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 `stargazer` to make a summary table of your logistic regression results form part (a).
(c) Use `ggplot2` to plot the logistic regression function from part (a) along with the data, jittered appropriately.
(d) Discuss your results from parts (a)-(c). In particular: based on `fit1`, are the x-variables in this regression statistically significant predictor of `switch`? Do the signs of the coefficients make sense? Explain.
(e) Calculate the predicted probability of switching wells for the *average* household in the dataset.
(f) Calculate the marginal effect of `dist100` for the *average* household in the dataset. Interpret your results.

3. Predictive performance of `fit1`

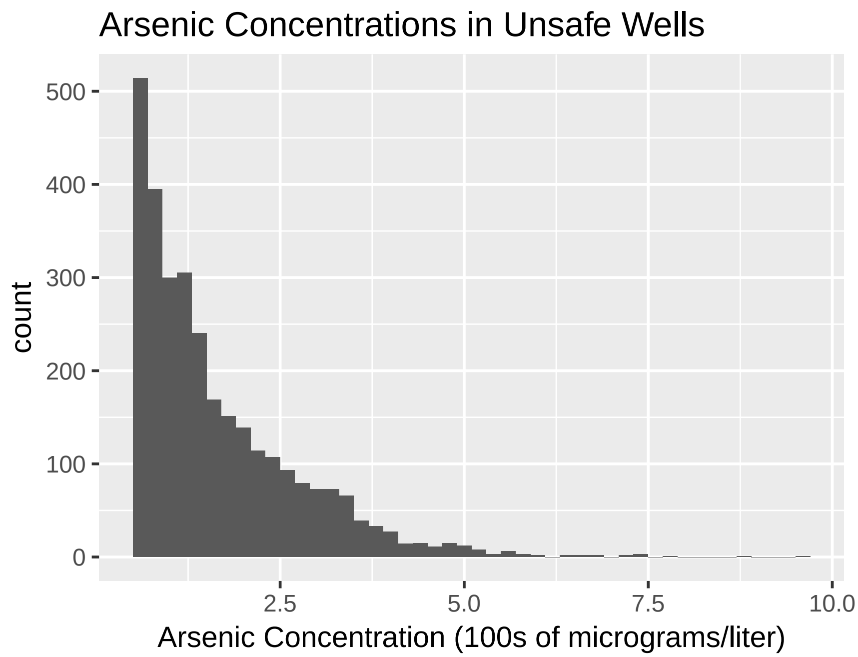
(a) Add a column called `pred1` to `wells` containing the predicted probabilities that `switch` equals one based on `fit1`.
(b) Use `pred1` to calculate the *Bayes error rate* of `fit1` based on the full training dataset, i.e. `wells`. Hint: you can do this using the `dplyr`

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

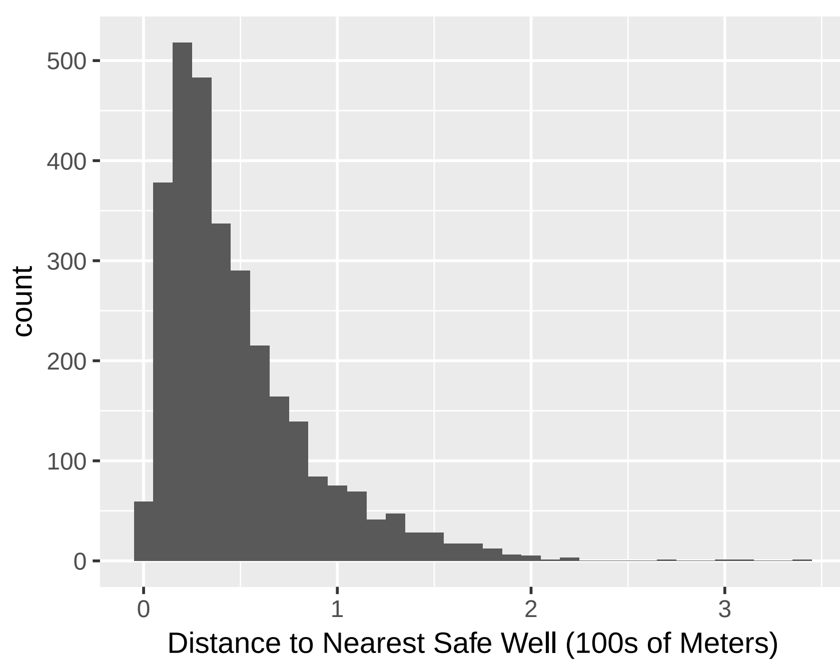
Solutions

1 - Preliminaries

```
#------- 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')
```




```r
#-------- 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)')
```

