Logistic Regression

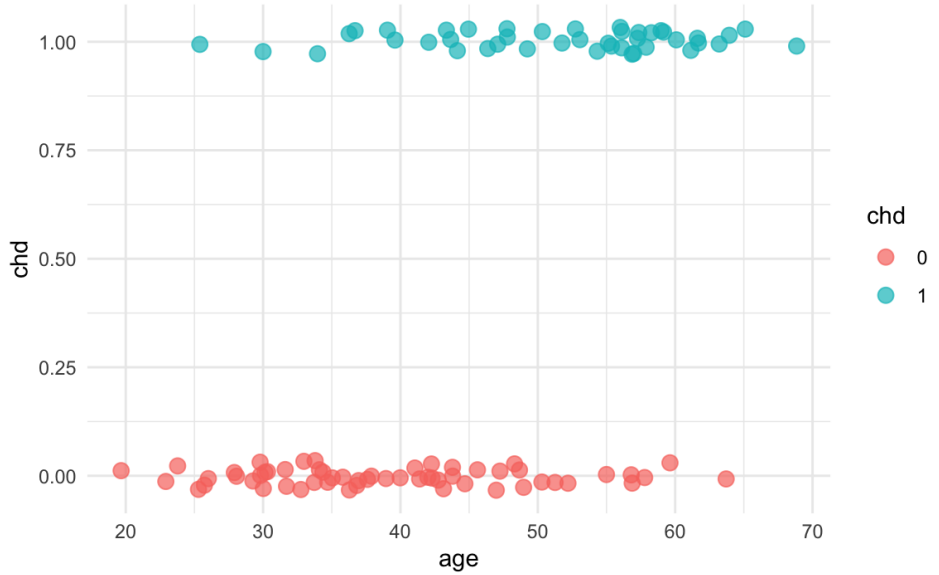
Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

May 22, 2023

Motivation

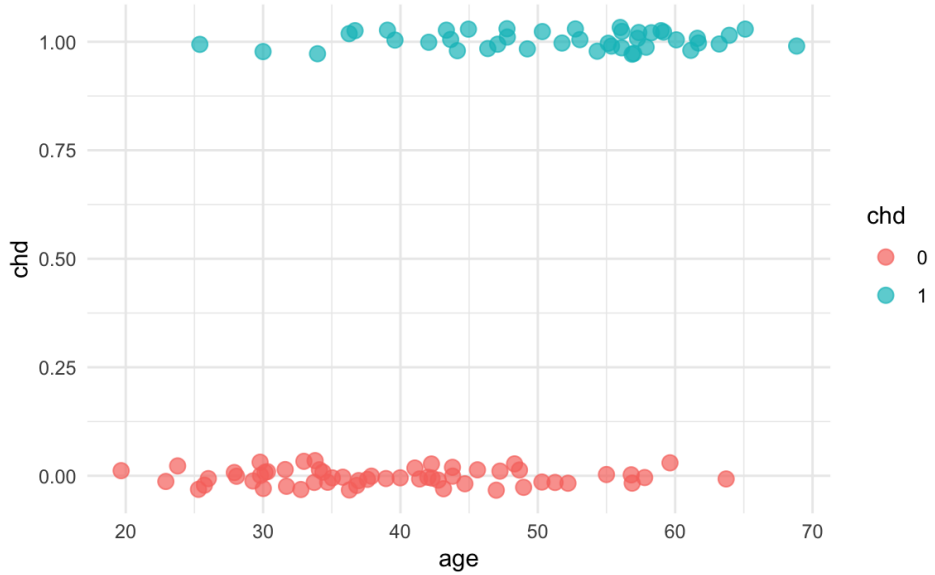
- Example: Predicting heart attack (coronary heart disease (chd)) with only one predictor: age.



- For age, we have values ranging from small x's to large x's.
- For the binary response, there are only 0's and 1's (for better visualization, there are some jitter).
- For $y_i = 0$, there are more small values x than large ones. For $y_i = 1$, there are more large values x than small ones.

Motivation

- Example: Predicting heart attack (coronary heart disease (chd)) with only one predictor: age.



- The goal is to fit a model that predicts chd from age.

