EBM Basic Simulation

Rachael Caelie (Rocky) Aikens

7/22/2020

Set up

We'll start with the basic cross-sectional model, with a single baseline characteristic, X_i. Here, X_i is binary and about half the population has $X_i = 1$, half has $X_i = 0$. (This could be sex, naturally, but we might consider some other discrete group like race, some comorbidity, sexuality, etc.)

$$\begin{aligned} X_i &\sim_{iid} \text{Bernoulli}(0.5) \\ Z_i &\sim_{iid} \text{Bernoulli}(0.15) \\ S_i | Z_i = 1 &\sim_{iid} \text{Uniform}(0, 1) \end{aligned}$$

Where $S_i = 0$ whenever $Z_i = 0$.

Additionally let's consider that a person's probability of diagnosis depends on not only the severity of their disease but their baseline characteristics:

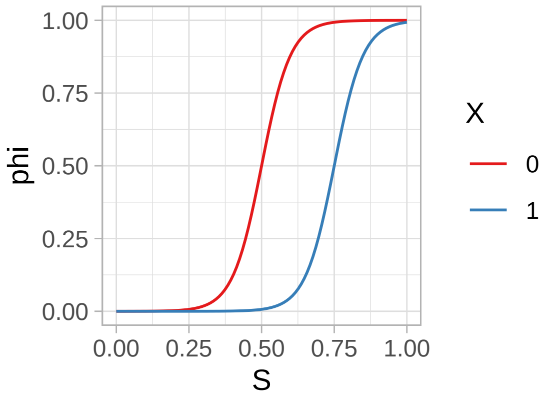
$$D_i \sim_{iid} \text{Bernoulli}\left(\phi(S_i, X_i)\right)$$

Without loss of generality, let's assume that people with $X_i = 0$ are more likely to be correctly diagnosed than people with $X_i = 1$. Explicitly, let's suppose:

$$\phi(S_i, X_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 S_i + \beta_2 X_i))} - c_{x_i},$$

where again, $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 X_i))}$ is a corrective constant to ensure that $\phi(0, X_i) = 0$. Let Let $\beta_0 = -10$, $\beta_1 = 20$, and $\beta_2 = -5$.

Inessence, this means that people are diagnosed according to a sigmoid function, with people with $X_i = 1$ needing a higher severity to be diagnosed with the same probability as people with $X_i = 0$, shown below:



A Naive Study

Again, we collect a dataset of 5×10^4, recording the baseline characteristic $\{S_i, D_i, X_i\}_{i=1}^{n}$. Now, our researcher wants to fit a logistic model for predicting disease status from X_i and S_i. However, since the true disease status is unkown, they use diagnosis as a proxy for disease. In regression shorthand they want to model:

$$D_i \sim S_i + X_i$$

The model result of such a study is shown below.

