IDS 702: Module 2.5

Logistic regression with multiple predictors I

Dr. Olanrewaju Michael Akande

Logistic regression with multiple predictors: motivating example

- In many developing countries, people get their drinking water from wells.
- Sometimes these wells are contaminated with the chemical arsenic, which when consumed in high concentrations causes skin and bladder cancer, as well as cardiovascular disease.
- Fortunately, in many cases people living near contaminated wells have the opportunity to get water from nearby uncontaminated wells.

The contaminated wells analysis

- In one study, several researchers measured the concentrations of arsenic in wells in a particular region of Bangladesh.
- They labeled wells as safe or unsafe based on the measurements.
- The researchers encouraged people drinking from unsafe wells to switch to safe wells.
- Several years later, the researchers returned to the area with the goal of seeing who had switched from unsafe to safe wells.
- They recorded information on a sample of 3020 individuals who had wells at their homes that were unsafe.
- Let's address the question: what predicts why people switch wells?
- The data is in the file `arsenic.csv` on Sakai.

The contaminated wells analysis

Data description

Variable	Description
Switch	1 = if respondent switched to a safe well 0 = if still using own unsafe well
Arsenic	amount of arsenic in well at respondent's home (100s of micro-grams per liter)
Dist	distance in meters to the nearest known safe well
Assoc	1 = if any members of household are active in community organizations 0 = otherwise
Educ	years of schooling of the head of household

Treat `switch` as the response variable and others as potential predictors.

Logistic regression with multiple predictors

- We can then formally extend the logistic regression model we had before to allow for multiple predictors.
- We still have

$$\Pr[y_i = 1|x_i] = \pi_i \text{ and } \Pr[y_i = 0|x_i] = 1 - \pi_i,$$

or

$$y_i|x_i \sim \text{Bernoulli}(\pi_i)$$

as before, but with

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

now in both cases.

- Let's fit the model to our motivating example.

The contaminated wells analysis: EDA

```
arsenic <- read.csv("data/arsenic.csv",header=T,
                    colClasses=c("numeric","numeric","numeric","factor","numeric"))
head(arsenic)
```

```
##   switch arsenic   dist assoc educ
## 1      1    2.36 16.826     0    0
## 2      1    0.71 47.322     0    0
## 3      0    2.07 20.967     0   10
## 4      1    1.15 21.486     0   12
## 5      1    1.10 40.874     1   14
## 6      1    3.90 69.518     1    9
```

```
summary(arsenic[,-1])
```

```
##     arsenic           dist          assoc         educ
##  Min.   :0.510   Min.   :  0.387   0:1743   Min.   : 0.000
##  1st Qu.:0.820   1st Qu.: 21.117   1:1277   1st Qu.: 0.000
##  Median :1.300   Median : 36.761            Median : 5.000
##  Mean   :1.657   Mean   : 48.332            Mean   : 4.828
##  3rd Qu.:2.200   3rd Qu.: 64.041            3rd Qu.: 8.000
##  Max.   :9.650   Max.   :339.531            Max.   :17.000
```

```
table(arsenic$switch)
```

```
##
##    0    1
## 1283 1737
```

Move to the R script here.

What's next?

Move on to the readings for the next module!