$$\texttt{chd} = f(\text{age}) + \varepsilon$$

First Approach: Fitting a Line

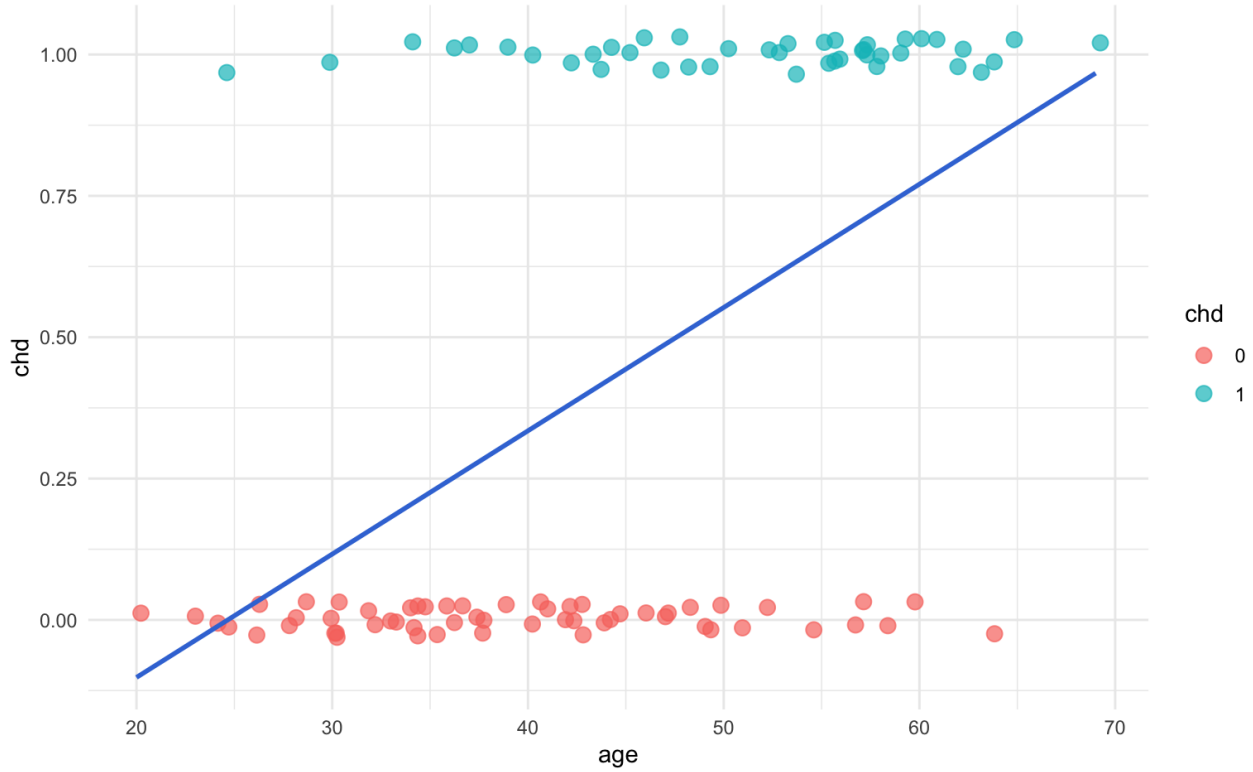
- We could try to fit a linear model:

$$\hat{\mathbf{y}} = b_0 + b_1\mathbf{x} = \mathbf{X}\mathbf{y}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

First Approach: Fitting a Line



- There seems to be some positive relation between age and `chd`.
- The regression line extends beyond the range [0,1], which does not make sense.

Second Approach: Harsh Thresholding

- We can set some threshold c and compare $\hat{y}_i = b_0 + b_1 x_i$ to c.

$$\hat{y}_i = \begin{cases} 1 & \text{if} \quad b_0 + b_1 x_i \geq c \\ 0 & \text{if} \quad b_0 + b_1 x_i < c \end{cases}$$

- We can arrange the terms to get:

$$\begin{aligned} \hat{y}_i = \hat{f}(x_i) = b_0 + b_1 x_i - c = (b_0 - c) + b_1 x_i \\ = b_0' + b_1 x_i = \mathbf{b}^\top \mathbf{x}_i \end{aligned}$$

- By paying attention to the sign of the signal $\mathbf{b}^\top \mathbf{x}$, we can transform our fitted model into:

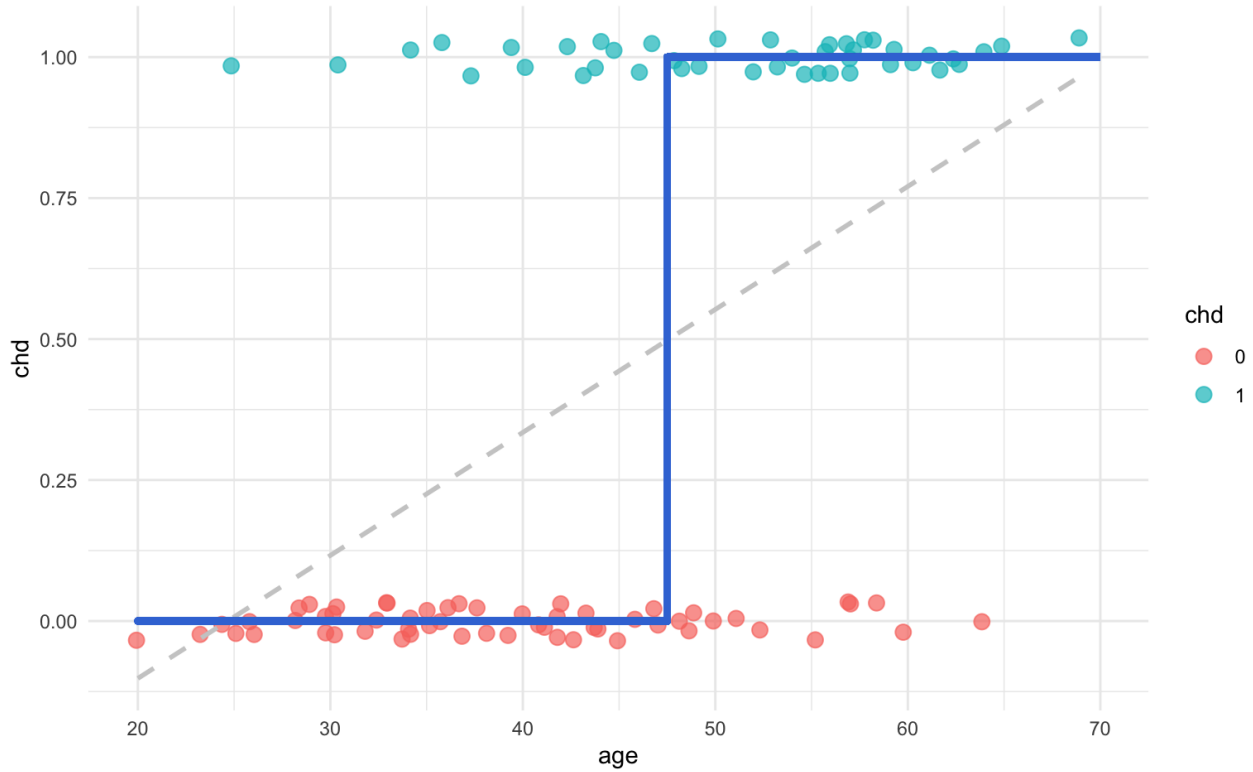
$$\hat{y} = \texttt{sign}(\mathbf{b}^\top \mathbf{x})$$

