PHP2517: Applied Multilevel Data Analysis

Homework 2

Antonella Basso

May 9, 2022

Data

The "**wells.vill**" dataset includes well-switching data from a study that a research team from the US and Bangladesh conducted to understand the impact several factors have on the decisions of households in Bangladesh about whether to change their source of drinking water. A detailed description of this dataset can be found in Section 5.4. of the book. The `wells.vill.csv` dataset has information on the following variables:

- **id**: Well ID (number)
- **switch**: Whether the household switched to a new well (Yes=1, 0=No)
- **arsenic**: Arsenic level in respondent's well (in micrograms per leter, μg/L)
- **dist**: Distance to the nearest safe well (in meters)
- **assoc**: Whether any members of the household are active in community organizations (Yes=1, 0=No)
- **educ**: Education level of the head of the household (higher number indicate higher level)
- **village**: Village ID (letter)

*Note: The safety standard for arsenic is 0.5 μg/L.

Question 1:

a. Exploratory Data Analysis (EDA): Explore your data and provide appropriate descriptive statistics and plots for summarizing and presenting the information collected in this study, with emphasis to the primary research question, i.e., to assess the effect of important factors on the probability of switching wells.

b. Formulate an appropriate multilevel regression model predicting the probability of switching well using the arsenic level and the log transformation of distance (to the nearest well), allowing intercepts to vary across villages.

c. Extend the model in (b) to allow the coefficient on arsenic to vary across village. Discuss your results.

d. Create graphs of the probability of switching wells as a function of arsenic level for eight of the villages.

e. Compare the fit of the models in (b) and (c).

*Note: For each of the multilevel models you fit in this question please:

- Write the model using appropriate multilevel notation.
- Fit the model using the software of your preference, and present the results.
- Interpret the results from model fitting (both fixed and random effects).

Solution

a. **Exploratory Data Analysis (EDA)**

Overview of Data:

- 3,020 total observations[1]
- 1,737 total switches made (57.52% of households)
- 1,277 households have at least 1 active member in community organizations (42.29% of households)
- 18 unique education levels (0-17)
- 15 unique villages (A-O)
- arsenic levels range between 0.51-9.65 μg/L
- distances to the nearest safe well range between 0.387-339.531 meters

Descriptive Statistics[2]:

Table 1: Descriptive Statistics by Switch Group

Switch Group	Households	Mean Arsenic	Arsenic Var	Mean Distance	Distance Var	Assoc.	Prop. Assoc.
0	1283	1.4201	0.9099	53.6115	1869.679	569	0.4435
1	1737	1.8319	1.3886	44.4322	1158.324	708	0.4076

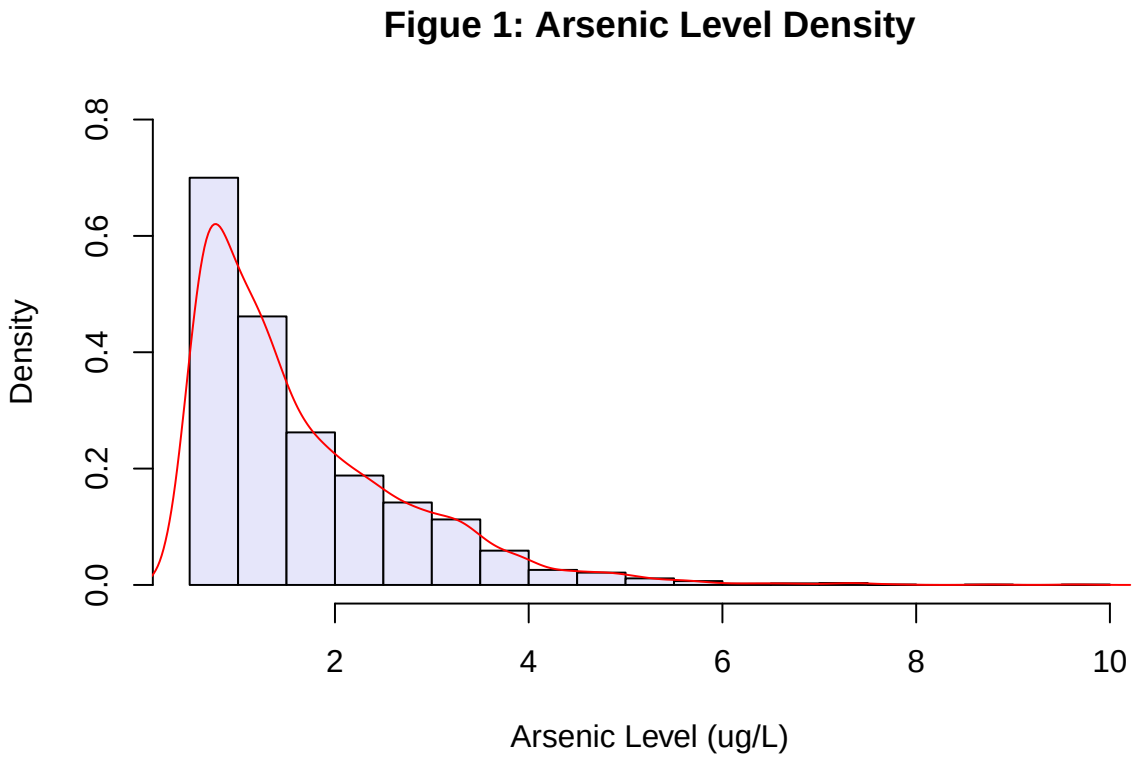
Table 2: Descriptive Statistics by Village

	Primary Outcome, Y=1			Other Predictors			
Village	Households	Switched	Prop. Switched	Mean Arsenic	Arsenic Var	Mean Distance	Distance Var
A	258	86	0.3333	0.5526	0.0008	31.4652	618.9645
B	456	225	0.4934	0.6943	0.0033	37.4696	867.2220
C	343	181	0.5277	0.8960	0.0033	42.6142	1108.1227
D	315	177	0.5619	1.1065	0.0034	42.9330	947.3526
E	267	150	0.5618	1.2991	0.0032	51.2644	1646.0816
F	199	126	0.6332	1.4948	0.0032	52.6660	2016.8196
G	165	97	0.5879	1.7030	0.0034	56.3453	1920.0057
H	147	93	0.6327	1.9070	0.0032	57.7630	1800.3008
I	284	184	0.6479	2.2448	0.0197	56.6242	2038.1190
J	214	144	0.6729	2.7443	0.0217	59.3330	1730.7547
K	170	122	0.7176	3.2516	0.0196	64.0308	1784.5568
L	89	62	0.6966	3.7526	0.0220	58.9390	1744.4521
M	71	56	0.7887	4.4785	0.0831	54.0685	1373.0419
N	34	28	0.8235	5.6121	0.2850	42.3963	1099.0493
O	8	6	0.7500	7.8612	0.7899	58.9441	1848.8157

[1]Where each observation corresponds to a unique household and well ID.

[2]Where: Village = Group j; Households = n_j; Switched = Frequency $\sum_j Y_{ij}$; % Switched = Relative Frequency $\frac{1}{n_j}\sum_j Y_{ij}$.

