Modeling a Binary Response

Logistic regression – Stat 230

Framingham Heart Study

- Long-term study of cardiovascular disease (heart attack, stroke, and other heart problems)
- First study of cardiovascular disease to follow initially healthy people over a long period of time
- Identified the idea that body measurements (blood pressure, cholesterol, etc) could predict cardiovascular disease.
- Participants from a population of free living (not hospitalized) subjects in the community of Framingham, Mass.
- Started in 1948, participants have been examined every other year since the inception
- In 3rd generation of participants!

Modeling overview

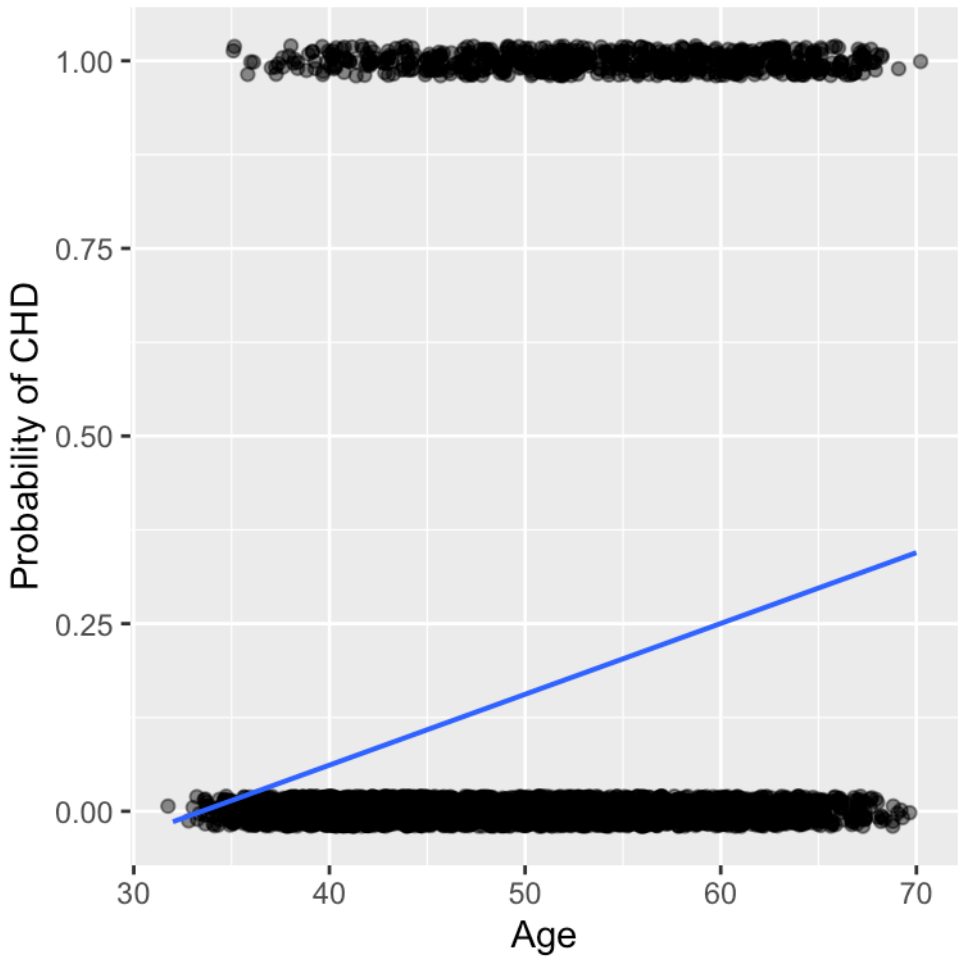
Goal: Determine whether participant experienced coronary heart disease (CHD) in 10-year window after their exam

- Y = CHD (0 = no, 1 = yes)
- X = participant's age, sex, total cholesterol, and systolic blood pressure

