Lecture 11
Intro to logistic regression

BIOST 515

February 10, 2004

Modeling binary data

Often in medical studies, we encounter outcomes that are not continous, but instead fall into 1 of 2 categories. For example:

- Disease status (disease vs. no disease)
- Alive or dead
- Low birth weight
- Improved health status

In these cases, we have a binary outcome

$$y_i = \begin{cases} 0 & \text{with probability } 1-\pi_i \\ 1 & \text{with probability } \pi_i \end{cases},$$

where

$$E[y_i] = \pi_i$$

and

$$\text{var}[y_i] = \pi_i(1-\pi_i).$$

Usually, one f the categories is the outcome of interest, like death or disease. This category is usually coded as 1.

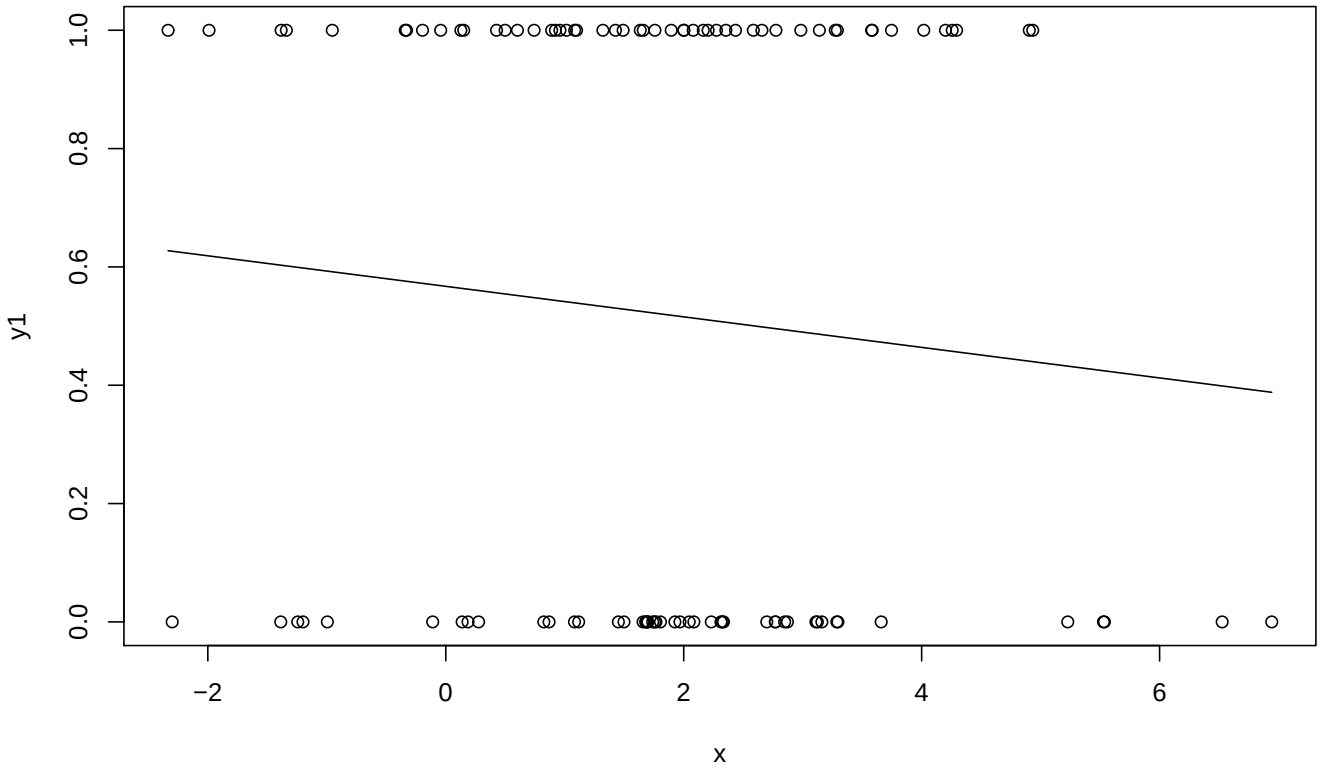
We can use linear regression to model this outcome, but this can present several problems as we will see.