EDA Plots:


Figue 1: Arsenic Level Density
Density
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10
Arsenic Level (ug/L)

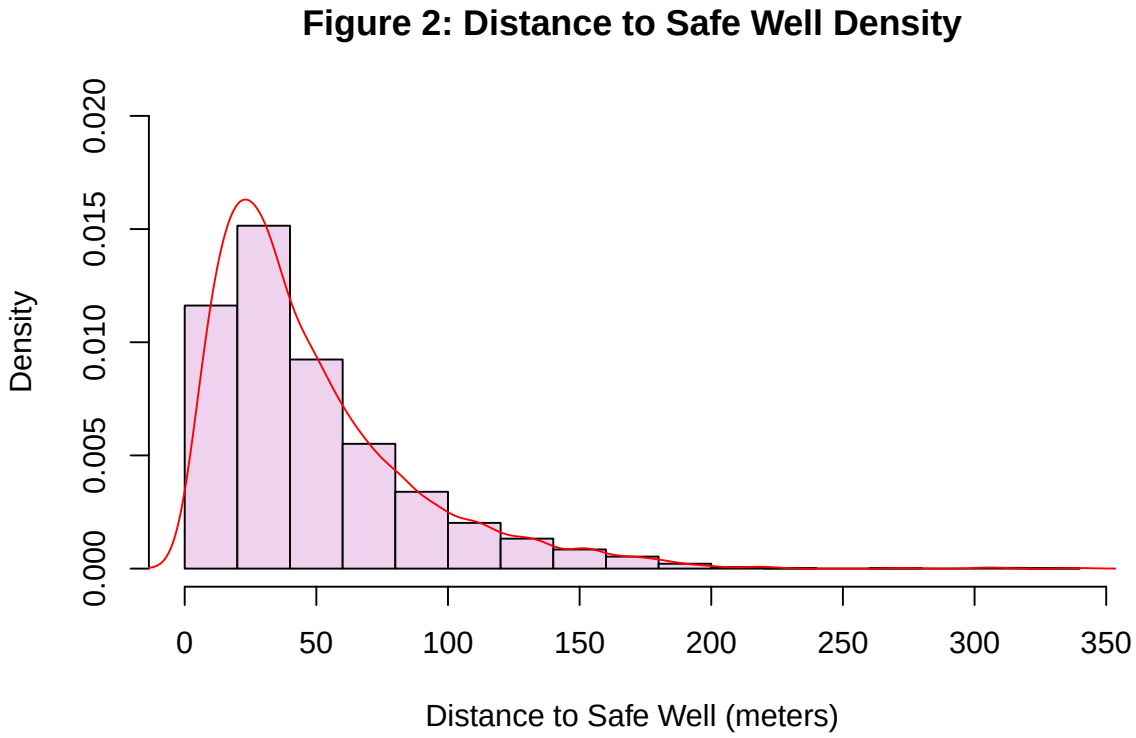


Figure 2: Distance to Safe Well Density
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
250
300
350
Distance to Safe Well (meters)

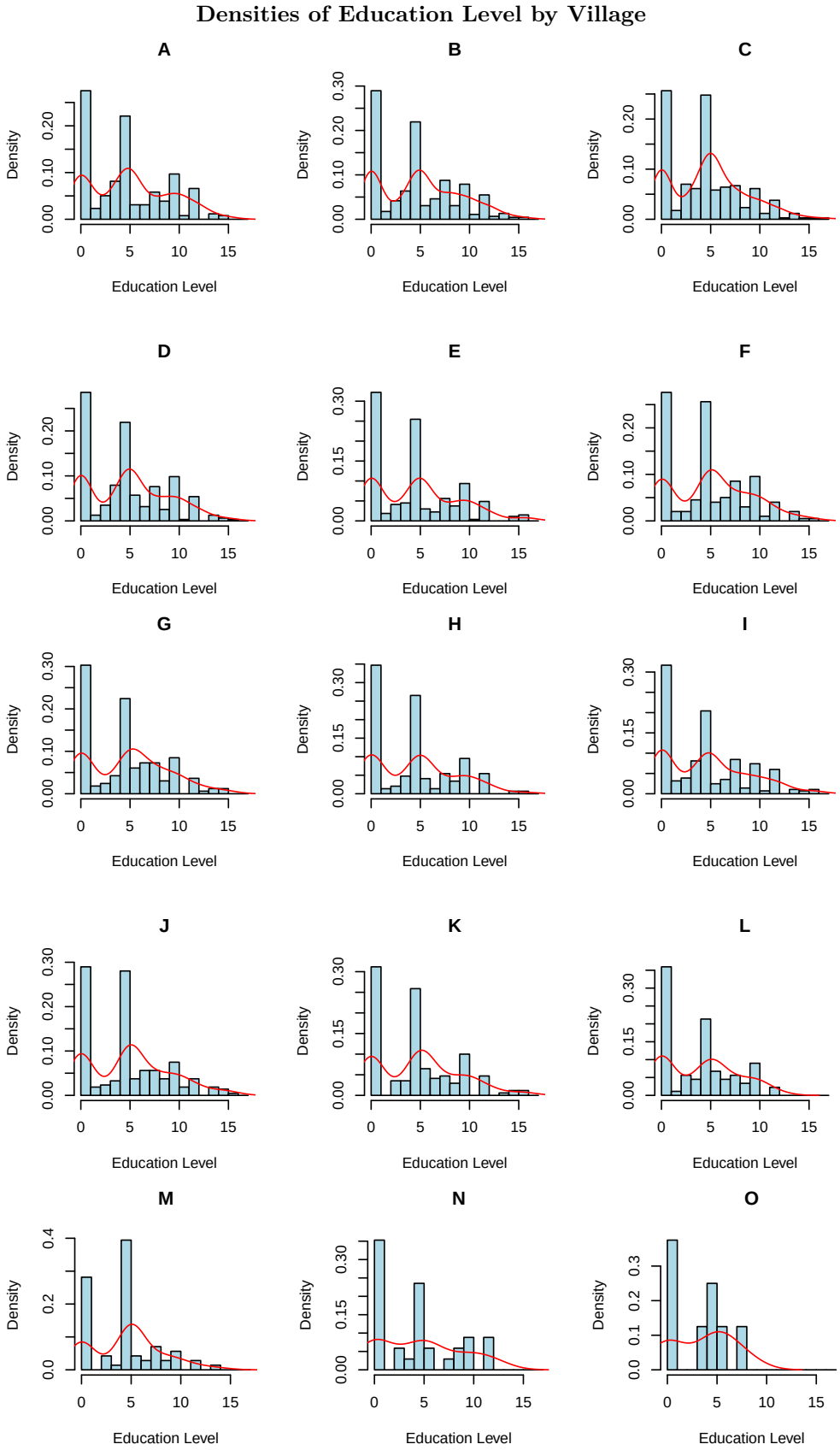

Densities of Education Level by Village
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Density
Education Level