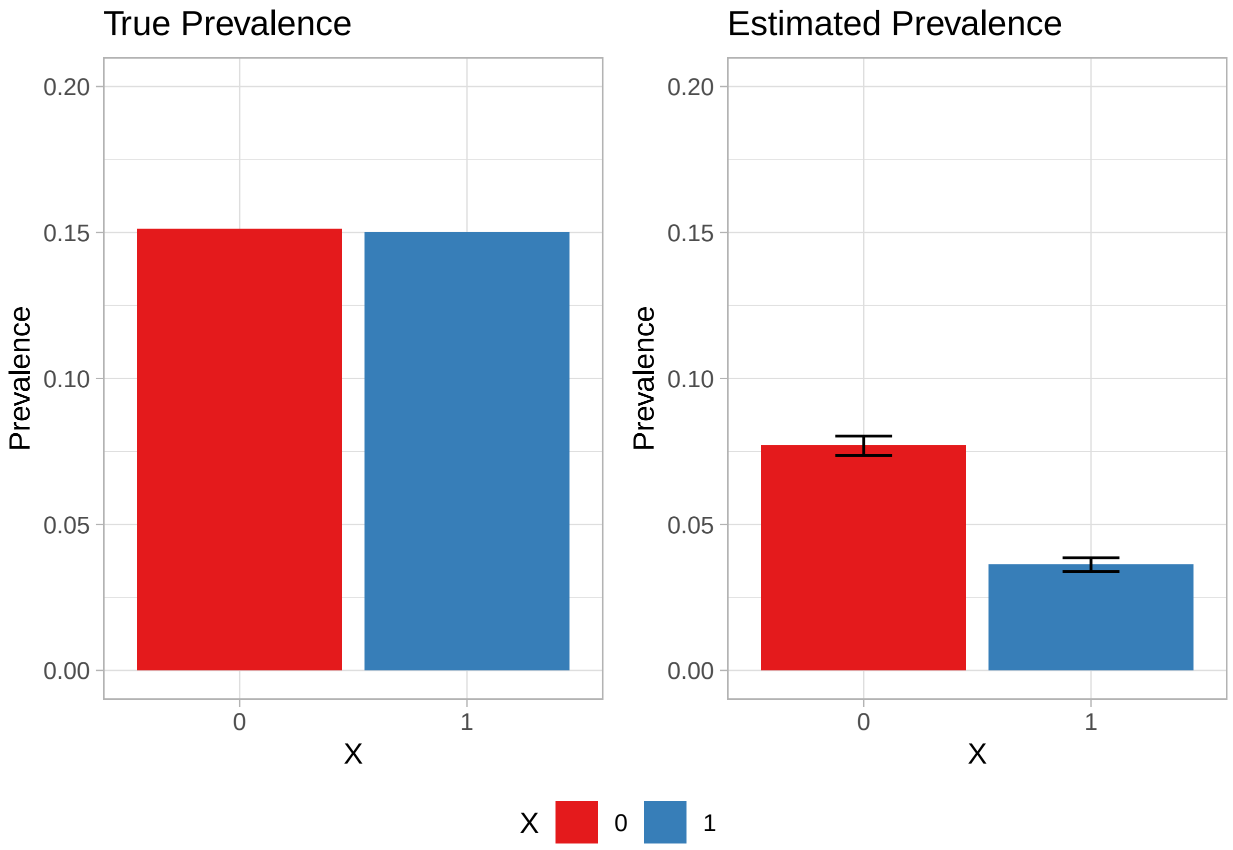
```
##
## Call:
## glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2474  -0.0080  -0.0080  -0.0006   4.2520
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.3584     0.3054  -33.91   <2e-16 ***
## x            -5.1981     0.1805  -28.79   <2e-16 ***
## severity     20.8046     0.5956   34.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21626.8  on 49999  degrees of freedom
## Residual deviance:  2376.6  on 49997  degrees of freedom
## AIC: 2382.6
##
## Number of Fisher Scoring iterations: 11
```

	OR	2.5 %	97.5 %
(Intercept)	0.000000e+00	0.000000e+00	0.000000e+00
x	6.000000e-03	4.000000e-03	8.000000e-03
severity	1.084723e+09	3.499006e+08	3.614885e+09

The researcher records a very high AIC and reports that this model is quite effective at predicting who has the disease. They write a nice paper, talking about how predictive modeling is the gateway to evidence-based personalized medicine. Maybe they suggest that future work might leverage the Awesome Power of Machine Learning. When they interpret their results, they might say that $X_i = 0$ significantly reduces predicted risk of disease.