Using the linear model approach, we relate the expected value of y_i to a predictor x_i as

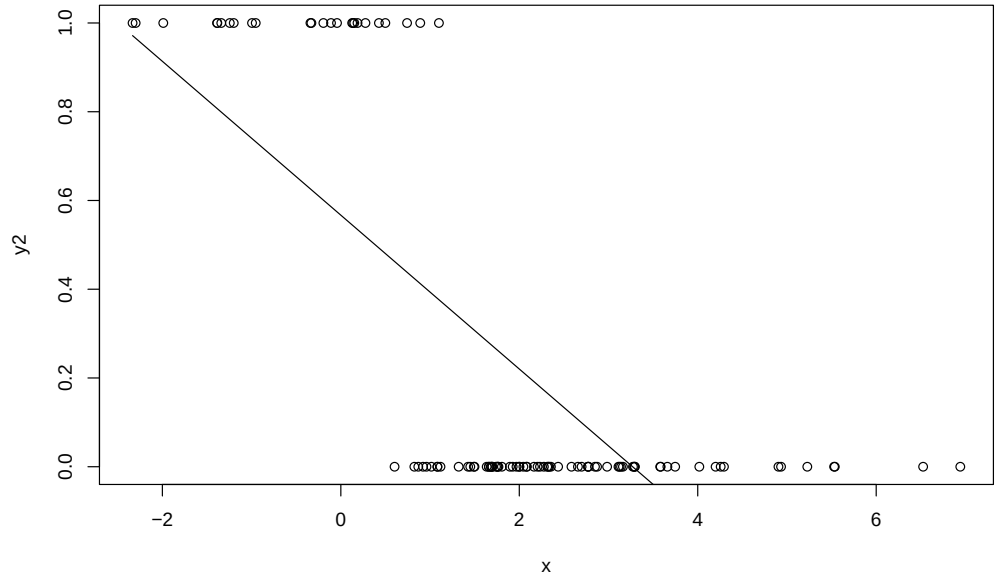
$$E[y_i] = \beta_0 + \beta_1 x_i$$

Just looking at this relationship, we can see a potential problem. What is it?

Over small ranges of the predictor or when the relationship between the predictor and the outcome is not strong, this may not be troubling.



However, if the association is strong, potential problems are more evident.



We could put constraints on the βs that would prevent this from happening, but this would be complicated and probably not the best way to address this problem.

The next obvious problem comes from the relationship

$$\begin{aligned} \text{var}[y_i] &= \pi_i(1-\pi_i) \\ &= E[y_i](1-E[y_i]) \\ &= (\beta_0+\beta_1 x_i)(1-\beta_0-\beta_1 x_i) \end{aligned}$$

What is this problem?

We may be able to do a transformation to fix this problem, but it would be better to use the information we have about the mean-variance relationship to build a more appropriate regression model.

Review of 2×2 tables

	Disease	
Exposure	Yes	No
Yes	π_{11}	π_{12}
No	π_{21}	π_{22}

where π_{ij} =Pr (exposure=i & disease = j).

Two of the most commonly used summaries of association are the relative risk and the odds ratio.

Relative risk

$$RR = \frac{\text{Pr(Disease|Exposure)}}{\text{Pr(Disease|No Exposure)}} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{21}/(\pi_{21}+\pi_{22})}$$

$$= \frac{\pi_{11}(\pi_{21}+\pi_{22})}{\pi_{21}(\pi_{11}+\pi_{12})}$$

Odds ratio

Given exposure, the odds of getting the disease are

$$\frac{\text{Pr(Disease|Exposure)}}{\text{Pr(No Disease|Exposure)}} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{12}/(\pi_{11}+\pi_{12})} = \frac{\pi_{11}}{\pi_{12}}.$$

The odds ratio can then be expressed as

$$OR = \frac{\text{Odds of Disease|Exposure}}{\text{Odds of Disease|No Exposure}} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{12}} = \frac{\pi_{11}}{\pi_{21}}.$$

Regression models for probability of disease

How do we relate the outcome, y, to an exposure, x? Recall the first lecture, when we discussed relating functions of the mean to linear functions of predictors (exposures). We will take that approach to modeling the outcome in this case by modeling

$$\begin{aligned} g(E[y_i|x_i]) = g(\pi_i) &= \beta_0 + \beta_1 x_i \\ E[y_i|x_i] = \pi_i &= g^{-1}(\beta_0 + \beta_1 x_i), \end{aligned}$$

where $g()$ is called a *link function*. How do we interpret π_i?