Figure 3: Household Counts by Village

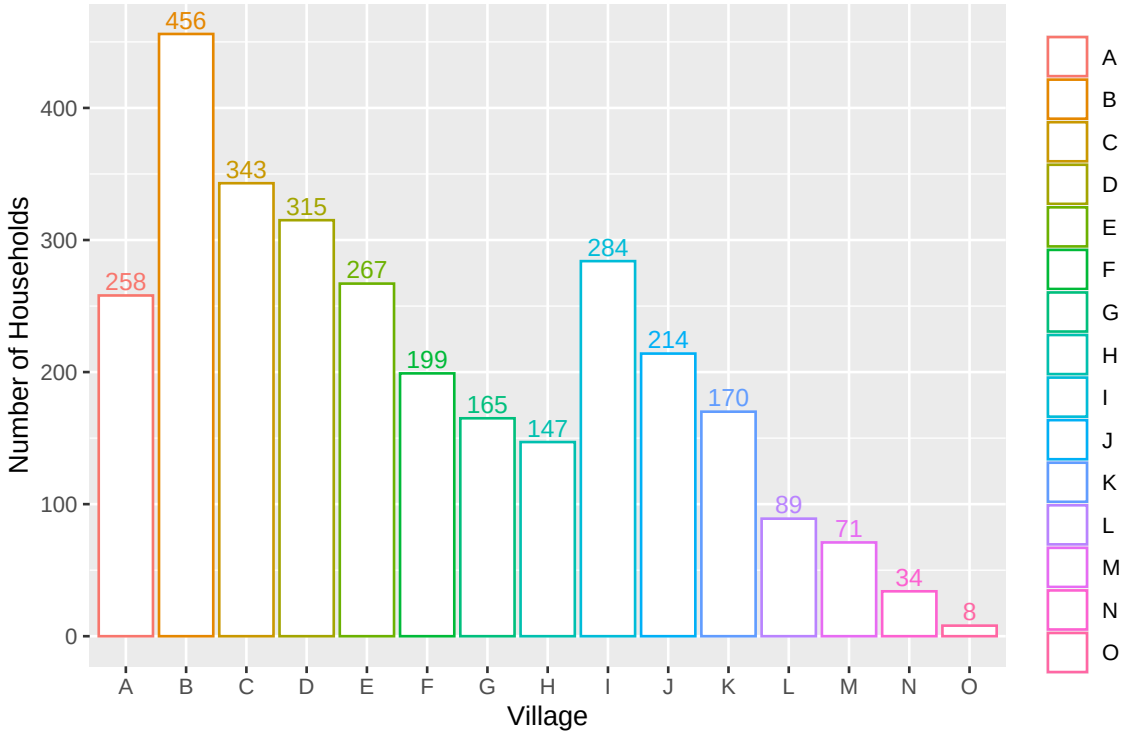


Figure 4: Mean Arsenic Levels by Village

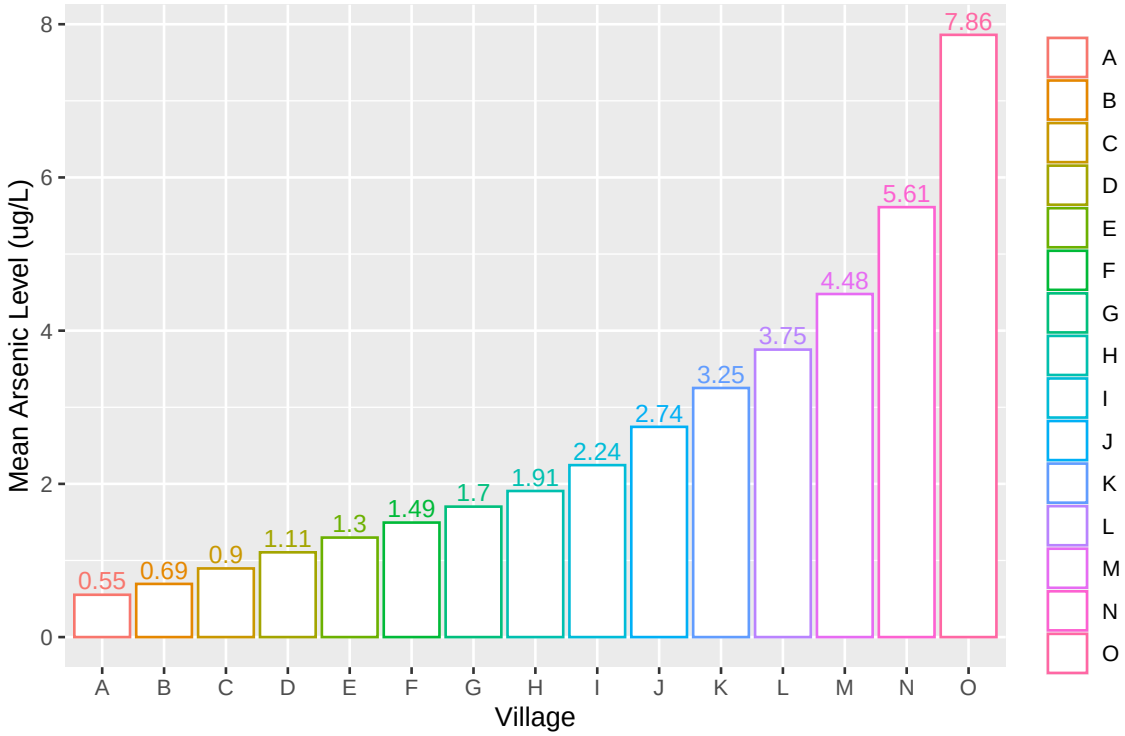


Figure 5: Mean Distance to Safe Well and Proportion Switched by Village

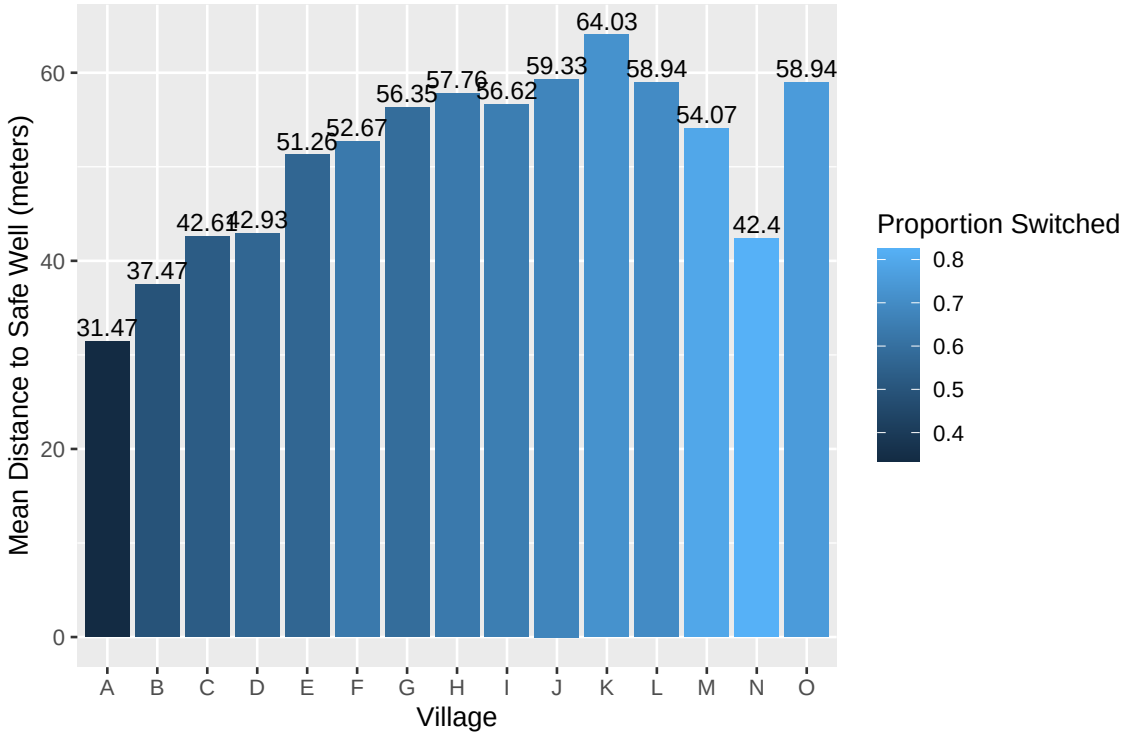


