Logistic Regression: Set Up

Response variable is categorical w/ 2 categories

↳ if more than 2 categories, use multinomial logistic regression

Explanatory variables: could be anything; today, 1 quantitative

Example: for each baby in a sample, I record:

- whether or not the baby has bronchopulmonary dysplasia (BPD)
- the baby's birth weight in grams

Question answered by logistic regression:
what is the relationship between a baby's birth weight and the probability they have BPD.

Actual response variable for model:

$$y_i = \begin{cases} 1 & \text{if baby number } i \text{ has BPD} \\ 0 & \text{if not} \end{cases}$$

x_i = birth weight for baby number i.

Note: y_i definitely doesn't follow a normal distribution!

We cannot write $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ | $y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$

We need a probability distribution for ~~???~~ a response (2)
variable that is either 0 or 1.

Bernoulli random variables

$Y \sim \text{Bernoulli}(p)$ means that:

- Y is a "random variable"
  (imagine selecting a baby at random and recording whether or not it has BPD)
- Y is either 0 or 1
- The probability that $Y=1$ is p.
  (ex: p could represent the proportion of babies in the population with BPD)

  Note: p is between 0 and 1

Example: Suppose 1% of babies in population have BPD. I pick a baby at random and record

$$Y = \begin{cases} 1 & \text{if baby has BPD} \\ 0 & \text{if not} \end{cases}$$

$$Y \sim \text{Bernoulli}(0.01)$$

Example: I flip a coin once and record

$$Y = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$

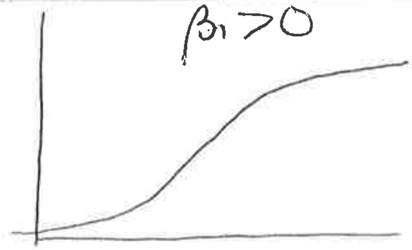
$$Y \sim \text{Bernoulli}(0.5)$$

~~Idea 20~~

Logistic Regression Model:

Idea: $Y_i \sim \text{Bernoulli}(p_i)$, where p_i is specific to baby #i, based on its birth weight.

We need a way to take a birth weight x_i and turn it into a probability.



Logistic function:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Note: $e^a > 0$ for any a.

So $\frac{e^a}{1+e^a} > 0 \quad \Rightarrow \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} > 0$

Also, $e^a < 1+e^a$, so $\frac{e^a}{1+e^a} < 1$

$$\Rightarrow \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} < 1$$

Logistic Regression Model:

$$Y_i \sim \text{Bernoulli}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

Interpretation of β_1:

Def.: Odds that $Y=1$:

$$\text{Odds}(Y=1) = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$$

Ex: • If $P(Y=1) = 0.75$, ~~Odds(Y=1) =~~

$$\text{Odds}(Y=1) = \frac{0.75}{1-0.75} = \frac{0.75}{0.25} = 3$$

The probability that $Y=1$ is 3 times the probability that $Y=0$.

• If $P(Y=1) = 0.5$,

$$\text{Odds}(Y=1) = \frac{0.5}{1-0.5} = \frac{0.5}{0.5} = 1$$

The probability that $Y=1$ is the same as the probability that $Y=0$

• If $P(Y=1) = 0.1$,

$$\text{Odds}(Y=1) = \frac{0.1}{0.9} = \frac{1}{9}$$

The probability that $Y=1$ is $\frac{1}{9}$ the probability that $Y=0$.

Odds in logistic regression:

$$\text{Odds}(y_i=1) = \frac{P(y_i=1)}{P(y_i=0)} = \frac{P(y_i=1)}{1-P(y_i=1)}$$

$$= \frac{\left(\dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)}{\left(1-\dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)}$$

$$= \frac{\left(\dfrac{e^{\beta_0+\beta_1 x_i}}{\cancel{1+e^{\beta_0+\beta_1 x_i}}}\right)}{\left(\dfrac{1+\cancel{e^{\beta_0+\beta_1 x_i}}-\cancel{e^{\beta_0+\beta_1 x_i}}}{\cancel{1+e^{\beta_0+\beta_1 x_i}}}\right)}$$

$$= e^{\beta_0+\beta_1 x_i}$$

How do the odds change if x_i increases by 1 unit?

$$e^{\beta_0+\beta_1(x_i+1)} = e^{\beta_0+\beta_1 x_i+\beta_1}$$

$$= e^{\beta_0+\beta_1 x_i} \cdot e^{\beta_1}$$

Interpretation of $\hat{\beta}_1$:

A 1 unit increase in x multiplies the estimated odds that $y=1$ by $e^{\hat{\beta}_1}$.