Classification and Logistic Regression

Practicum

- The **Practicum** is posted. It is due at 11:59pm on Wednesday December 13th.
- **The Rules**:
  - ❑ All work must be your own. Collaboration of any kind is is not permitted.
  - ❑ You may use any resources you like, but you may not post to message boards or other online resources asking for help.
  - ❑ We will answer general, clarifying questions in office hours.
  - ❑ If you have a question for us, post a **PRIVATE** message on Piazza.

Previously on CSCI 3022

Given data $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ for $i = 1, 2, \ldots, n$ fit a MLR model of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim N(0, \sigma^2)$$

After learning weights, if we want to make a prediction about a new data point

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_p x_p$$

Regression as Prediction

- So far, we've learned about various forms of **regression**
- We've viewed regression in terms of learning a relationship between one or more features and a response

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- We've also talked about regression as a way to make **predictions**

Predicting Survival

- Based on our previous experience, it might be tempting to use linear regression as a classifier

Example:

	age	outcome
0	55	survived
1	56	survived
2	57	survived
3	58	survived
4	59	died
5	60	died
6	61	died
7	62	died

classes

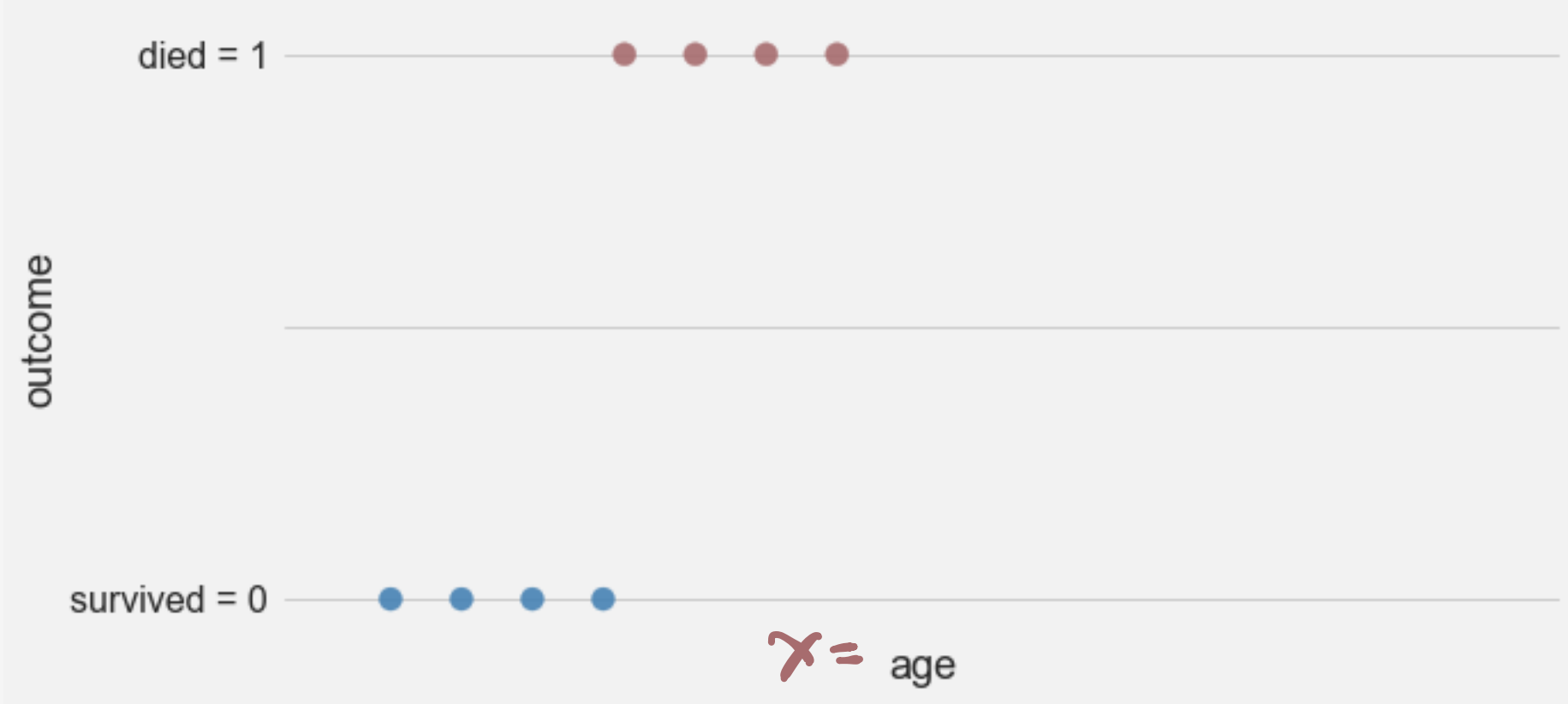
Recode the outcomes as y = {0, 1}

	age	outcome
0	55	0
1	56	0
2	57	0
3	58	0
4	59	1
5	60	1
6	61	1
7	62	1

- Perform linear regression to take a feature x and predict the response y

Predicting Survival

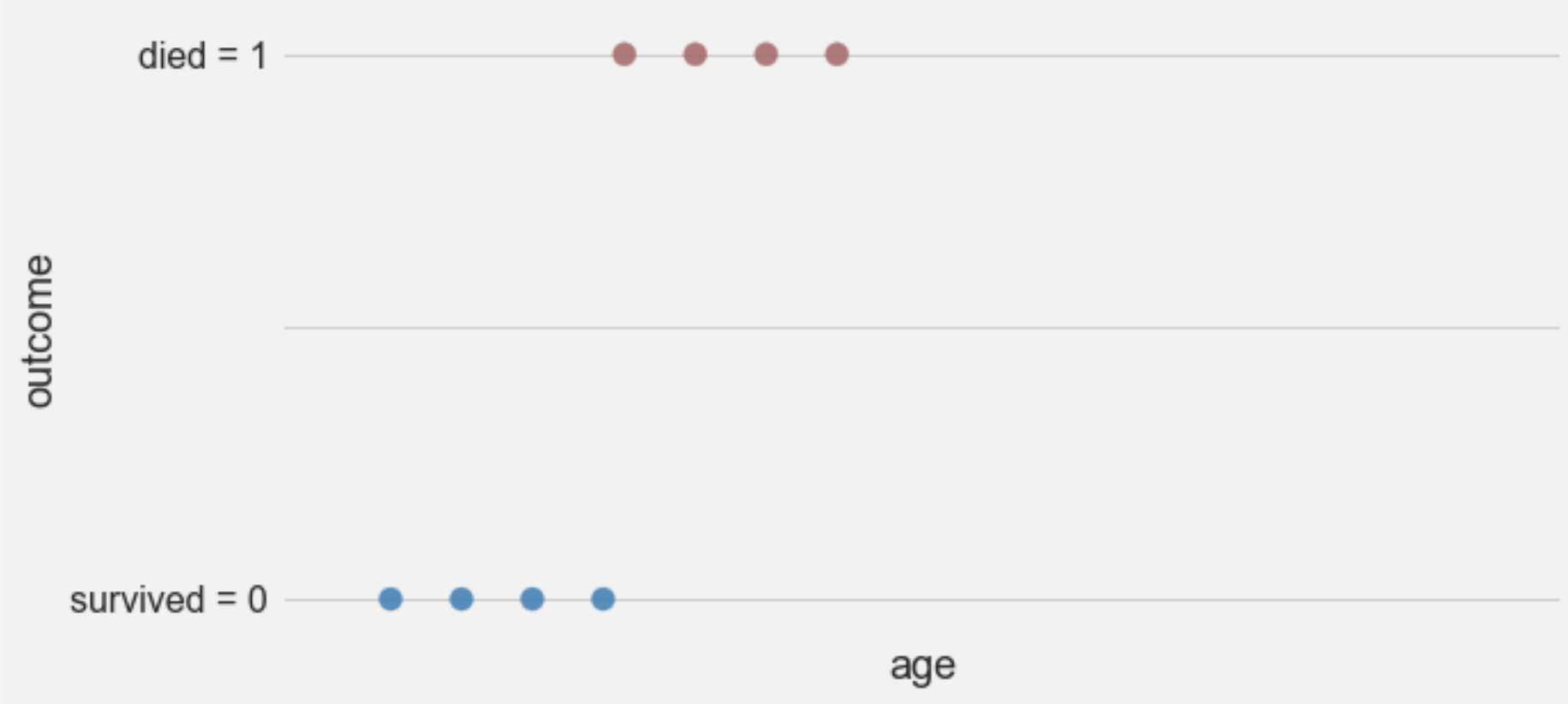
Example: Suppose you want to predict whether a passenger on the Titanic survived or not using passenger age as the sole feature.



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

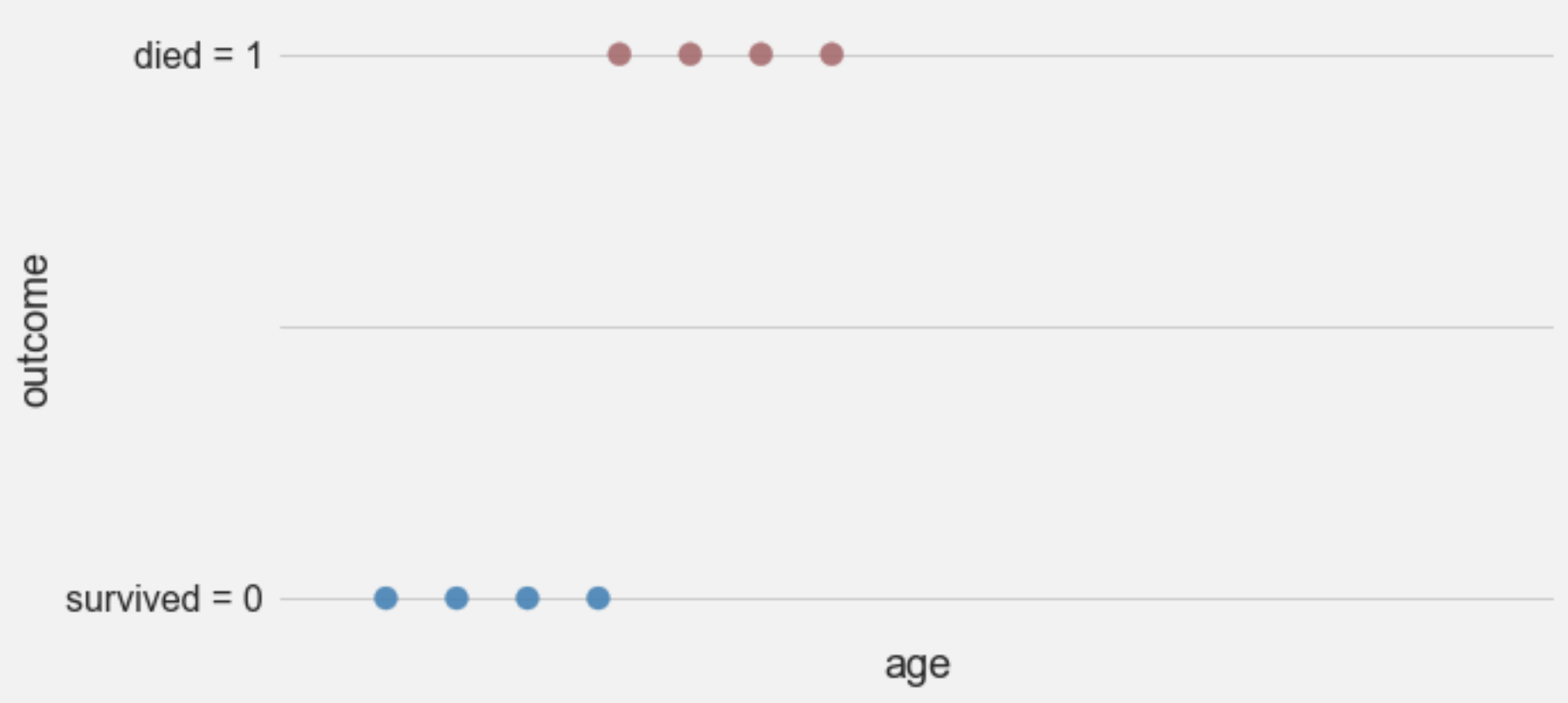
The output from the model is the prediction: $y = \{0, 1\} =$ {survived, died}