Figure 6: Mean Arsenic Levels by Distance and Switch Group

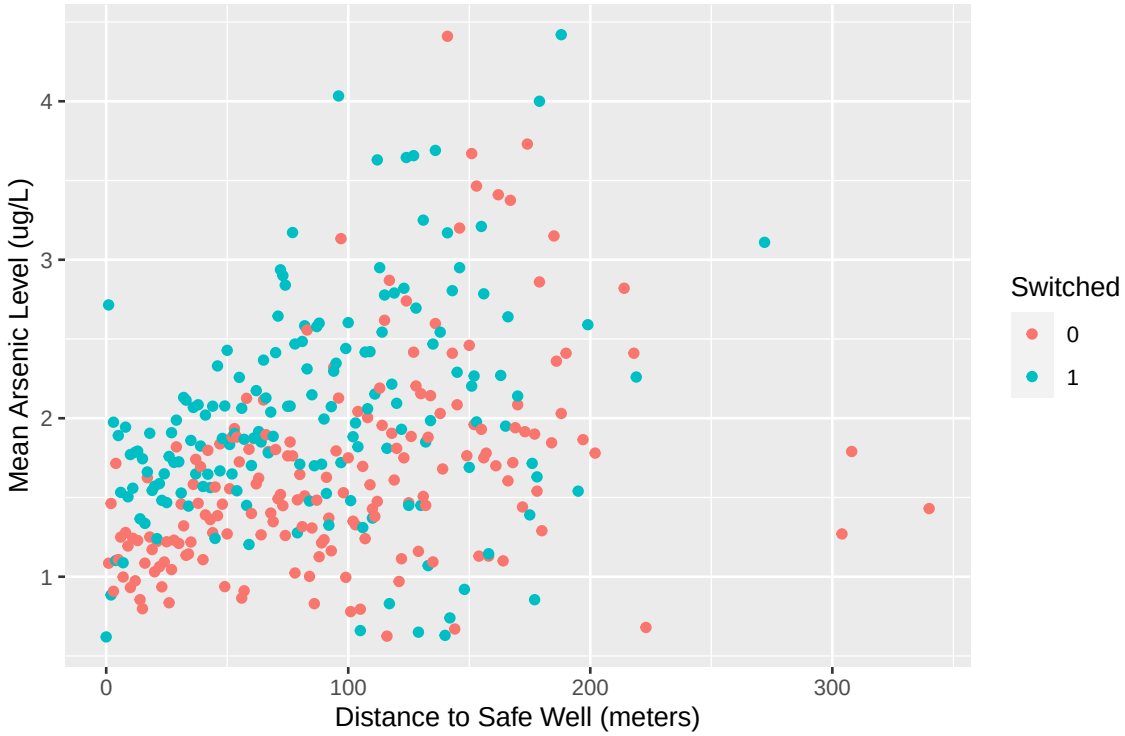


Figure 7: Mean Arsenic Levels by Education Level and Switch Group

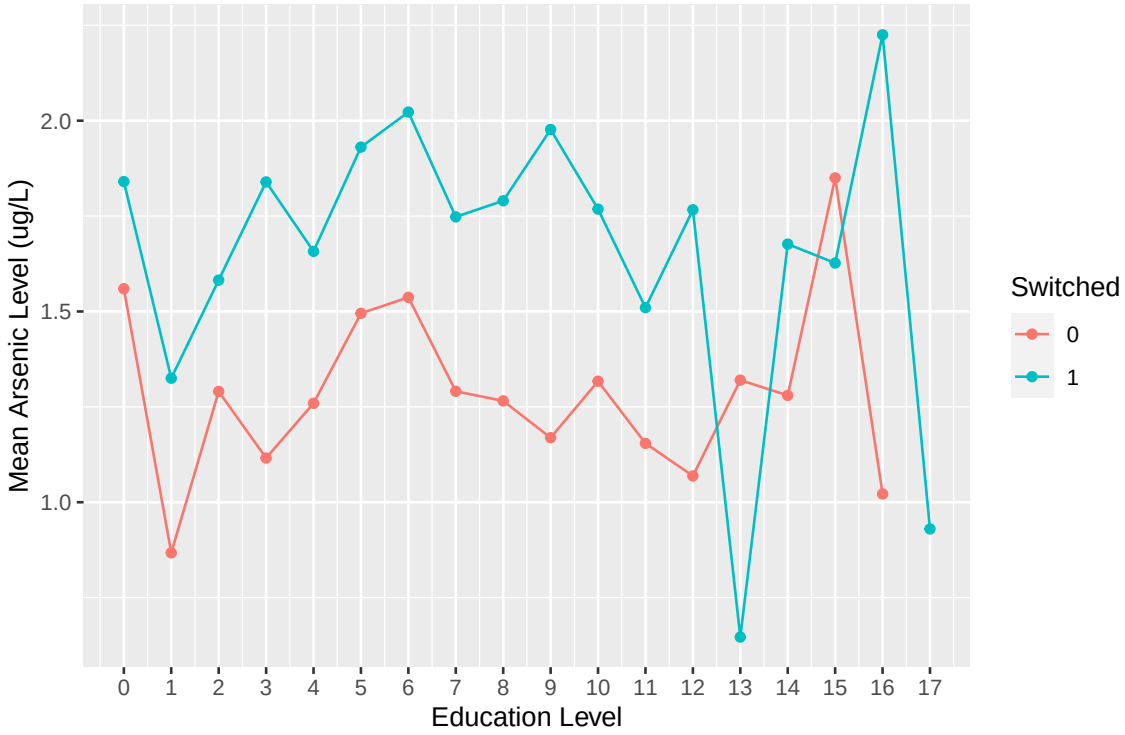
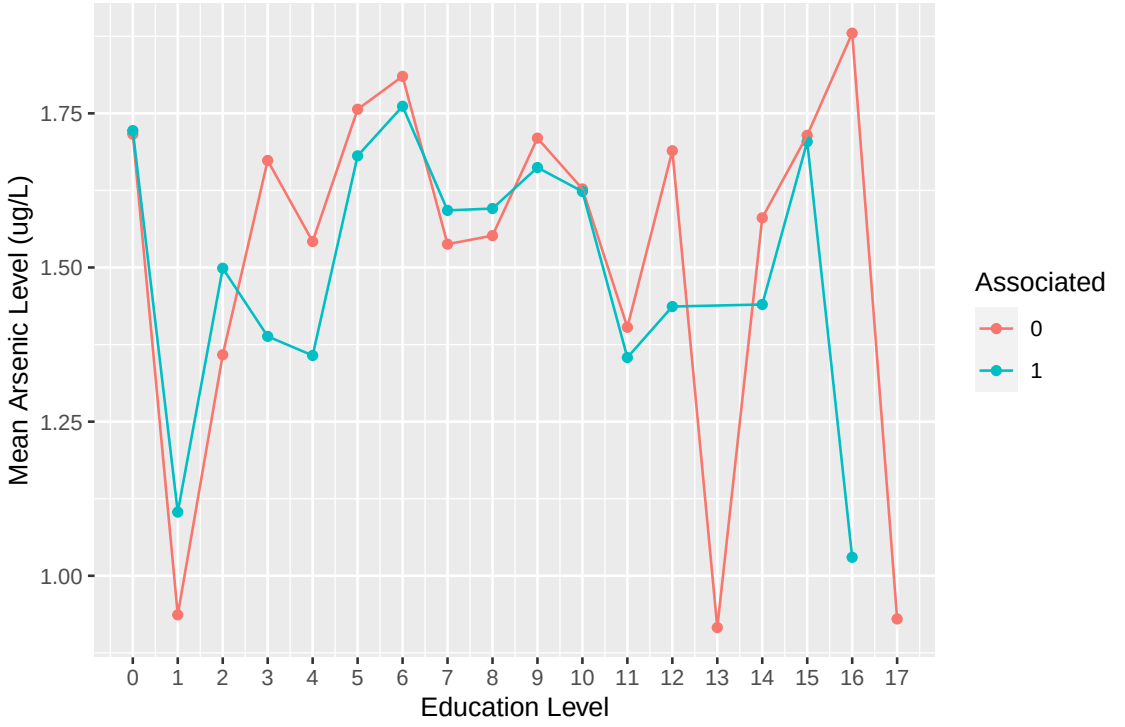