Distribution of y

In this case, we know that y_i follows a bernoulli distribution

$$\begin{aligned} p(y_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= [g^{-1}(\beta_0+\beta_1 x_i)]^{y_i}[1-g^{-1}(\beta_0+\beta_1 x_i)]^{1-y_i}. \end{aligned}$$

Relating π_i to exposure

We will first look at the case where the exposure is dichotomous (exposed/unexposed =1/0) One way we may relate π_i to the exposures is through the log-link, $g = \log$. This gives the following relationship

$$\log(\pi_i) = \beta_0 + \beta_1 x_i.$$

When a subject is exposed, $x_i = 1$ and $\pi_i = \pi_{D|E}$ (probability of disease given exposure). In the 2×2 table, this was $\pi_{11}/(\pi_{11} + \pi_{12})$. Therefore,

$$\log(\pi_{D|E}) = \beta_0 + \beta_1.$$

When a subject is unexposed, $x_i = 1$ and $\pi_i = \pi_{D|E^c}$ (prob-

ability of disease given no exposure). In the 2×2 table, this was $\pi_{21}/(\pi_{21} + \pi_{22})$. Therefore,

$$\log(\pi_{D|E^c}) = \beta_0.$$

We can then get the relative risk as follows.

$$\begin{aligned} \log(\pi_{D|E}) - \log(\pi_{D|E^c}) &= \beta_0 + \beta_1 - \beta_0 \\ \log\left(\frac{\pi_{D|E}}{\pi_{D|E^c}}\right) &= \beta_1 \\ RR &= \exp(\beta_1). \end{aligned}$$

What are some potential drawbacks of this modeling scheme?

Logistic regression

In logistic regression, we use the logit link, which is defined as

$$g(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

This is equivalent to modeling the log odds. We relate $E[y_i|x_i]$ to the exposure using

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i.$$

When a subject is exposed, $x_i = 1$ and $\pi_i = \pi_{D|E}$ (probability of disease given no exposure). Therefore,

$$\text{logit}(\pi_{D|E}) = \beta_0 + \beta_1.$$

This is equivalent to the log of the odds of disease given exposure.

When a subject is unexposed, $x_i = 0$ and $\pi_i = \pi_{D|E^c}$ (probability of disease given no exposure). Therefore,