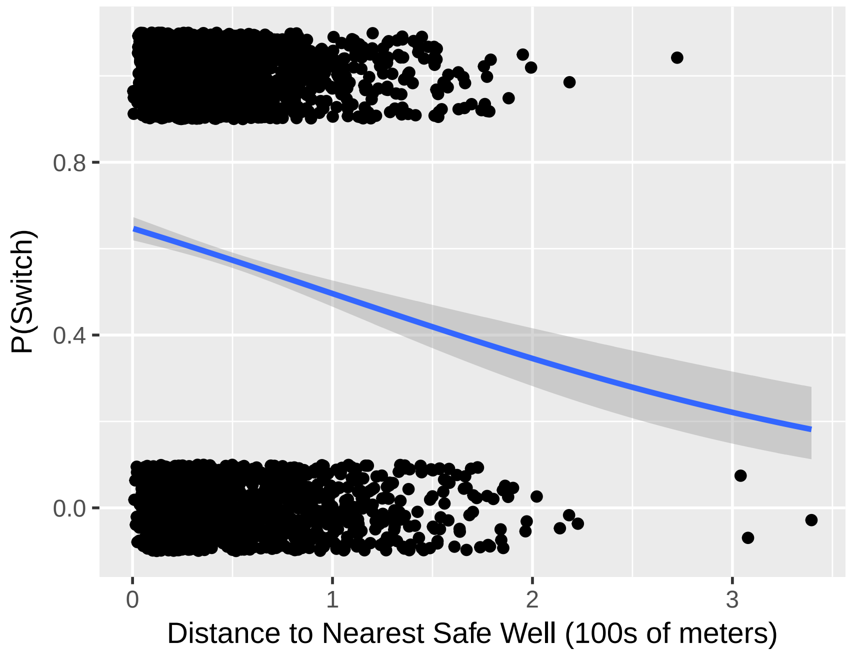


2 - First Regression: `fit1`

```r
#---------- 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)")
```




```
#--------- 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
```

```
#---------- In-sample predictions from fit1
wells <- wells %>%
  mutate(pred1 = predict(fit1, type = 'response'))

#---------- Bayes Error Rate from fit1
wells %>%
  summarize(error_rate = mean((pred1 > 0.5 & switch == 0) | (pred1 <= 0.5 & switch == 1)))
```

```
# A tibble: 1 x 1
  error_rate
       <dbl>
1      0.405
```

Make a nice table with lots more regressions with stargazer!

```
library(stargazer)
fit2 <- glm(switch ~ I(log(arsenic)), wells, family = binomial(link = 'logit'))
fit3 <- glm(switch ~ dist100 + I(log(arsenic)), wells, family = binomial(link = 'logit'))
fit4 <- glm(switch ~ dist100 + I(log(arsenic)) + dist100:I(log(arsenic)), wells, family = binomial(link
fit5 <- glm(switch ~ dist100 + I(log(arsenic)) + dist100:I(log(arsenic)) + educ, wells, family = binomi
stargazer(fit1, fit2, fit3, fit4, fit5, type = 'latex', digits = 2,
          title = 'Logistic Regression Results',
          header = FALSE)
```

Table 2: Logistic Regression Results

	Dependent variable:				
	switch				
	(1)	(2)	(3)	(4)	(5)
dist100	−0.62***		−0.98***	−0.87***	−0.88***
	(0.10)		(0.11)	(0.13)	(0.13)
I(log(arsenic))		0.71***	0.88***	0.98***	0.99***
		(0.06)	(0.07)	(0.11)	(0.11)
educ					0.04***
					(0.01)
dist100:I(log(arsenic))				−0.23	−0.21
				(0.18)	(0.18)
Constant	0.61***	0.10**	0.53***	0.49***	0.29***
	(0.06)	(0.04)	(0.06)	(0.07)	(0.08)
Observations	3,020	3,020	3,020	3,020	3,020
Log Likelihood	−2,038.12	−1,994.64	−1,949.18	−1,948.39	−1,938.44
Akaike Inf. Crit.	4,080.24	3,993.29	3,904.37	3,904.77	3,886.88
Note:	*p<0.1; **p<0.05; ***p<0.01				