Figure 8: Mean Arsenic Levels by Education Level and Association Group



b. Predicting the probability of switching using arsenic level and log transformation of distance (to the nearest well) with random intercepts across villages.

Multilevel Model 1:

```
glmer(switch ~ arsenic + log(dist) + (1|village),
      family=binomial(link="logit"), data=wells)
```

$$\text{logit}(\text{E}[Y_{ij}|X]) = \beta_{0j} + \beta_1 X_{ij}^{\text{arsenic}} + \beta_2 \log(X_{ij}^{\text{dist}}) + \epsilon_{ij}$$
$$\beta_{0j} = \beta_0 + b_{0j}$$
$$\beta_{0j} \sim \text{N}(\beta_0, \sigma^2_{\beta_0}), \quad b_{0j} \sim \text{N}(0, \sigma^2_{\beta_0})$$

Where:

- Y_{ij} is the predicted outcome (`switch`=1, or `switch`=0) and $\text{E}[Y_{ij}|X] = \mu_Y$ is the probability of switching ($P(Y_{ij} = 1)$) for the i^{th} household in the j^{th} village, such that,
$$\text{logit}(\mu_Y) = \log\left(\frac{\mu_Y}{1-\mu_Y}\right) = \log\Big[\text{odds}\big[P(Y_{ij} = 1)\big]\Big];$$
- β_{0j} is the random (village-specific) intercept, such that $\beta_{0j} = \beta_0 + b_{0j}$, where β_0 is the log-odds that $Y_{ij} = 1$ when all $X_{ij} = 0$ and $b_{0j} = 0$, and b_{0j} is the (random) effect of being in the j^{th} village on the log-odds that $Y_{ij} = 1$;
- β_1 is fixed slope for arsenic level, which gives the effect of a 1-unit increase in X_{ij}^{arsenic} on the log-odds that $Y_{ij} = 1$ for the i^{th} household in the j^{th} village; and
- β_2 is fixed slope for the log transformation of distance to the nearest safe well, which gives the effect of a 1-unit increase in $\log(X_{ij}^{\text{dist}})$ on the log-odds that $Y_{ij} = 1$ for the i^{th} household in the j^{th} village.

Coefficient Estimates:

$$\beta_0 = 0.91, \ \sigma^2_{\beta_0} = (0.22)^2$$
$$\beta_1 = 0.38$$
$$\beta_2 = -0.34$$

Interpretation: An intercept coefficient of 0.91 gives the expected log-odds of switching for any village adjusting for arsenic level and log(distance) to the nearest safe well. Moreover, a (residual) variance of ≈ 0.047, shows us the extent of the between-village variation in the log-odds that $Y_{ij} = 1$ after accounting for arsenic level and log(distance). The remaining fixed-effects coefficients indicate that each unit increase in arsenic level and log(distance) to the nearest safe well multiply the odds of switching by factors of $e^{\beta_1} = 1.462285$ and $e^{\beta_2} = 0.7117703$, respectively, adjusting for other model covariates.

c. Predicting the probability of switching using arsenic level and log transformation of distance (to the nearest well) with varying intercepts and arsenic level coefficients across villages.

Multilevel Model 2:

```
glmer(switch ~  arsenic + log(dist) + (1+arsenic|village),
      family=binomial(link="logit"), data=wells)
```

$$\begin{aligned}
\text{logit}(\text{E}[Y_{ij}|X]) &= \beta_{0j} + \beta_{1j}X_{ij}^{\text{arsenic}} + \beta_2 \log(X_{ij}^{\text{dist}}) + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + b_{0j}, \quad \beta_{0j} \sim \text{N}(\beta_0, \sigma^2_{\beta_0}), \quad b_{0j} \sim \text{N}(0, \sigma^2_{\beta_0}) \\
\beta_{1j} &= \beta_1 + b_{1j}, \quad \beta_{1j} \sim \text{N}(\beta_1, \sigma^2_{\beta_1}), \quad b_{1j} \sim \text{N}(0, \sigma^2_{\beta_1})
\end{aligned}$$

Where:

- β_{1j} is random (village-specific) slope for arsenic level, such that $\beta_{1j} = \beta_1 + b_{1j}$, where β_1 is the mean/expected effect of a 1-unit increase in arsenic level on the log-odds that $Y_{ij} = 1$ across villages, and b_{1j} is the corresponding (random) effect (of a 1-unit increase in X_{ij}^{arsenic}) on the log-odds scale attributed to being in the j^{th} village; and
- all else is as previously stated in part (b).

Coefficient Estimates:

$$\begin{aligned}
\beta_0 = 1.03, \ \sigma^2_{\beta_0} = 0.15 \\
\beta_1 = 0.37, \ \sigma^2_{\beta_1} = 0.021 \\
\beta_2 = -0.35
\end{aligned}$$