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

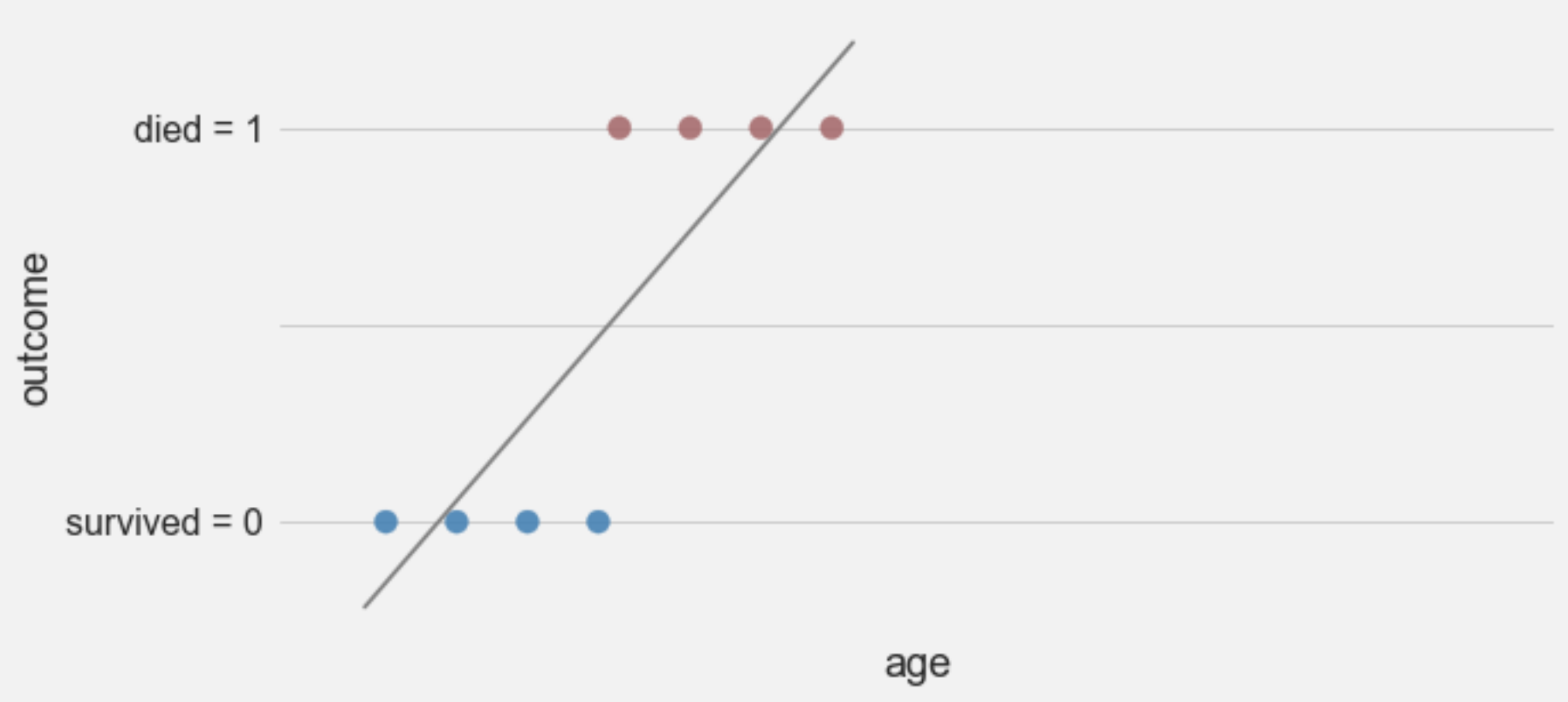
Question: How should we model the relation between feature and response?