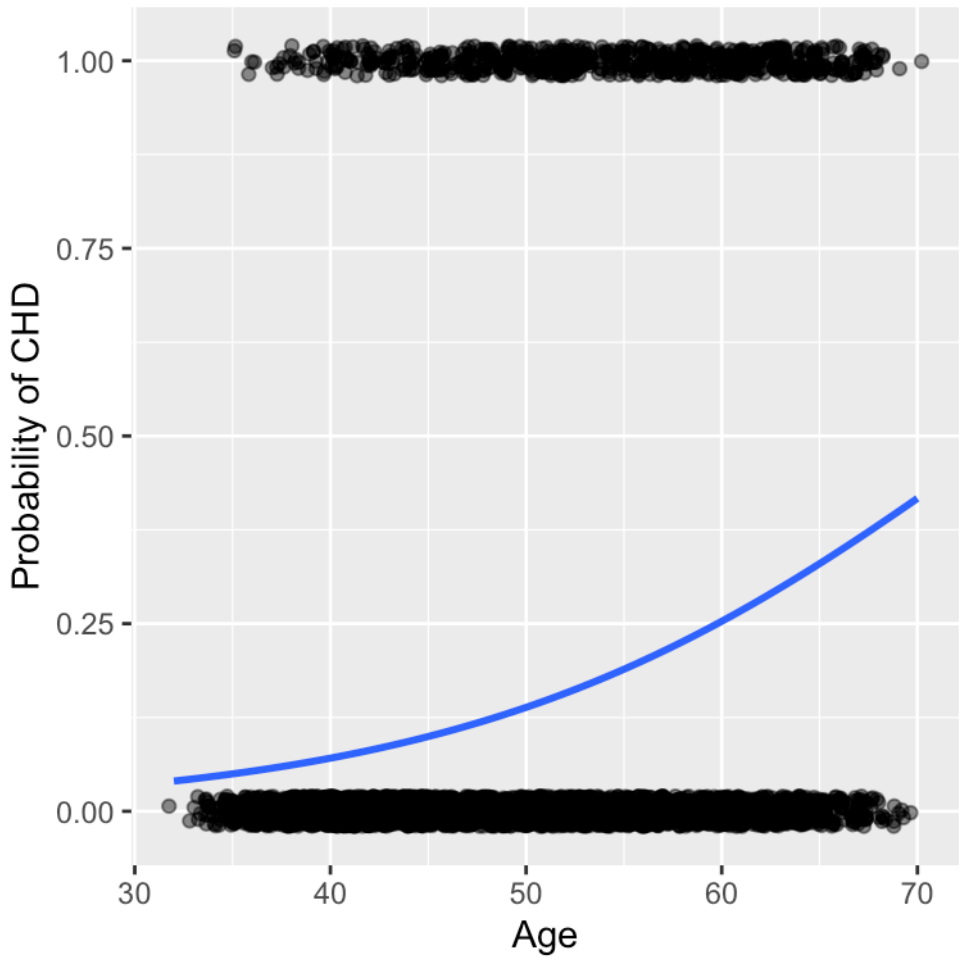
Strategy: model the probability of CHD given these factors

How can we model a binary response?

Linear regression



Logistic regression



Generalized linear model

Random component:

What distribution does the response variable follow?

Linear predictor (systematic component):

What function of the predictors should we use?

Link function:

How can we relate the expectation of Y and our linear predictor?

Random component

What distribution does the response variable follow?

Possible values: $Y = 0 \text{ or } 1$

Bernoulli distribution

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}, \; y = 0, 1$$

Linear predictor

What function of the predictors should we use?

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Link function

How can we relate the expectation of Y and our linear predictor?