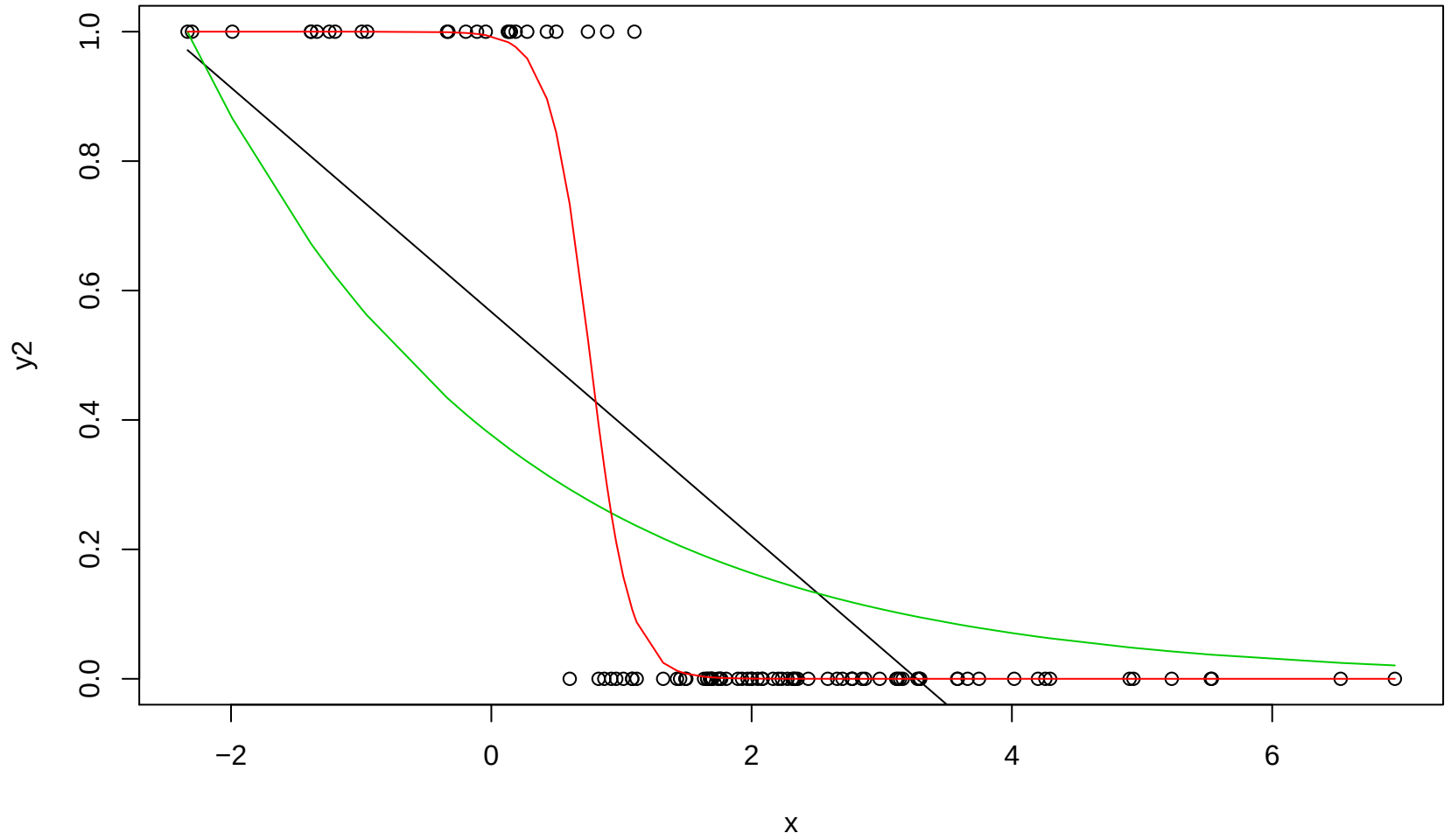
$$\text{logit}(\pi_{D|E^c}) = \beta_0.$$

This is equivalent to the log of the odds of disease given no exposure.

Calculating the odds ratio

We can calculate the odds ratio as follows

$$\begin{aligned}
\text{logit}(\pi_{D|E}) - \text{logit}(\pi_{D|E^c}) &= \beta_0 + \beta_1 - \beta_0 \\
\log\left(\frac{\pi_{D|E}}{1-\pi_{D|E}}\right) - \log\left(\frac{\pi_{D|E^c}}{1-\pi_{D|E^c}}\right) &= \beta_1 \\
\left(\frac{\pi_{D|E}}{1-\pi_{D|E}} / \frac{\pi_{D|E^c}}{1-\pi_{D|E^c}}\right) &= \beta_1 \\
OR &= \exp(\beta_1)
\end{aligned}$$


y2
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
x

CHS Example

In this example, we will look at coronary heart disease. We code

$$y_i = \begin{cases} 1, & \text{disease} \\ 0, & \text{no disease} \end{cases}.$$

The exposure is male gender. Our observed proportions are

	CHD	
Exposure	Yes	No
Male	0.098	0.322
Female	0.102	0.478

Pr(Disease)$= 0.098 + 0.102 = 0.2$

Pr(Disease|Male)=0.098/(0.098+0.322)=0.233

Pr(Disease|Female)=0.102/(0.102+0.478)=0.176

Pr(Disease|Male)$-$Pr(Disease|Female)$= 0.233 - 0.176 = 0.057$

$$RR = 0.098/(0.098 + 0.322)/(0.102/(0.102 + 0.478)) = 1.32$$
$$OR = 0.098/0.322/(0.102/0.478) = 1.43$$

Because this is a simple 2×2 table, our estimates from linear regression and glm with log and logit links should match.