Predicting Survival

The input to the model is a single feature: $x_1 = \text{age}$

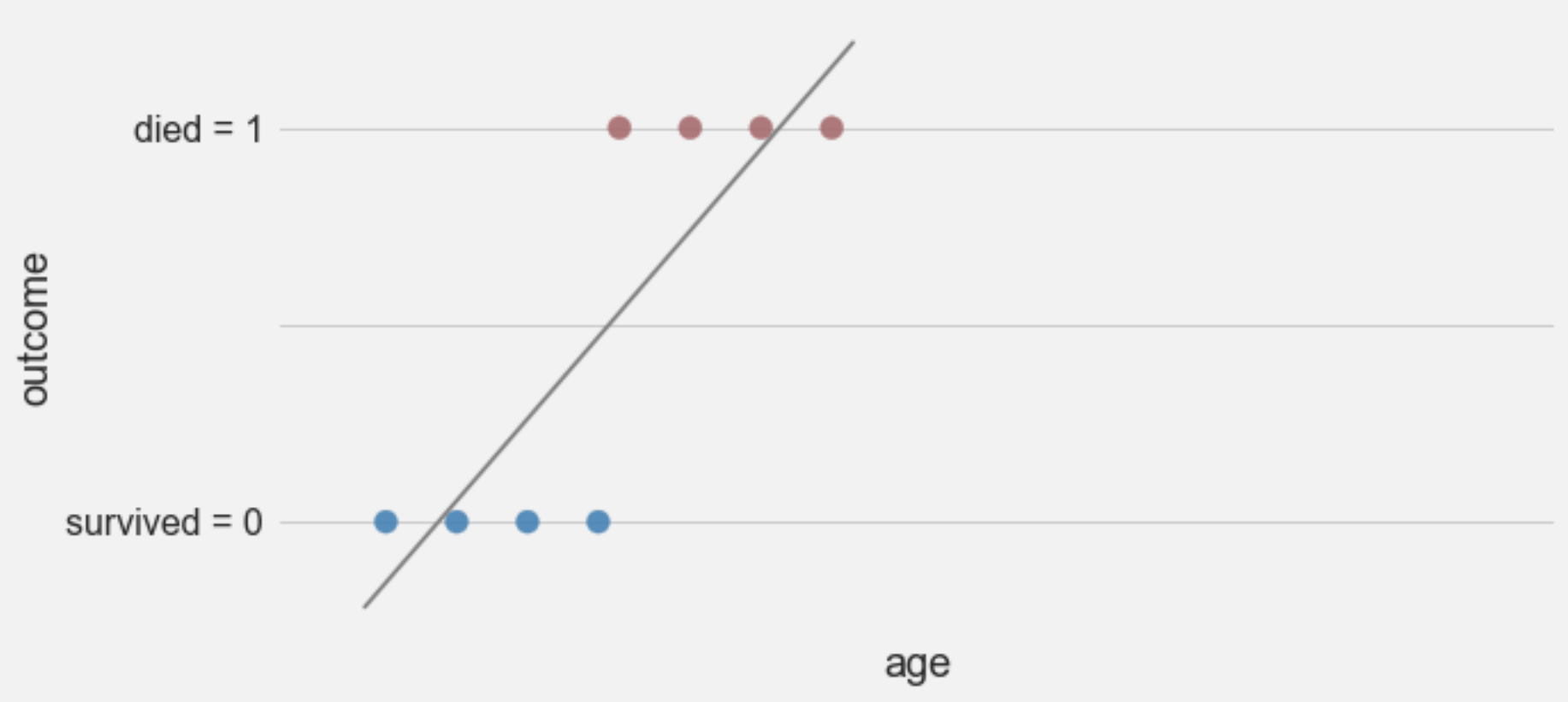
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

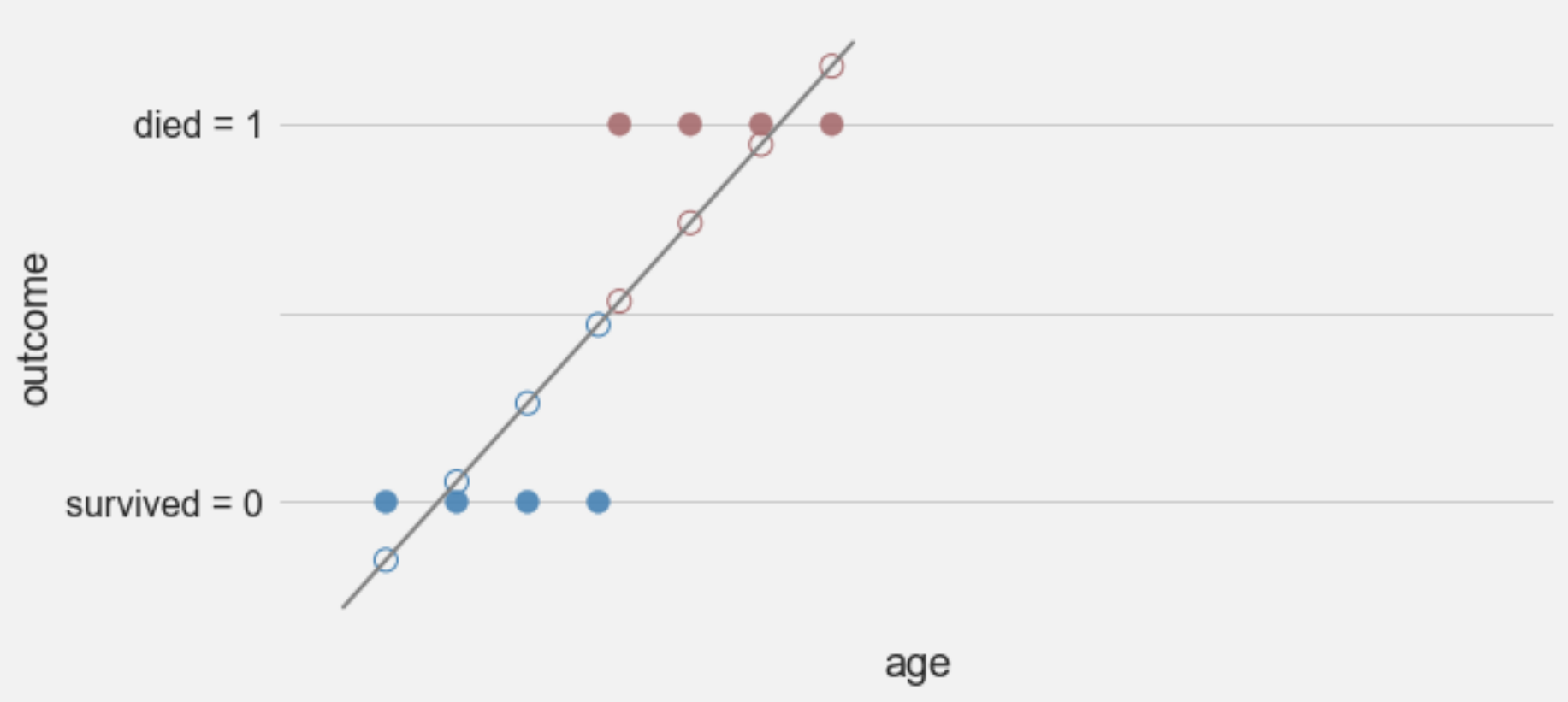
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Survival

