Lab #10 - Logistic Regression Part II

Econ 224

September 27th, 2018

Contaminated Wells in Bangladesh

Today we'll work with a dataset containing household-level information from Bangladesh: `wells.csv`. You can download the dataset from the course website at http://ditraglia.com/econ224/wells.csv.

Here is some background on the dataset from Gelman and Hill (2007):

> Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic ... a research team from the United States and Bangladesh measured all the wells [in a small region] and labeled them with their arsenic level as well as a characterization of "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.

Our goal is to predict which households will switch wells using the following information:

Name	Description
`dist`	Distance to closest known safe well (meters)
`arsenic`	Arsenic level of respondent's well (100s of micrograms/liter)
`switch`	Dummy variable: equals 1 if switched to a new well
`assoc`	Dummy variable: equals 1 if any member of the household is active in community organizations
`educ`	Education level of head of household (years)

To be clear, our dataset contains *only* information for households with an arsenic level of 0.5 or above, as these are the households that were encouraged to switch wells.

Exercises

1. Preliminaries:

   (a) Load the data and store it in a tibble called `wells`.
   (b) Use `ggplot2` to make a histogram of `arsenic`. Be sure to label your plot appropriately. Comment on your findings.
   (c) Create a variable called `dist100` that contains the same information as `dist` but measured in *hundreds of meters* rather than in meters.
   (d) Use `ggplot2` to make a histogram of `dist100`. Be sure to label your plot appropriately. Comment on your findings.

2. First Regression: `fit1`

(a) Run a logistic regression using `dist100` to predict `switch` and store the result in an object called `fit1`.
(b) Use `ggplot2` to plot the logistic regression function from `fit1` along with the data, jittered appropriately.
(c) Discuss your results from parts (a) and (b). In particular: based on `fit1`, is `dist100` a statistically significant predictor of `switch`? Does the sign of its coefficient make sense? Explain.
(d) Based on `fit1`, calculate the predicted probability of switching wells for a household for which the nearest safe well is zero meters away.
(e) Based on `fit1`, calculate the predicted probability of switching wells for a household for which the nearest safe well is $\overline{\text{dist100}}$ hundred meters away, where a is the sample mean of `dist100`.

3. Second Regression: `fit2`
4. Error Rates:

Solutions

1-2: Load Data, Histogram of `arsenic`

```
#------- Load data
library(tidyverse)
library(ggplot2)
wells <- read_csv('~/econ224/labs/wells.csv')

#------- Histogram of arsenic
ggplot(wells) +
  geom_histogram(aes(x = arsenic), binwidth = 0.2) +
  xlab('Arsenic Concentration (100s of micrograms/liter)') +
  ggtitle('Arsenic Concentrations in Unsafe Wells')
```
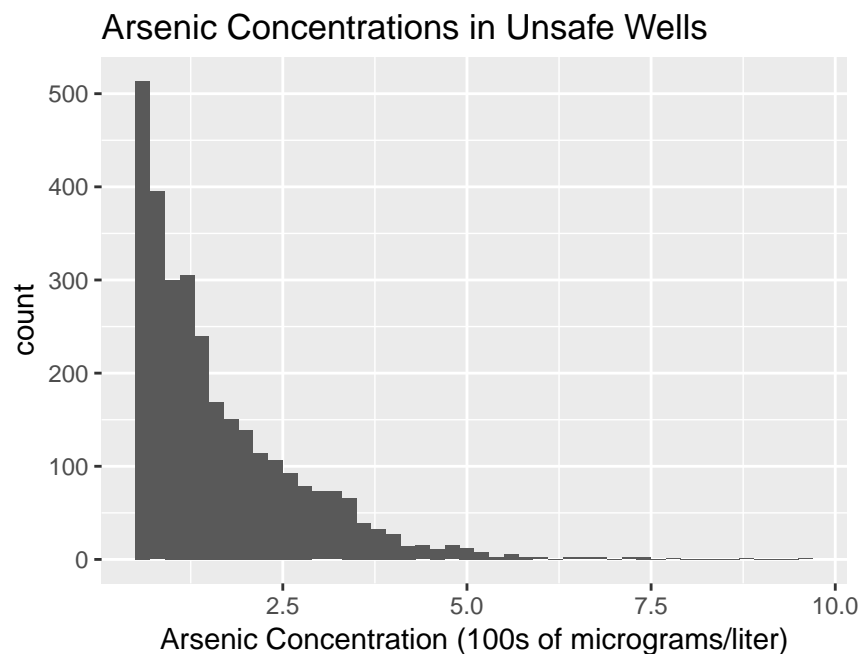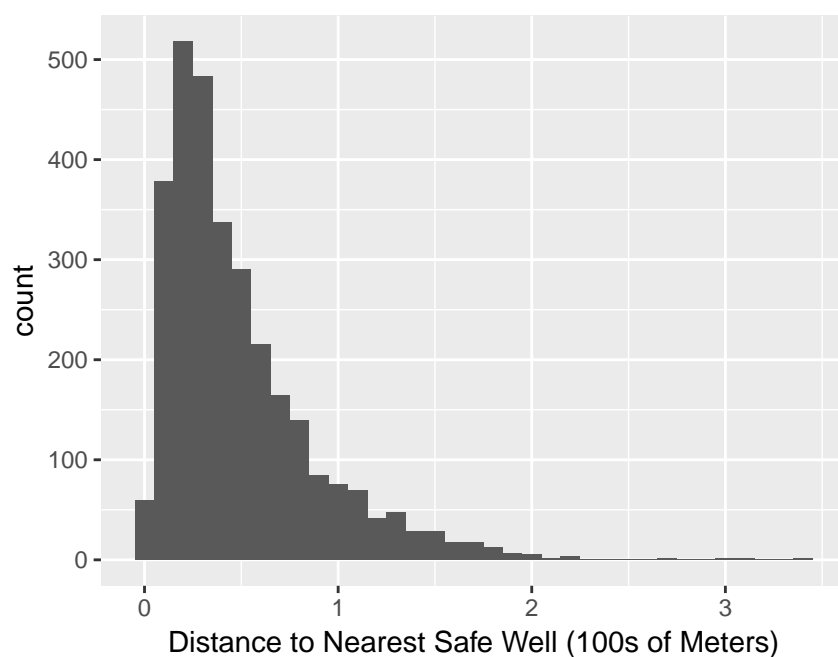
3-4: Histogram of `dist100`

```
#-------- Create dist100
wells <- wells %>%
  mutate(dist100 = dist / 100)

#-------- Plot histogram of dist100
ggplot(wells) +
  geom_histogram(aes(x = dist100), binwidth = 0.1) +
  xlab('Distance to Nearest Safe Well (100s of Meters)')
```


