------------------
## Observations 9,282 9,282 9,282
## R2 -0.068 0.053 -0.015
## Adjusted R2 -0.069 0.053 -0.016
## Residual Std. Error (df = 9276) 0.477 0.449 0.465
## ===============================================================
## Note: *p<0.1; **p<0.05; ***p<0.01
```



```
###################### AUTOMATING OUTPUTTING TABLES TO XLSX FILE ################

rc_water_def_results_table <- read_html("RC_Water_Def_Results.html") %>%
 html_table(fill = TRUE) %>%
 pluck(1) %>%
 filter(if_any(c(X1, X2, X3, X4), ~ . != ""))

write_xlsx(rc_water_def_results_table, "rc_water_def_results_table.xlsx")

print


## function (x, ...)
## UseMethod("print")
## <bytecode: 0x000001ca1f53e4a8>
## <environment: namespace:base>


###################### AUTOMATING OUTPUTTING TABLES TO XLSX FILE ################
```

```
key_results_table <- read_html("Key_Results.html") %>%
 html_table(fill = TRUE) %>%
 pluck(1) %>%
 filter(if_any(c(X1, X2, X3), ~ . != ""))

write_xlsx(key_results_table, "key_results_table.xlsx")
```