The input to the model is a single feature: $x_1 = \text{age}$

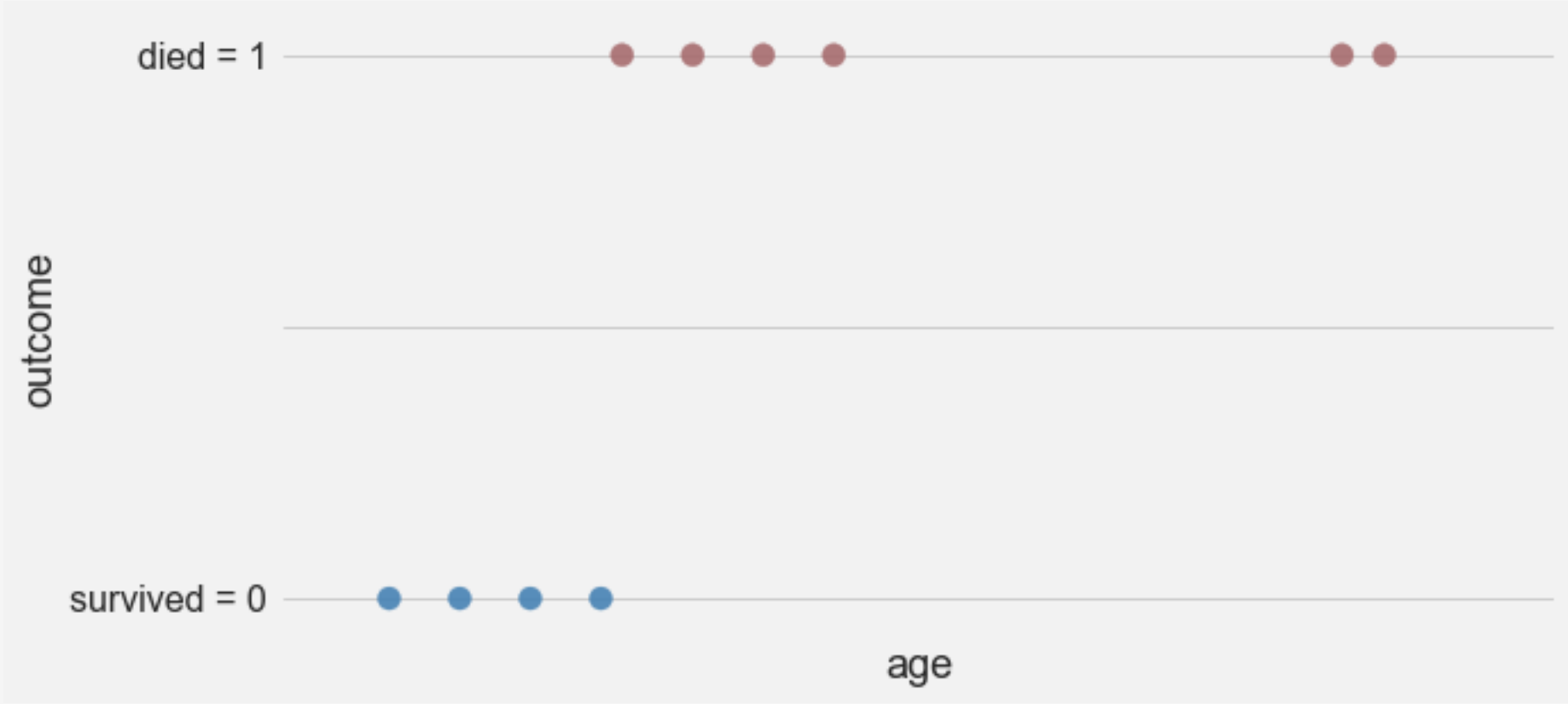
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