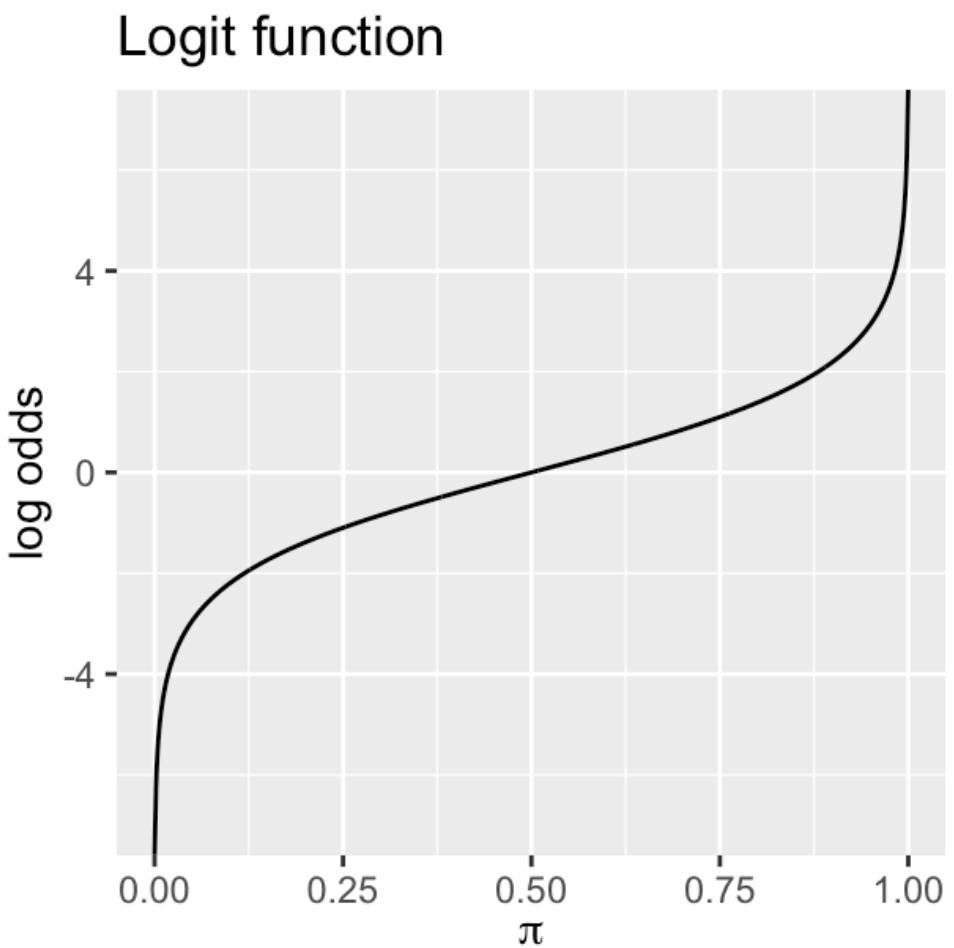
$E(Y) = \pi \longrightarrow$ between 0 and 1

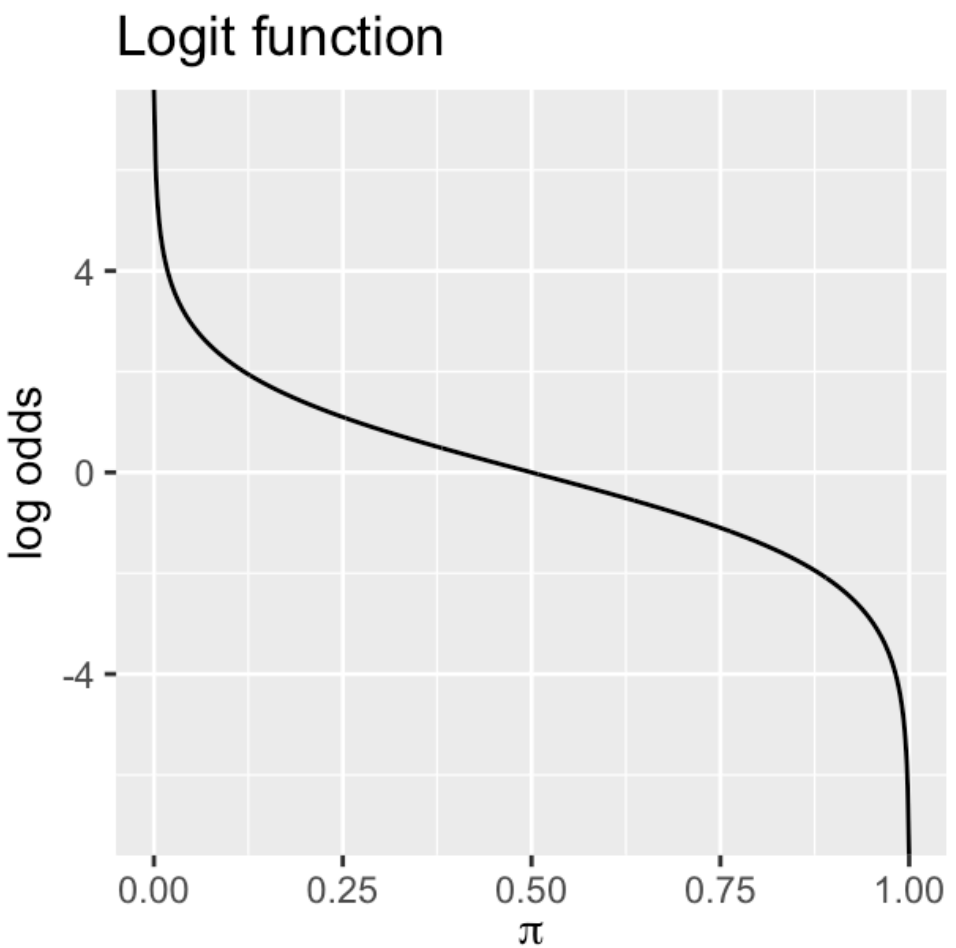
$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \longrightarrow$ between $-\infty$ and ∞

What function takes numbers between 0 and 1 as input and maps them to the real line?