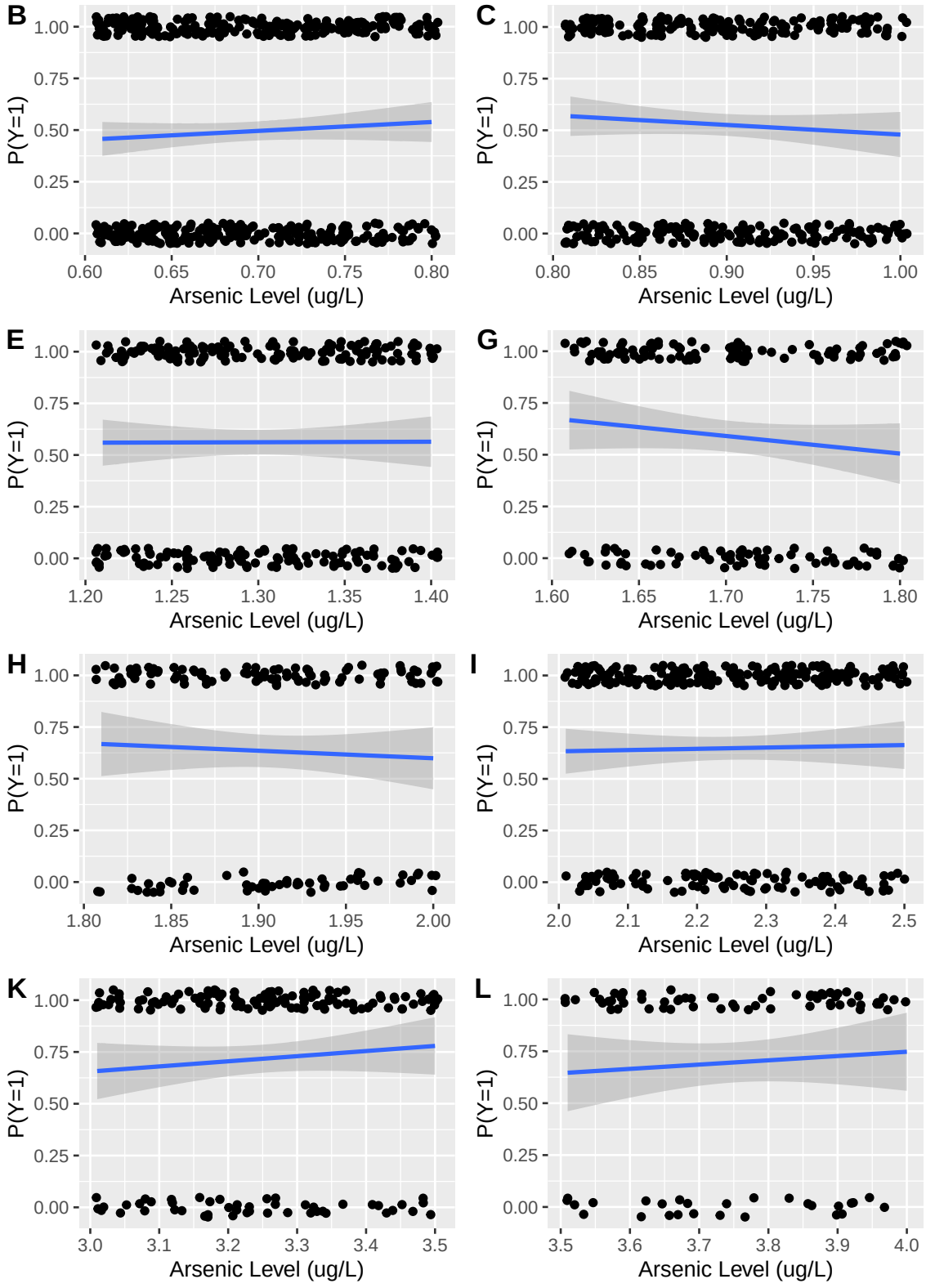
Interpretation: A β_1 coefficient of 0.37 gives the expected change in log-odds of switching for a 1-unit increase in arsenic level, adjusting for log(distance) to the nearest safe well. Moreover, a corresponding (residual) variance of ≈ 0.021, shows us the extent of the between-village variation in this expected change in log-odds of switching for each unit increase in arsenic level, adjusting for log(distance). And, as before, exponentiating this coefficient gives us the factor by which the odds of switching is multiplied for each unit increase in arsenic level, adjusting for log(distance), namely, $e^{\beta_1} = 1.447735$. All else is as previously stated in part (a).

Results: Given the model estimates, it is clear that allowing slope coefficients for arsenic level to vary by village has very minimal impact on the fixed effects (see Table 3 below). Moreover, the small β_1 residual variance ($\sigma^2_{\beta_1} = 0.021$) indicates that there is little between-village variation with respect to the effect of arsenic level on the log-odds of switching (and hence, minimal deviance from the mean arsenic level effect across villages).

Table 3: Fixed Effects Comparison

	(Intercept)	arsenic	log(dist)
Model 1	0.9132	0.3804	-0.3411
Model 2	1.0292	0.3671	-0.3458

d. Graphs to illustrate the probability of switching as a function of arsenic level for a randomly selected set of eight villages (of fifteen total).



e. Comparing the fit of models 1 and 2 from parts (b) and (c), respectively.

Model Estimation Comparisons[3]:

Table 4: Estimated Coefficients

	Model 1			Model 2		
	(Intercept)	arsenic	log(dist)	(Intercept)	arsenic	log(dist)
A	0.369	0.38	-0.341	0.127	0.707	-0.346
B	0.856	0.38	-0.341	0.831	0.442	-0.346
C	0.946	0.38	-0.341	0.976	0.387	-0.346
D	0.994	0.38	-0.341	1.048	0.360	-0.346
E	0.980	0.38	-0.341	1.040	0.363	-0.346
F	1.131	0.38	-0.341	1.326	0.255	-0.346
G	0.964	0.38	-0.341	1.050	0.359	-0.346
H	1.035	0.38	-0.341	1.165	0.316	-0.346
I	1.001	0.38	-0.341	1.090	0.344	-0.346
J	0.967	0.38	-0.341	1.107	0.338	-0.346
K	0.995	0.38	-0.341	0.935	0.402	-0.346
L	0.816	0.38	-0.341	1.218	0.296	-0.346
M	0.917	0.38	-0.341	1.075	0.350	-0.346
N	0.863	0.38	-0.341	1.189	0.307	-0.346
O	0.849	0.38	-0.341	1.309	0.261	-0.346

Table 5: Random Effect Variabilities

	Variance	Std.Dev.	Covariance	Correlation
Model 1				
(Intercept)	0.046	0.215	NA	NA
Model 2				
(Intercept)	0.150	0.388	NA	NA
arsenic	0.021	0.146	-0.057	-1

[3]Aside from fixed effects, which can be observed in Table 3.

Model Predictions and GoF Results:

Table 6: Observed vs. Predicted (First 10)

		Model 1 Predictions		Model 2 Predictions	
id	Observed	Probability	Value	Probability	Value
1	1	0.7182	1	0.7163	1
2	1	0.4528	0	0.4529	0
3	0	0.6792	1	0.6792	1
4	1	0.5952	1	0.599	1
5	1	0.5367	1	0.5402	1
6	1	0.7013	1	0.7121	1
7	1	0.6456	1	0.6438	1
8	1	0.7026	1	0.7011	1
9	1	0.7091	1	0.7078	1
10	1	0.6114	1	0.6142	1

Note:
Model 1 had 1,891/3,020 correct predictions.
Model 2 had 1,888/3,020 correct predictions.

Table 7: Likelihood Ratio Test

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Model 1	4	3956.686	3980.738	-1974.343	3948.686	NA	NA	NA
Model 2	6	3953.957	3990.035	-1970.978	3941.957	6.729	2	0.035

Note:
Assuming Model 1 is nested in Model 2.
Where the null hypothesis is that both models fit the data equally well.