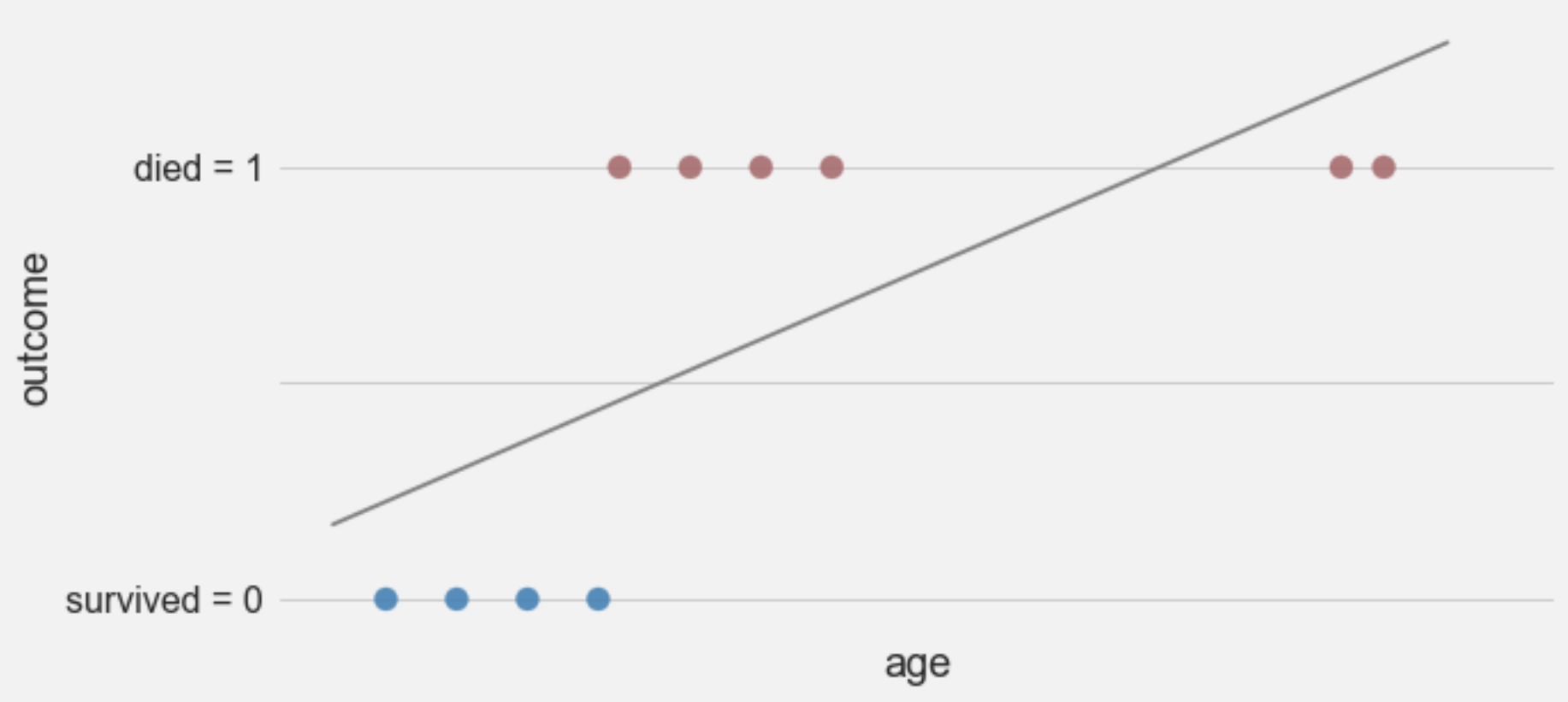
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Survival

The input to the model is a single feature: $x_1 = \text{age}$

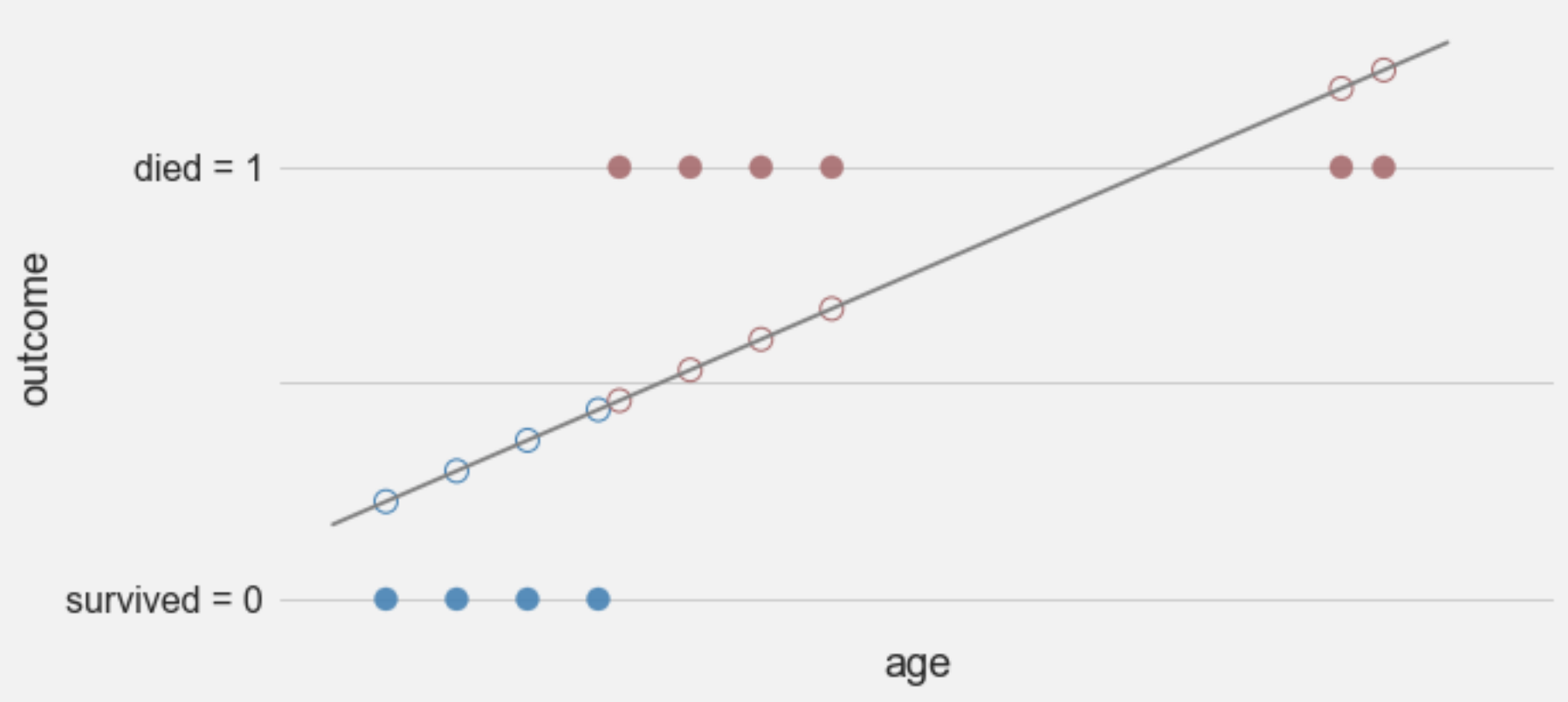
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Survival