5-7: Generate, plot and interpret `fit1`

```
#---------- Generate and summarize fit1
fit1 <- glm(switch ~ dist100, family = binomial(link = 'logit'), wells)
summary(fit1)
```

```
Call:
glm(formula = switch ~ dist100, family = binomial(link = "logit"),
    data = wells)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4406  -1.3058   0.9669   1.0308   1.6603

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.60596    0.06031  10.047  < 2e-16 ***
```

```
dist100      -0.62188     0.09743  -6.383 1.74e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 4076.2  on 3018  degrees of freedom
AIC: 4080.2

Number of Fisher Scoring iterations: 4
```

```r
#---------- Plot fit1 with jittering
ggplot(wells, aes(x = dist100, y = switch)) +
  geom_jitter(height = 0.1) +
  stat_smooth(method='glm',
              method.args = list(family = "binomial"),
              formula = y ~ x) +
  xlab('Distance to Nearest Safe Well (100s of meters)') +
  ylab("P(Switch)")
```
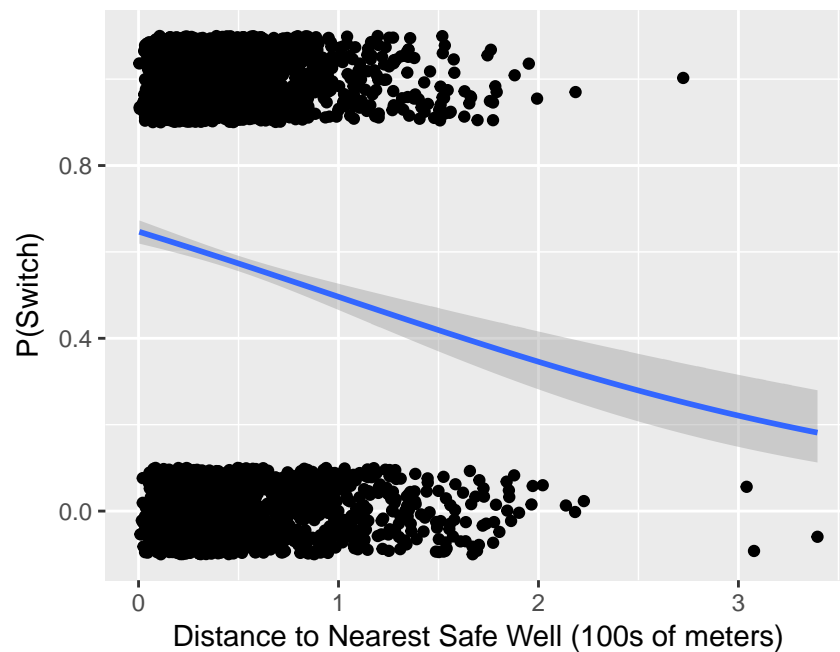

8-9: Probabilities of switching wells based on `fit1`

```r
#---------- Prob(switch) if dist100 = 0
predict(fit1, newdata = data.frame(dist100 = 0), type = 'response')
```

```
        1
0.6470185
```

4

```r
#--------- Prob(switch) at the avg of dist100
avgdist <- wells %>%
  summarize(avgdist = mean(dist100)) %>%
  pull(avgdist)
predict(fit1, newdata = data.frame(dist100 = avgdist), type = 'response')
```

```
        1
0.5757602
```