Second Approach: Harsh Thresholding

- We have a classification rule:

$$\hat{y}_i = \begin{cases} 1 & \text{if} \quad \texttt{sign}(b_0' + b_1 x_i) \geq 0 \\ 0 & \text{if} \quad \texttt{sign}(b_0' + b_1 x_i) < 0 \end{cases}$$

- The signal is still linear but we apply a non-linear transformation to it: $\phi(x) = \texttt{sign}(b_0' + b_1 x_i)$



Third Approach: Conditional Means

- For example, we consider a patient $x = 24$ years old, and we count the relative frequency of `chd` cases. We compute the conditional mean: $avg(y_i \mid x_i = 24)$
- Similarly, we could compute all conditional means for all age values:

$$(\bar{y} \mid x_i = 25), \quad (\bar{y} \mid x_i = 26), \quad \ldots (\bar{y} \mid x_i = 70),$$

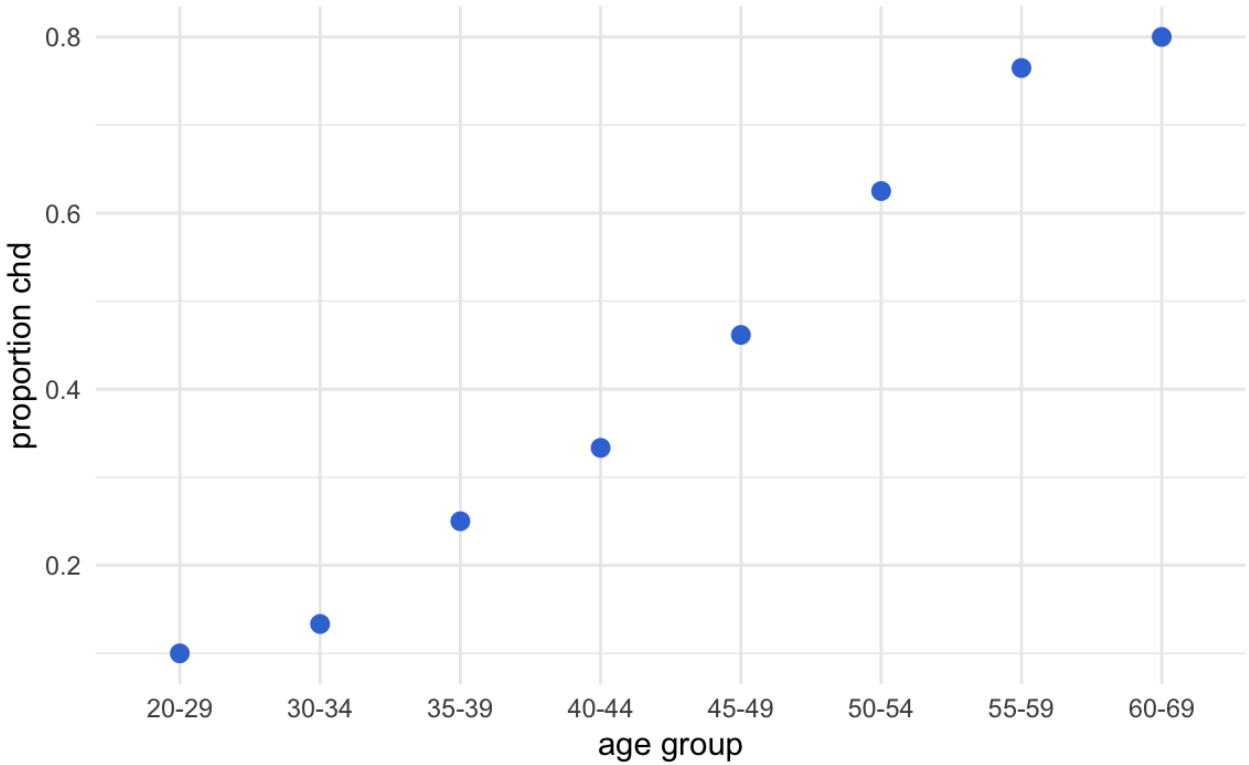
- If we do not have data points for a specific x-value, we can use groups of ages. For example, we compute the proportion of `chd` cases in the group of ages $20 - 29$ years:

$$avg(y_i \mid x_i \in \{20 - 29 \text{ years}\})$$

Third Approach: Conditional Means

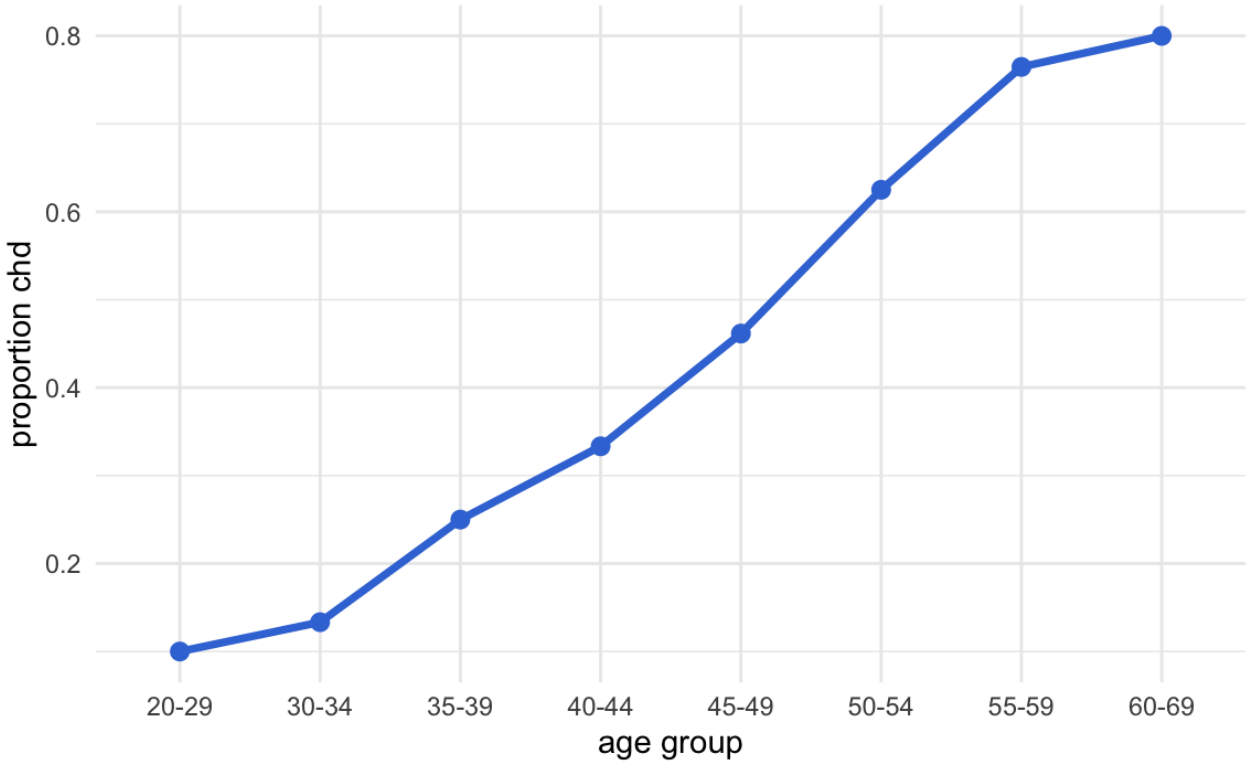
- For each age group, we calculate the proportion of `chd` cases:

$$avg(y_i \mid x_i \in \text{ age group}\})$$



Third Approach: Conditional Means

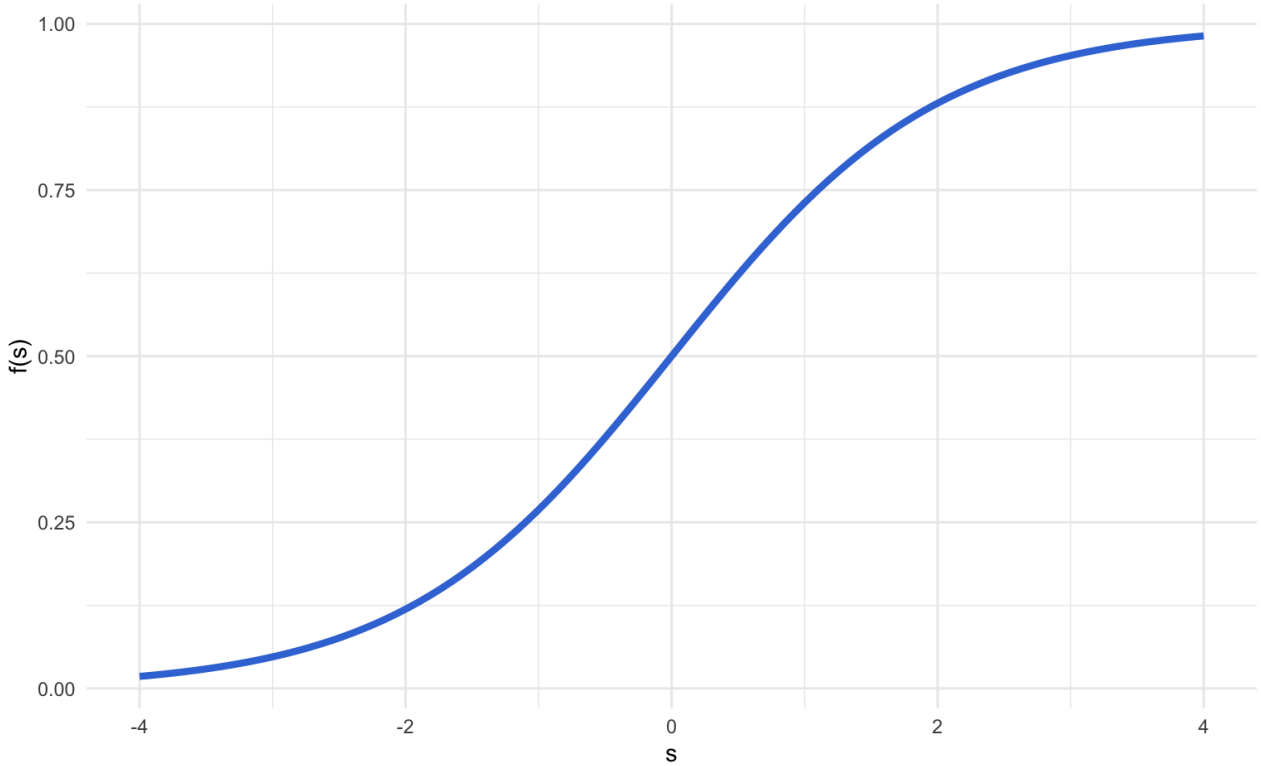
- Theoretical we are modeling the conditional expectations: $\mathbb{E}(y \mid x)$. This is the regression function.
- Connecting the averages, we have a sigmoid pattern:



Third Approach: Conditional Means

- The pattern can be approximated by some mathematical functions.
- The most popular is the logistic function:

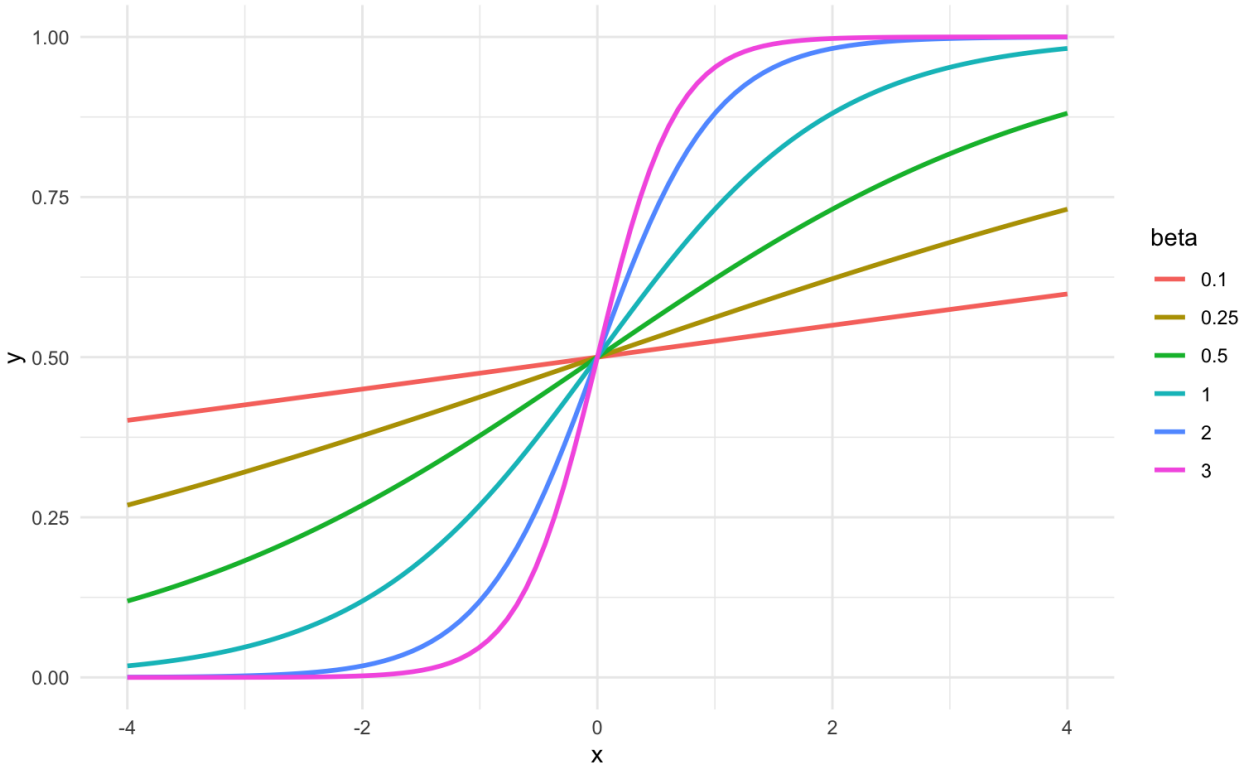
$$f(s) = \frac{e^s}{1 + e^s} = \frac{1}{1 + e^{-s}}$$



Third Approach: Conditional Means

- Replacing the signal s by a linear model $\beta_0 + \beta_1 x$, we have:

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



- Since probability values range inside [0,1], instead of using a line to approximate these values, we should use a more adequate curve.