Linear regression

```
lm1=lm(CHD~GENDER,data=chs)
summary(lm1)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1759	0.0235	7.49	0.0000
GENDER	0.0575	0.0362	1.59	0.1133

$$\pi_i = 0.1759 + 0.0575 x_i$$

GLM with log link

```
glm1=glm(CHD~GENDER,family=binomial(link="log"),data=chs)
summary(glm1)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−1.7381	0.1271	−13.67	0.0000
GENDER	0.2828	0.1783	1.59	0.1128

$$\log(\pi_i) = -1.7381 + 0.2828 x_i$$

How do we get the relative risk from this output?

Logistic regression (GLM with logit link)

```
glm2=glm(CHD~GENDER,family=binomial(link="logit"),data=chs)
summary(glm2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−1.5446	0.1542	−10.01	0.0000
GENDER	0.3551	0.2245	1.58	0.1138

$$\text{logit}(\pi_i) = -1.5446 + 0.3551 x_i$$

How do we get the odds ratio from this output?

Exposures measured on a continuous scale

So far, we've only replicated the information we could from a 2×2 table. What if, instead, we had an exposure that was measured on a continuous scale. Examples

- Age
- An enviromental toxin that is hypothesized to be related to some disease
- Score on an elementary school exam and subsequent enrollment in college

Relative risk regression with continuous predictor

$$log(\pi_i) = \beta_0 + \beta_1 x_i$$

$$\begin{array}{ll} x = c & \log(\pi(x_i = c)) = \beta_0 + \beta_1 c \\ x = c+1 & \log(\pi(x_i = c+1)) = \beta_0 + \beta_1 (c+1) \end{array}$$

$$\begin{aligned} \log(\pi(x_i = c)) - log(\pi(x_i = c+1)) &= \beta_0 + \beta_1 c - (\beta_0 + \beta_1 (c+1) \\ \log\left(\frac{\pi(x_i = c))}{\pi(x_i = c+1)}\right) &= \beta_1 (c - c - 1) \\ \frac{\pi(x_i = c)}{\pi(x_i = c+1)} &= \exp(-\beta_1) \end{aligned}$$

How do we interpret this?

Logistic regression with continuous predictor

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

$x = c$	$\text{logit}(\pi(x_i = c)) = \beta_0 + \beta_1 c$
$x = c + 1$	$\text{logit}(\pi(x_i = c + 1)) = \beta_0 + \beta_1 (c + 1)$

$$\text{logit}(\pi(x_i = c)) - logit(\pi(x_i = c + 1)) = \beta_0 + \beta_1 c - (\beta_0 + \beta_1 (c +$$

$$\frac{\text{odds}(\pi(x_i = c))}{\text{odds}(\pi(x_i = c + 1))} = \exp(-\beta_1)$$

Example

In this example, we will look at age as a predictor of CHD. The regression model is

$$g(E(CHD_i)) = \beta_0 + \beta_1 \text{age}_i.$$

If we use linear regression ($CHD_i = \beta_0 + \beta_1 \text{age}_i + \epsilon_i$), the results are

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.0127	0.2346	−0.05	0.9569
AGE	0.0029	0.0032	0.91	0.3636

Example with relative risk regression (GLM with log link)

```
glm1.2=glm(CHD~AGE,family=binomial(link="log"),data=chs)
summary(glm1.2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−2.6598	1.1230	−2.37	0.0179
AGE	0.0143	0.0152	0.94	0.3457

Example with logistic regression

```
glm1.2=glm(CHD~AGE,family=binomial(link="log"),data=chs)
summary(glm1.2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−2.6846	1.4356	−1.87	0.0615
AGE	0.0177	0.0195	0.91	0.3632