POL 213, Spring 2024
Problem Set 3

Professor: Lauren Peritz

Due: May. 31, 2024

Instructions:

- Responses should be typeset in LaTeXor RMarkdown. If you have difficulty, you may use Word but you will quickly find that it stinks.
- Submit your completed problem set as a single PDF via the course website. If you are not using RMarkdown, please include a copy of your code in your write-up.
- All work must be your own. Do not collaborate. You may ask Lily and me questions.
- Please DO NOT submit pages of copy-paste R output. Problem sets doing this will be graded as unsatisfactory.

---

1 Midterm exam

Go through your midterm exam and identify questions you missed. Using your textbook and notes, submit updated solutions to any questions you were unable to correctly solve in the exam.

2 Practice with Logistic Regression

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. In a project in the Araihazar region, researchers measured all the wells people used and labeled them as "safe" or "unsafe" according to whether their arsenic levels fell below or above 0.5 hundreds of micrograms per liter. People with unsafe wells were encouraged to swith to nearby private or community wells or construct new private ones. A few years later, the researchers returned to find out who had switched wells.

Download the "`wells.dat`" data. The data set consists $n = 3020$ households encouraged to switch with the following variables:

- `switch` = 1 if household i switched to a new well; 0 otherwise (Y)
- `arsenic` = contaminant level in i's well (X1)
- `distance` = meters to closest known safe well (X2)

- `assoc` = whether any members of the household are active in community organizations (X3)
- `educ` = education level of the head of household (X4)

Use these data to answer the following questions.

(a) Begin with a model in which the probability of a `switch` is a function of `arsenic` and `distance`. Write down the log likelihood function for your model by adapting the Chilean example we covered in class. It will be easier to rescale the distance in 100-meter units:

```
wells$dist100 <- wells$dist/100
```

(b) Run a logistic regression with just predictors $X1$ and $X2$. Run a second model with $X1$, $X2$, $X3$, and $X4$.

(c) Choose one of the models and explain why you prefer it.

(d) Interpret the results by answering the following questions / doing the following calculations: Are the parameters significantly different from 0? Give an odds ratio interpretation for at least one of the covariates.

(e) Using your preferred model, plot the distribution of 1s and 0s in your response variable and the fitted values for your estimates. Label the axes appropriately.

3 Derivation of Logit Model

Fox textbook, exercise 14.6. (Same across the 2nd and 3rd editions.)