Discussion: In line with the observations made regarding the second model, the random effects variabilities and coefficient estimates displayed here show us that there is little variation between villages with respect to the effect of arsenic level on the log-odds of switching. Moreover, the model predictions suggest that the first model (from part (b)) provided a slightly better fit to the data, producing probabilities closer to the true values and 3 more correct predictions than the second model in which we allow arsenic level slopes to vary by village. This result contradicts the results form the LRT in Table 7, which suggest that Model 2 provides a more accurate fit to the data (based on our rejection of the null hypothesis on the grounds that $p < 0.05$). However, it is important to note that the second model resulted in a singular fit (`boundary (singular) fit: see help('isSingular')`), which we can attribute to the correlation of -1 between random effects seen in Table 5. According to the R Documentation, this may be a due to the model assumming an overly complex random effects structure, which results in multicollinearity and overfitting. To get a better sense for the reasoning behind this sigularity pattern, we apply the `rePCA` function to the model, which performs PCA on a fitted mixed-effects model's random effects variance-covariance estimates to give the total variance explained by the model's random effects:

```
Importance of components:
                          [,1] [,2]
Standard deviation     0.4142    0
Proportion of Variance 1.0000    0
Cumulative Proportion  1.0000    1
```

This result shows us that 100% of the variance can be explained by the first principle component (the intercept random effect), form which we deduce that the second model suffers from overfitting and assumes a random effects structure that is too complex for the data available. Thus, according to the R Documentation, a standard inferential procedure such as a LRT may not only be inappropriate, but produce misleading results. For this reason, in addition to the remaining gathered evidence, it is safe to assume that Model 1 provides a superior fit over the second model from part (c).