Deployment

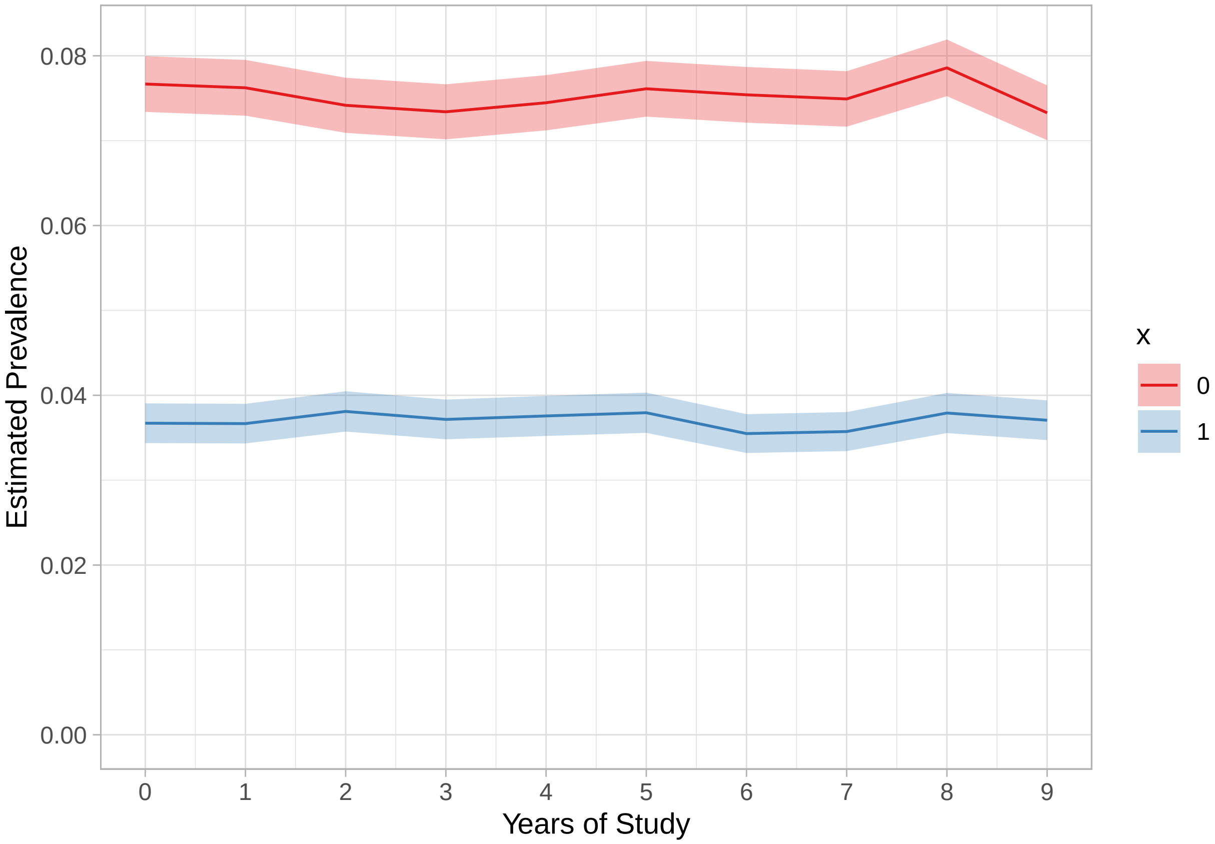
Doctors then read about this model and deploy it. In reality, we know that doctors usualy do not do exact numeric calculations to decide what diagnosis to give, but let's suppose for a moment that they perfectly follow this published model. In reality, some doctors will be less faithful and some more faithful.



Iterative EBM

```
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -10.076       -5.008       20.023
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21820
## Residual Deviance: 2400  AIC: 2406
```

```
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -10.075       -4.924       20.093
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21720
## Residual Deviance: 2410  AIC: 2416
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##      -9.831       -4.794       19.608
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21630
## Residual Deviance: 2508  AIC: 2514
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -10.349       -4.976       20.561
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21350
## Residual Deviance: 2358  AIC: 2364
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##      -10.27        -5.12        20.64
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21600
## Residual Deviance: 2361  AIC: 2367
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -10.741       -5.189       21.458
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21880
## Residual Deviance: 2313  AIC: 2319
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -11.154       -5.546       22.192
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21390
## Residual Deviance: 2163  AIC: 2169
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -11.392       -5.589       22.653
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21370
## Residual Deviance: 2113  AIC: 2119
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##      -11.65        -5.72        23.29
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       22200
## Residual Deviance: 2155  AIC: 2161
##
## Call:  glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = curr_df)
##
## Coefficients:
## (Intercept)            x     severity
##     -11.704       -5.559       23.177
##
## Degrees of Freedom: 49999 Total (i.e. Null);  49997 Residual
## Null Deviance:       21350
## Residual Deviance: 2062  AIC: 2068
```


Estimated Prevalence
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
6
7
8
9
Years of Study
x
0
1