Logistic Regression Model

- We consider the model:

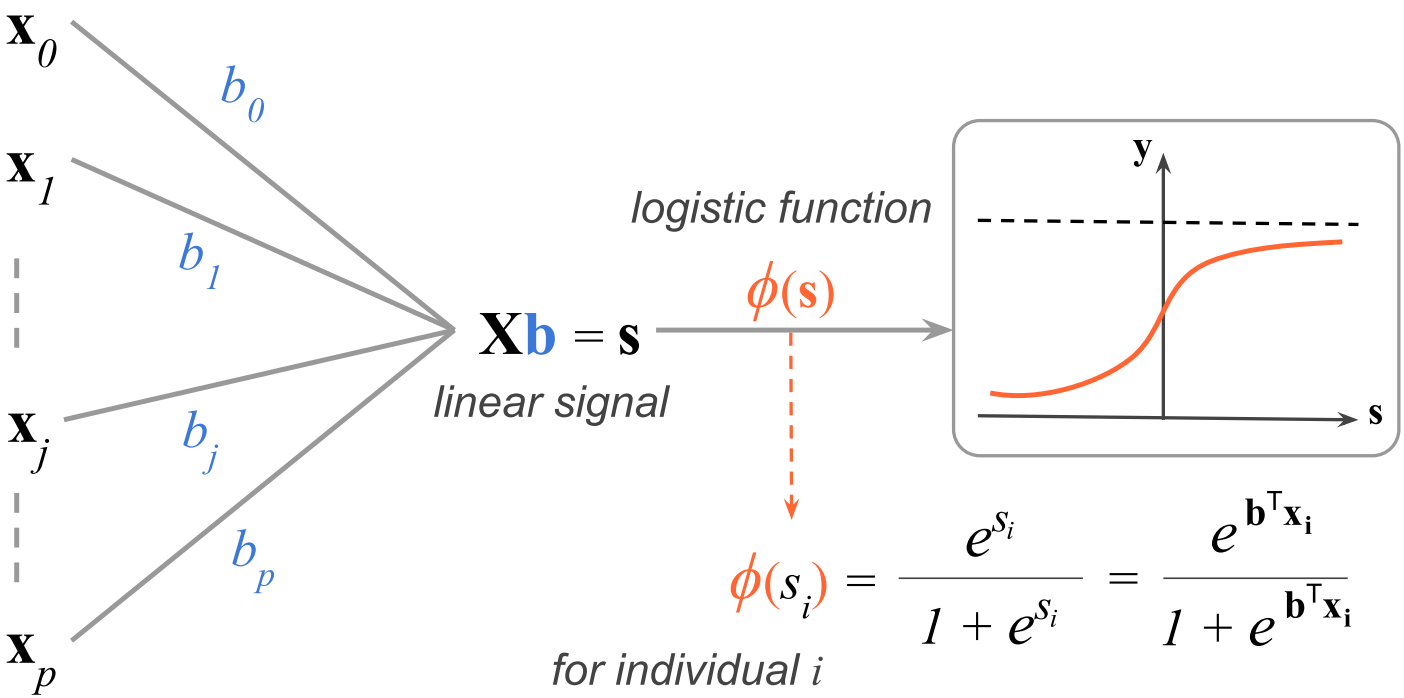
$$Prob(y \mid \mathbf{x}; \mathbf{b}) = f(\mathbf{x})$$

- However, we don't observe the true probability $\in [0, 1]$. We only observe the noisy targets $y_i \in \{0, 1\}$ that is generated by the probability $f(\mathbf{x})$.
- We model the probability by using a mathematical function that has the sigmoid shape, and the most popular function is the logistic function.

$$Prob(y_i \mid \mathbf{x}_i; \mathbf{b}) = \frac{e^{\mathbf{b}^\top \mathbf{x}_i}}{1 + e^{\mathbf{b}^\top \mathbf{x}_i}}$$

where $\mathbf{x}_i$ represents the vector of features of individual i.

Logistic Regression Model



$$Prob(y_i \mid \mathbf{x}_i; \mathbf{b}) = \begin{cases} h(\mathbf{x}_i) & \text{for } y_i = 1 \\ 1 - h(\mathbf{x}_i) & \text{for } y_i = 0 \end{cases}$$

where $h()$ denotes the logistic function $\phi()$:

$$h(\mathbf{x}) = \phi(\mathbf{b}^\top \mathbf{x})$$

The Criterion Being Optimized

$$Prob(y_i \mid \mathbf{x}_i; \mathbf{b}) = \begin{cases} h(\mathbf{x}_i) & \text{for } y_i = 1 \\ 1 - h(\mathbf{x}_i) & \text{for } y_i = 0 \end{cases}$$

- Assuming that our model is true, i.e., $h(x) = f(x)$, we ask "how likely is it that we observe the data we already observed (y_i)?"