The logit function

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$





Link function

How can we relate the expectation of Y and our linear predictor?

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \eta$$

Example: A quantitative predictor

Let's use total cholesterol as a predictor of CHD

$$\widehat{\eta}_i = \log\left(\frac{\widehat{\pi}(\texttt{totChol}_\texttt{i})}{1-\widehat{\pi}(\texttt{totChol}_\texttt{i})}\right) = -2.894 + 0.005\texttt{totChol}_i$$

$\widehat{\beta}_1 = 0.005$ is the expected change in the log odds associated with a 1-unit increase in total cholesterol…

We don't think in terms of log odds, so we need more interpretable quantities!

Log odds $\longrightarrow$ odds

Log odds:

$$\widehat{\eta} = \log\left(\frac{\widehat{\pi}(\mathrm{X})}{1-\widehat{\pi}(\mathrm{X})}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p$$

Odds of success:

$$\widehat{\omega} = \frac{\widehat{\pi}(\mathrm{X})}{1-\widehat{\pi}(\mathrm{X})} = e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p}$$

Example: A quantitative predictor

$$\hat{\eta}_i = \log\left(\frac{\hat{\pi}(\texttt{totChol})}{1-\hat{\pi}(\texttt{totChol})}\right) = -2.894 + 0.005\texttt{totChol}_i$$

How do we interpret $e^{\hat{\beta}_1} = e^{0.005} \approx 1.005$??

$$\frac{\hat{\pi}(\texttt{totChol})}{1-\hat{\pi}(\texttt{totChol})} = e^{-2.894+0.005\texttt{totChol}_i} = e^{-2.894} \cdot e^{0.005\texttt{totChol}_i}$$

So if `totChol` increases by one unit…

$$\frac{\hat{\pi}(\texttt{totChol}+1)/\{1-\hat{\pi}(\texttt{totChol}+1)\}}{\hat{\pi}(\texttt{totChol})/\{1-\hat{\pi}(\texttt{totChol})\}} = \frac{e^{-2.894} \cdot e^{0.005\texttt{totChol}+1}}{e^{-2.894} \cdot e^{0.005\texttt{totChol}}}$$

$$= \frac{e^{-2.894} \cdot e^{0.005\texttt{totChol}} \cdot e^{0.005(1)}}{e^{-2.894} \cdot e^{0.005\texttt{totChol}}}$$

$$= e^{0.005(1)} \approx 1.005$$

we expect the odds of CHD to increase by a factor of 1.005

Estimating probabilities

What's the probability that a person with total cholesterol of 240 mg/dL will develop CHD in the next 10 years?

$$\hat{\eta}(240) = \log\left(\frac{\hat{\pi}(240)}{1-\hat{\pi}(240)}\right) = -2.894 + 0.005(240) = -1.694$$

How do we get from the estimated log odds to a probability?

Log odds $\longrightarrow$ probability

Log odds:

$$\widehat{\eta} = \log\left(\frac{\widehat{\pi}(\mathrm{X})}{1-\widehat{\pi}(\mathrm{X})}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p$$

Odds of success:

$$\widehat{\omega} = \frac{\widehat{\pi}(\mathrm{X})}{1-\widehat{\pi}(\mathrm{X})} = e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p}$$

Probability of success:

$$\widehat{\pi}(X) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p}}$$

Estimating probabilities

What's the probability that a person with total cholesterol of 240 mg/dL will develop CHD in the next 10 years?

$$\hat{\eta}(240) = \log\left(\frac{\hat{\pi}(240)}{1-\hat{\pi}(240)}\right) = -2.894 + 0.005(240) = -1.694$$

How do we get from the estimated log odds to a probability?

$$\hat{\pi}(240) = \frac{e^{-1.694}}{1+e^{-1.694}} \approx 0.155$$

Fitted logistic regression model

$\text{logistic}(\hat{\eta}) = \dfrac{\exp(\hat{\eta})}{1 + \exp(\hat{\eta})}$ is the fitted curve on the probability scale