Code

```
## Importing Data

wells <- read.csv("/Users/antonellabasso/Desktop/PHP2517/DATA/wells.vill.csv")
wells
```

```
## Descriptive Statistics

str(wells)
nrow(wells) # 3020 observations/households/wells
apply(wells, 2, function(x) length(unique(x))) # 15 villages, 18 education levels

# missing data
wells_na <- apply(wells, 2, function(x) sum(is.na(x)))
wells_na[wells_na != 0] # no missing values of any kind

# general info
sum(wells$switch) # 1737 switches made (57.52%)
sum(wells$assoc) # 1277 households active in community org. (42.29%)
sort(unique(wells$educ)) # education levels range between 0-17
range(wells$arsenic) # arsenic levels range between 0.51-9.65 ug/L
range(wells$dist) # distances range between 0.387-339.531 meters

# variable summary
CreateTableOne(data=wells)

# number of observations for each village
wells_village <- as.data.frame(table(wells$village)) %>%
  rename("village"=Var1, "observations"=Freq)

# number of observations for each education level
wells_educ <- as.data.frame(table(wells$educ)) %>%
  rename("education"=Var1, "observations"=Freq)

# mode function
getmode <- function(x) {
  unique_x <- unique(x)
  unique_x[which.max(tabulate(match(x, unique_x)))]
}

# descriptive statistics wrt arsenic level and probability of switching by village
ds_village <- wells %>%
  group_by(village) %>%
  summarise(households=n(),
            switch1=sum(switch),
            switch_prob=mean(switch),
            mean_ars=mean(arsenic),
            var_ars=var(arsenic),
            mean_dist=mean(dist),
            var_dist=var(dist),
            assoc1=sum(assoc),
            assoc_prob=mean(assoc),
            mode_educ=getmode(educ)) #%>%
```

```
  #mutate(num_individuals=count/observations)

# descriptive statistics wrt arsenic level by switch group
ds_switch <- wells %>%
  group_by(switch) %>%
  summarise(households=n(),
            mean_ars=mean(arsenic),
            var_ars=var(arsenic),
            mean_dist=mean(dist),
            var_dist=var(dist),
            assoc1=sum(assoc),
            assoc_prob=mean(assoc),
            mode_educ=getmode(educ))

# education levels by village
educ_village <- wells %>%
  group_by(village, educ) %>%
  summarise(count=n()) %>%
  full_join(ds_village[, 1:2]) %>%
  summarise(educ=educ, count=count, prop=count/households) %>%
  rename("households"=count)

# factorizing categorical variables
#cd4$id <- as.factor(cd4$id)
#cd4$trt <- as.factor(cd4$trt)
```

```
## EDA Tables

# descriptive statistics with respect switch group
dstats1 <- ds_switch %>%
  rename("Switch Group"=switch,
         "Households"=households,
         "Mean Arsenic"=mean_ars,
         "Arsenic Var"=var_ars,
         "Mean Distance"=mean_dist,
         "Distance Var"=var_dist,
         "Assoc."=assoc1,
         "Prop. Assoc."=assoc_prob,
         "Mode Education"=mode_educ)

# descriptive statistics by village (group-level)
dstats2 <- ds_village[, -c(11)] %>%
  rename("Village"=village,
         "Households"=households,
         "Switched"=switch1,
         "Prop. Switched"=switch_prob,
         "Mean Arsenic"=mean_ars,
         "Arsenic Var"=var_ars,
         "Mean Distance"=mean_dist,
         "Distance Var"=var_dist,
         "Assoc."=assoc1,
         "Prop. Assoc."=assoc_prob)
dstats2[, c(4:8, 10)] <- lapply(dstats2[, c(4:8, 10)], function(x) round(x, 4))
```

```
dstats2
```

```
## Density Plots (Histograms)

# Arsenic Level Density
hist(wells$arsenic,
     probability=TRUE, ylim=c(0, 0.8), col="lavender", breaks=20,
     main="Figue 1: Arsenic Level Density", xlab="Arsenic Level (ug/L)")
lines(density(wells$arsenic), col="red")
```

```
# Distance to Safe Well Density
hist(wells$dist,
     probability=TRUE, ylim=c(0, 0.02), col="thistle2", breaks=20,
     main="Figure 2: Distance to Safe Well Density", xlab="Distance to Safe Well (meters)")
lines(density(wells$dist), col="red")
```

```
# Densities of Education Level by Village
par(mfrow=c(2, 3))
colnames <- LETTERS[1:15]
for (i in colnames) {
    hist(wells[which(wells$village==i), "educ"],
         probability=TRUE, col="lightblue",
         xlim=c(0, 17), breaks=seq(0, 17, 1),
         main=i, xlab="Education Level")
    dens <- density(wells[which(wells$village==i), "educ"])
    lines(dens, col="red")
}
```

```
## EDA Plots

# Household Counts by Village
p1 <- ggplot(ds_village, aes(x=village, y=households, color=village)) +
  geom_bar(stat="identity", fill="white") +
  geom_text(aes(label=households),
            vjust=-0.3, size=3.5, show.legend=FALSE) +
  labs(title="Figure 3: Household Counts by Village",
       x="Village",
       y="Number of Households",
       color="",
       fill="")

# Mean Arsenic Levels by Village
p2 <- ggplot(ds_village, aes(x=village, y=mean_ars, color=village)) +
  geom_bar(stat="identity", fill="white") +
  geom_text(aes(label=round(mean_ars, 2)),
            vjust=-0.3, size=3.5, show.legend=FALSE) +
  labs(title="Figure 4: Mean Arsenic Levels by Village",
       x="Village",
       y="Mean Arsenic Level (ug/L)",
       color="",
       fill="")

# Spread of Arsenic Levels by Village
p3 <- ggplot(wells, aes(x=village, y=arsenic, color=village)) +
  geom_boxplot() +
```

```
  labs(title="",
       x="Village",
       y="Arsenic Level (ug/L)",
       color="",
       fill="")

# Mean Distance to Safe Well and Proportion Switched by Village
p4 <- ggplot(ds_village, aes(x=village, y=mean_dist)) +
  geom_bar(stat="identity", aes(fill=switch_prob)) +
  geom_text(aes(label=round(mean_dist, 2)),
            vjust=-0.3, size=3.5, show.legend=FALSE) +
  labs(title="Figure 5: Mean Distance to Safe Well and Proportion Switched by Village",
       x="Village",
       y="Mean Distance to Safe Well (meters)",
       color="",
       fill="Proportion Switched")

# Proportion Switched and Mean Distance to Safe Well by Village
p4b <- ggplot(ds_village, aes(x=village, y=switch_prob)) +
  geom_bar(stat="identity", aes(fill=mean_dist)) +
  geom_text(aes(label=round(switch_prob, 2)),
            vjust=-0.3, size=3.5, show.legend=FALSE) +
  labs(title="Proportion Switched and Mean Distance to Safe Well by Village",
       x="Village",
       y="Proportion Switched",
       color="",
       fill="Mean Distance")

# Mean Arsenic Levels by Distance and Switch Group
mean_arsenic <- wells %>%
  group_by(round_dist=round(dist), switch) %>%
  summarise(mean=mean(arsenic), .groups="keep")
p5 <- ggplot(mean_arsenic, aes(x=round_dist, y=mean, color=as.factor(switch))) +
  geom_point() +
  labs(title="Figure 6: Mean Arsenic Levels by Distance and Switch Group",
       x="Distance to Safe Well (meters)",
       y="Mean Arsenic Level (ug/L)",
       color="Switched")

# Mean Arsenic Levels by Education Level and Switch Group
mean_arsenic2 <- wells %>%
  group_by(educ, switch) %>%
  summarise(mean=mean(arsenic), .groups="keep")
p6 <- ggplot(mean_arsenic2, aes(x=educ, y=mean, color=as.factor(switch))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks=seq(0, 17, 1)) +
  labs(title="Figure 7: Mean Arsenic Levels by Education Level and Switch Group",
       x="Education Level",
       y="Mean Arsenic Level (ug/L)",
       color="Switched")

# Mean Arsenic Levels by Education Level and Assoc. Group
```

```
mean_arsenic3 <- wells %>%
  group_by(educ, assoc) %>%
  summarise(mean=mean(arsenic), .groups="keep")
p7 <- ggplot(mean_arsenic3, aes(x=educ, y=mean, color=as.factor(assoc))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks=seq(0, 17, 1)) +
  labs(title="Figure 8: Mean Arsenic Levels by Education Level and Association Group",
       x="Education Level",
       y="Mean Arsenic Level (ug/L)",
       color="Associated")
```

```
## MODEL 1
m1 <- glmer(switch ~ arsenic + log(dist) + (1|village),
            family=binomial(link="logit"), data=wells)

summary(m1)
display(m1)
fixef(m1)
coef(m1)
```

```
m1_coef <- coef(m1)$village
b_0 <- mean(m1_coef[,1])
sd(b_0-m1_coef[,1])
mean(m1_coef[,1])
var(m1_coef[,1])
#se.coef(m1)
#se.fixef(m1)
#se.ranef(m1)
```

```
## MODEL 2
m2 <- glmer(switch ~ arsenic + log(dist) + (1+arsenic|village),
            family=binomial(link="logit"), data=wells)

summary(m2)
display(m2)
fixef(m2)
coef(m2)
```

```
# comparing fixed effects between models
models <-  c("Model 1", "Model 2")
comp_fe <- as.data.frame(rbind(fixef(m1), fixef(m2))) # fixed effects
rownames(comp_fe) <- models
```

```
## Randomly Selected 8 Villages
set.seed(47)
villages_8 <- sample(unique(wells$village), 8, replace=FALSE)

## Probability Plots
prob_plot1 <- ggplot(wells[which(wells$village==villages_8[1]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")
```

```
prob_plot2 <- ggplot(wells[which(wells$village==villages_8[2]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

prob_plot3 <- ggplot(wells[which(wells$village==villages_8[3]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

prob_plot4 <- ggplot(wells[which(wells$village==villages_8[4]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

prob_plot5 <- ggplot(wells[which(wells$village==villages_8[5]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

prob_plot6 <- ggplot(wells[which(wells$village==villages_8[6]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

prob_plot7 <- ggplot(wells[which(wells$village==villages_8[7]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

prob_plot8 <- ggplot(wells[which(wells$village==villages_8[8]), ],
                     aes(x=arsenic, y=switch)) +
  geom_jitter(position=position_jitter(height=0.05)) +
  stat_smooth(method="glm", family="binomial") +
  labs(x="Arsenic Level (ug/L)", y="P(Y=1)")

plot_grid(prob_plot3, prob_plot2, prob_plot6, prob_plot7,
          ncol=2, labels=villages_8[c(3, 2, 6, 7)])

plot_grid(prob_plot1, prob_plot5, prob_plot8, prob_plot4,
          ncol=2, labels=villages_8[c(1, 5, 8, 4)])
```

```
## Model Comparison Tables

models <-  c("Model 1", "Model 2")

# model estimations
comp_fe <- as.data.frame(rbind(fixef(m1), fixef(m2))) # fixed effects
```

```r
comp_re <- cbind(ranef(m1)$village, ranef(m2)$village) # random effects
comp_coef <- cbind(coef(m1)$village, coef(m2)$village) # all estimated coefficients
comp_aic <- as.data.frame(rbind(summary(m1)$AIC, summary(m2)$AIC)) # performance metrics

# variabilities
VarCorr(m1)
VarCorr(m2)
rnames <- c("(Intercept) ", "(Intercept)", "arsenic")
Variance <- c(0.04637038, 0.15024372, 0.02132066)
Std.Dev. <- c(0.2153378, 0.3876128, 0.1460159)
Covariance <- c(NA, NA, -0.05659766)
Correlation <- c(NA, NA, -1.0000000)
comp_varcor <- as.data.frame(cbind(Variance, Std.Dev., Covariance, Correlation))
rownames(comp_varcor) <- rnames

# model predictions
pred_m1 <- as.vector(fitted(m1)) # model 1
pred_m2 <- as.vector(fitted(m2)) # model 2
obsvpred <- as.data.frame(cbind(id=wells$id,
                                village=wells$village,
                                arsenic=wells$arsenic,
                                observed=wells$switch,
                                pred_m1=round(pred_m1, 4),
                                pred2_m1=round(pred_m1),
                                pred_m2=round(pred_m2, 4),
                                pred2_m2=round(pred_m2)))
obsvpred_t <- obsvpred[1:10, c(1, 4:8)] %>%
  rename("Observed"=observed,
         "Probability"=pred_m1, "Value"=pred2_m1,
         "Probability "=pred_m2, "Value "=pred2_m2)
sum(obsvpred$observed!=obsvpred$pred2_m1) # model 1: 1129 incorrect predictions
sum(obsvpred$observed!=obsvpred$pred2_m2) # model 2: 1132 incorrect predictions
sum(obsvpred$observed==obsvpred$pred2_m1) # model 1: 1891 correct predictions of 3020
sum(obsvpred$observed==obsvpred$pred2_m2) # model 2: 1888 correct predictions of 3020

# GoF/LRT (model 1 nested in model 2)
anova(m1, m2) # m2 is a better fit (test invalid due to singularity of m2)
comp_LRT <- as.data.frame(anova(m1, m2))

rownames(comp_fe) <- models
rownames(comp_aic) <- models
rownames(comp_LRT) <- models
```