$$\begin{aligned} Prob(\mathbf{y} \mid x_1, x_2, \ldots, x_p; \mathbf{b}) &= \prod_{i=1}^{n} P(y_i \mid \mathbf{b}^\top \mathbf{x}_i) \\ &= \prod_{i=1}^{n} h(\mathbf{b}^\top \mathbf{x}_i)^{y_i} [1 - h(\mathbf{b}^\top \mathbf{x}_i)]^{1-y_i} \end{aligned}$$

The Criterion Being Optimized

- Take the logarithm (log-likelihood)

$$
\begin{aligned}
l(\mathbf{b}) &= \ln\left[L(\mathbf{b})\right] \\
&= \sum_{i=1}^{n} \ln\left[P(y_i \mid \mathbf{b}^\top \mathbf{x}_i)\right] \\
&= \sum_{i=1}^{n} \{y_i \ln\left[h(\mathbf{b}^\top \mathbf{x}_i)\right] + (1-y_i) \ln\left[1-h(\mathbf{b}^\top \mathbf{x}_i)\right]\} \\
&= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{e^{\mathbf{b}^\top \mathbf{x}_i}}{1+e^{\mathbf{b}^\top \mathbf{x}_i}}\right) + (1-y_i) \ln\left(1-\frac{e^{\mathbf{b}^\top \mathbf{x}_i}}{1+e^{\mathbf{b}^\top \mathbf{x}_i}}\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \mathbf{b}^\top \mathbf{x}_i - \ln\left(1+e^{\mathbf{b}^\top \mathbf{x}_i}\right)\right]
\end{aligned}
$$

- Differentiating and setting to 0 yields an equation for which no closed-form solution exists.

The Criterion Being Optimized

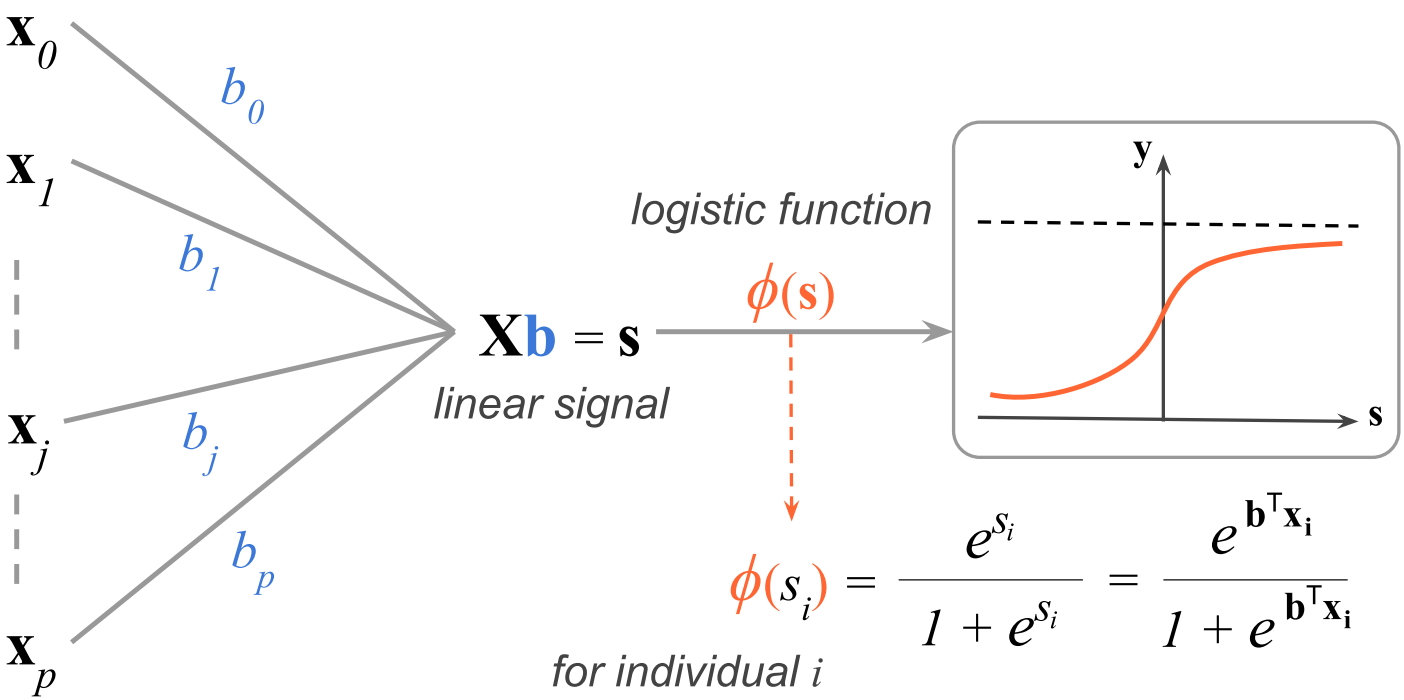
- We need to use gradient ascent.
- The gradient of the likelihood function is:

$$\begin{aligned}\nabla l(\mathbf{b}) &= \sum_{i=1}^{n}\left[y_i\mathbf{x}_i - \left(\frac{e^{\mathbf{b}^\top\mathbf{x}_i}}{1+e^{\mathbf{b}^\top\mathbf{x}_i}}\right)\mathbf{x}_i\right] \\ &= \sum_{i=1}^{n}[y_i\mathbf{x}_i - \phi(\mathbf{b}^\top\mathbf{x}_i)\mathbf{x}_i] \\ &= \sum_{i=1}^{n}[y_i - \phi(\mathbf{b}^\top\mathbf{x}_i)]\mathbf{x}_i\end{aligned}$$

