Figure 5.7 *Wells in an area of Araihazar upazila, Bangladesh. Light and dark dots represent wells with arsenic greater than and less than the safety standard of 0.5 (in units of hundreds of micrograms per liter). (The wells are located where people live. The empty areas between the wells are mostly cropland.) Safe and unsafe wells are intermingled in most of the area, which suggests that users of unsafe wells can switch to nearby safe wells.*

Background

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure.

Any locality can include wells with a range of arsenic levels, as can be seen from the map in Figure 5.7 of all the wells in a collection of villages in a small area of Bangladesh. The bad news is that even if your neighbor's well is safe, it does not mean that yours is safe. However, the corresponding good news is that, if your well has a high arsenic level, you can probably find a safe well nearby to get your water from—if you are willing to walk the distance and your neighbor is willing to share. (The amount of water needed for drinking is low enough that adding users to a well would not exhaust its capacity, and the surface water in this area is subject to contamination by microbes, hence the desire to use water from deep wells.)

In the area shown in Figure 5.7, a research team from the United States and Bangladesh measured all the wells and labeled them with their arsenic level as well as a characterization as "safe" (below 0.5 in units of hundreds of micrograms per liter, the Bangladesh standard for arsenic in drinking water) or "unsafe" (above 0.5). People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction.

A few years later, the researchers returned to find out who had switched wells. We shall perform a logistic regression analysis to understand the factors predictive of well switching among the users of unsafe wells. In the notation of the previous section, our outcome variable is

$$y_i = \begin{cases} 1 & \text{if household } i \text{ switched to a new well} \\ 0 & \text{if household } i \text{ continued using its own well.} \end{cases}$$
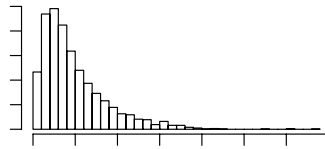
Figure 5.8 *Histogram of distance to the nearest safe well, for each of the unsafe wells in the Araihazar dataset (see Figure 5.7).*

We consider the following inputs:

- A constant term
- The distance (in meters) to the closest known safe well
- The arsenic level of respondent's well
- Whether any members of the household are active in community organizations
- The education level of the head of household.

We shall first fit the model just using distance to nearest well and then put in arsenic concentration, organizational membership, and education.

Logistic regression with just one predictor

We fit the logistic regression in R:

R code
```
fit.1 <- glm (switch ~ dist, family=binomial(link="logit"))
```

Displaying this yields

R output
```
glm(formula = switch ~ dist, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)   0.6060  0.0603
dist         -0.0062  0.0010
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

The coefficient for `dist` is -0.0062, which seems low, but this is misleading since distance is measured in meters, so this coefficient corresponds to the difference between, say, a house that is 90 meters away from the nearest safe well and a house that is 91 meters away.

Figure 5.8 shows the distribution of `dist` in the data. It seems more reasonable to rescale distance in 100-meter units:

R code
```
dist100 <- dist/100
```

and refitting the logistic regression yields

R output
```
glm(formula = switch ~ dist100, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)    0.61    0.06
dist100       -0.62    0.10
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```