The input to the model is a single feature: $x_1 = \text{age}$

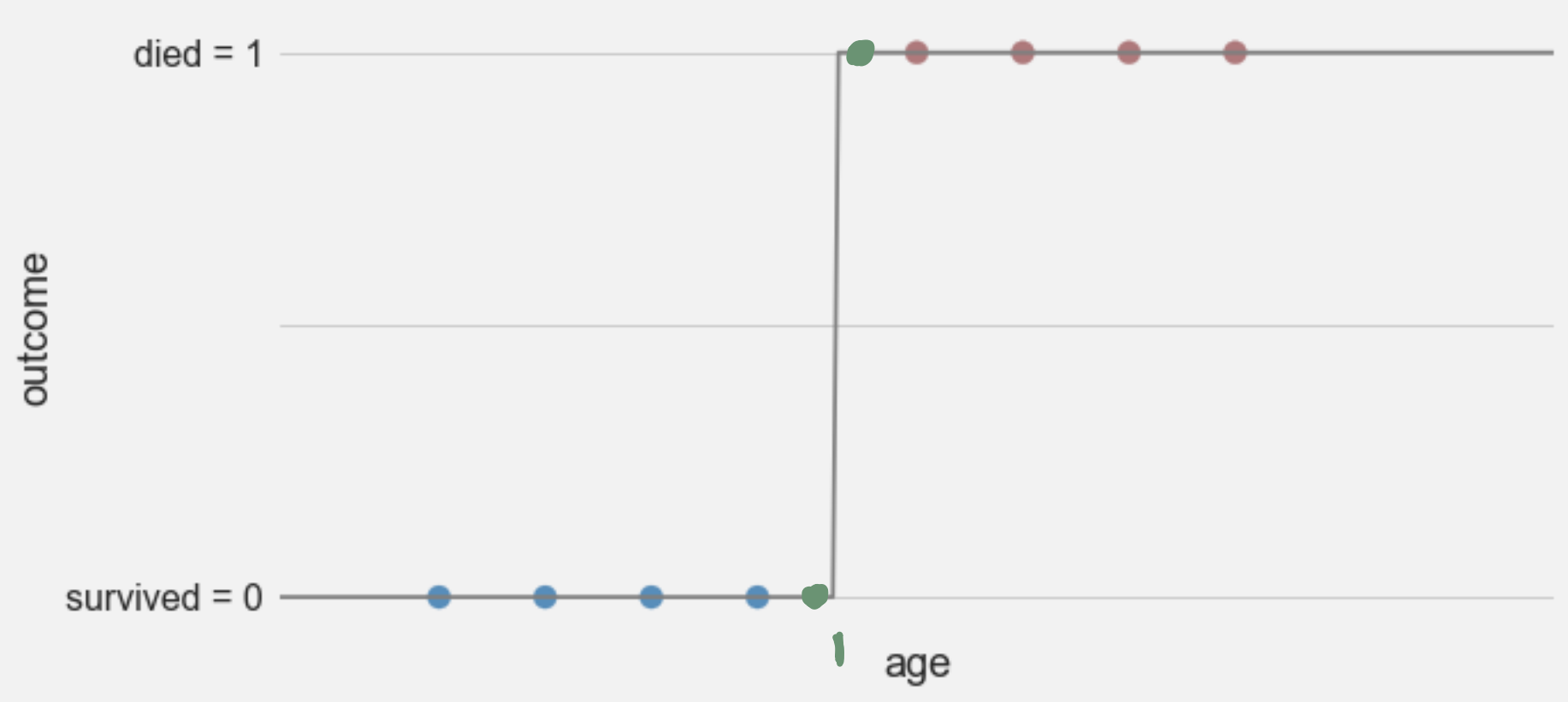
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Survival