- In gradient ascent, we would update the parameters $\mathbf{b}$:

$$\mathbf{b}^{(s+1)} = \mathbf{b}^{(s)} + \alpha\nabla l(\mathbf{b}^{(s)})$$

Another Way to Solve Logistic Regression



What it used to be $y_i = 0$, let's encode it as $y_i = -1$

$$Prob(y_i \mid \mathbf{x}_i; \mathbf{b}) = \begin{cases} h(\mathbf{x}_i) & \text{for } y_i = 1 \\ 1 - h(\mathbf{x}_i) & \text{for } y_i = -1 \end{cases}$$

where $h()$ denotes the logistic function $\phi()$: $h(\mathbf{x}) = \phi(\mathbf{b}^\top \mathbf{x})$

Another Way to Solve Logistic Regression

- The logistic function has the property:

$$\phi(-s) = \frac{e^{-s}}{1+e^{-s}} = \frac{1-1+e^{-s}}{1+e^{-s}} = 1 - \frac{1}{1+e^{-s}} = 1 - \phi(s)$$

- We can update the expression for the conditional probability:

$$Prob(y_i \mid \mathbf{x}_i; \mathbf{b}) = \begin{cases} h(\mathbf{x}_i) & = \phi(y_i \mathbf{b}^\top \mathbf{x}) \text{ for } y_i = 1 \\ 1 - h(\mathbf{x}_i) & = \phi(y_i \mathbf{b}^\top \mathbf{x}) \text{ for } y_i = -1 \end{cases}$$

- This implies that we only need one case regardless of y_i:

$$Prob(y_i \mid \mathbf{x}_i; \mathbf{b}) = \phi(y_i \mathbf{b}^\top \mathbf{x})$$

Another Way to Solve Logistic Regression

- We can compute the likelihood as previously:

$$Prob(\mathbf{y} \mid x_1, x_2, \ldots, x_p; \mathbf{b}) = \prod_{i=1}^{n} P(y_i \mid \mathbf{b}^\top \mathbf{x}_i)$$
$$= \prod_{i=1}^{n} \phi(y_i \mathbf{b}^\top \mathbf{x}_i)$$

- then log-likelihood:

$$l(\mathbf{b}) = \ln[L(\mathbf{b})] = \sum_{i=1}^{n} \ln[\phi(y_i \mathbf{b}^\top \mathbf{x}_i])$$
$$\Rightarrow \frac{1}{n} \sum_{i=1}^{n} \ln[\phi(y_i \mathbf{b}^\top \mathbf{x}_i)]$$

- Instead of maximizing the log-likelihood, we can minimize the negative log-likelihood

$$\min_{\mathbf{b}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln[\phi(y_i \mathbf{b}^\top \mathbf{x}_i)] \right\}$$

Another Way to Solve Logistic Regression

$$\min_{\mathbf{b}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left[\phi(y_i \mathbf{b}^\top \mathbf{x}_i) \right] \right\} \Leftrightarrow \min_{\mathbf{b}} \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ln \left(1 + e^{-y_i \mathbf{b}^\top \mathbf{x}_i} \right)}_{\text{pointwise error}} \right\}}_{E_{in}(\mathbf{b})}$$

- Focus on the product term between the response and the linear signal: $y_i \mathbf{b}^\top \mathbf{x}_i = y_i \mathbf{s}$ where $\mathbf{s}$ represents "signal".
- A small signal means that the probability $\phi(\mathbf{b}^\top \mathbf{x}_i)$ will be small.
- A large signal means that the probability $\phi(\mathbf{b}^\top \mathbf{x}_i)$ will be large.
- The term y_i can be either -1 or $+1$.

Another Way to Solve Logistic Regression

$$\min_{\mathbf{b}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left[\phi(y_i \mathbf{b}^\top \mathbf{x}_i) \right] \right\} \Leftrightarrow \min_{\mathbf{b}} \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ln \left(1 + e^{-y_i \mathbf{b}^\top \mathbf{x}_i} \right)}_{\text{cross-entropy error}} \right\}}_{E_{in}(\mathbf{b})}$$

- With correct predictions, $e^{-y_i \mathbf{b}^\top \mathbf{x}_i}$ will be small, and will give a small error.
- With incorrect predictions, $e^{-y_i \mathbf{b}^\top \mathbf{x}_i}$ will be large, and will give a large error.
- We need to find $\mathbf{b}$ to minimize $E_{in}(\mathbf{b})$. We need to use gradient descent with the gradient w.r.t. $\mathbf{b}$ can be computed as:

$$\nabla E_{in}(\mathbf{b}) = -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{e^{-y_i \mathbf{b}^\top \mathbf{x}_i}}{1 + e^{-y_i \mathbf{b}^\top \mathbf{x}_i}} \right) y_i \mathbf{x}_i = -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{1 + e^{y_i \mathbf{b}^\top \mathbf{x}_i}} \right) y_i \mathbf{x}_i$$