The input to the model is a single feature: $x_1 = \text{age}$

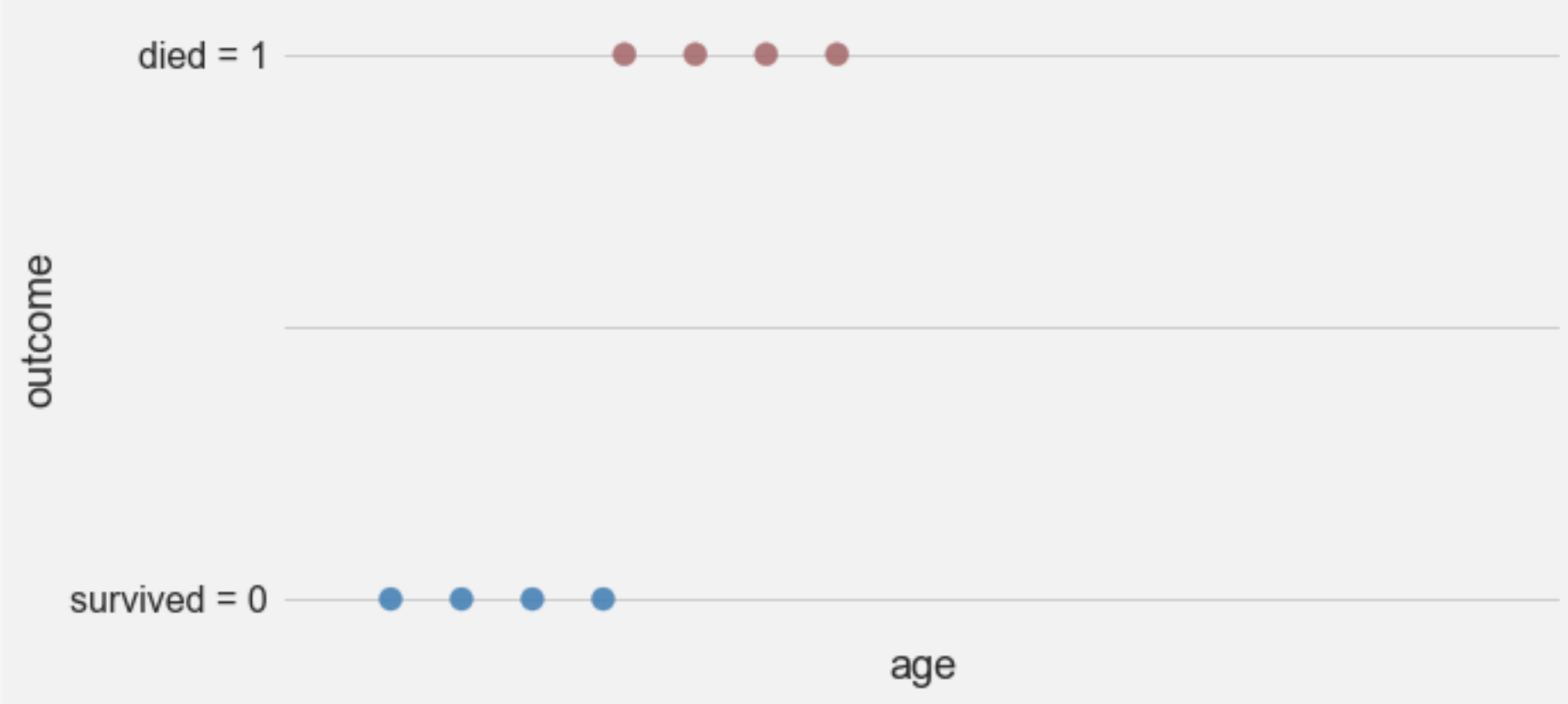
Idea: Use a piecewise function (aka perceptron) $y = \begin{cases} 1 & \text{if } \beta_0 + \beta_1 x_1 > 0 \\ 0 & \text{otherwise} \end{cases}$



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

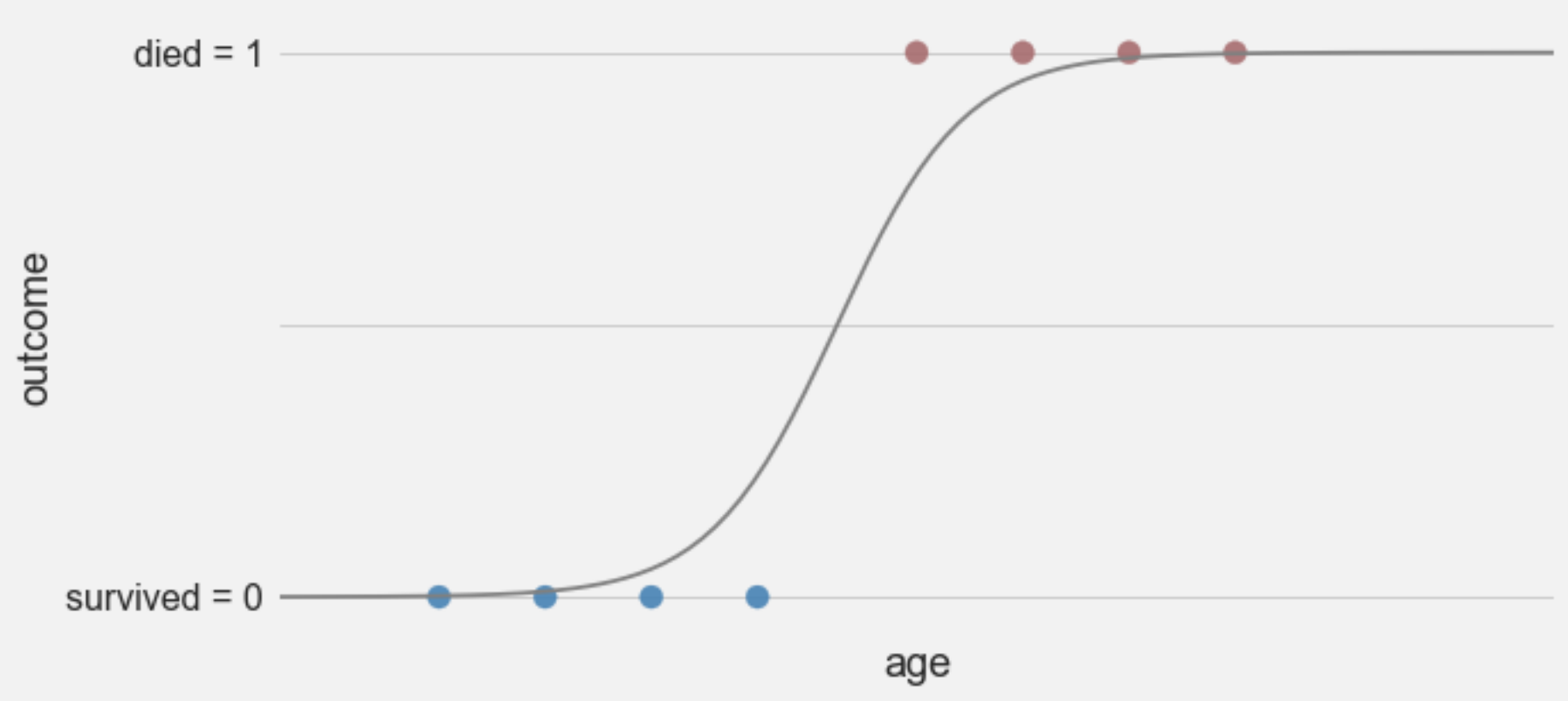
Idea: Need something that behaves more like a probability …



Predicting Survival

The input to the model is a single feature: $x_1 = \text{age}$

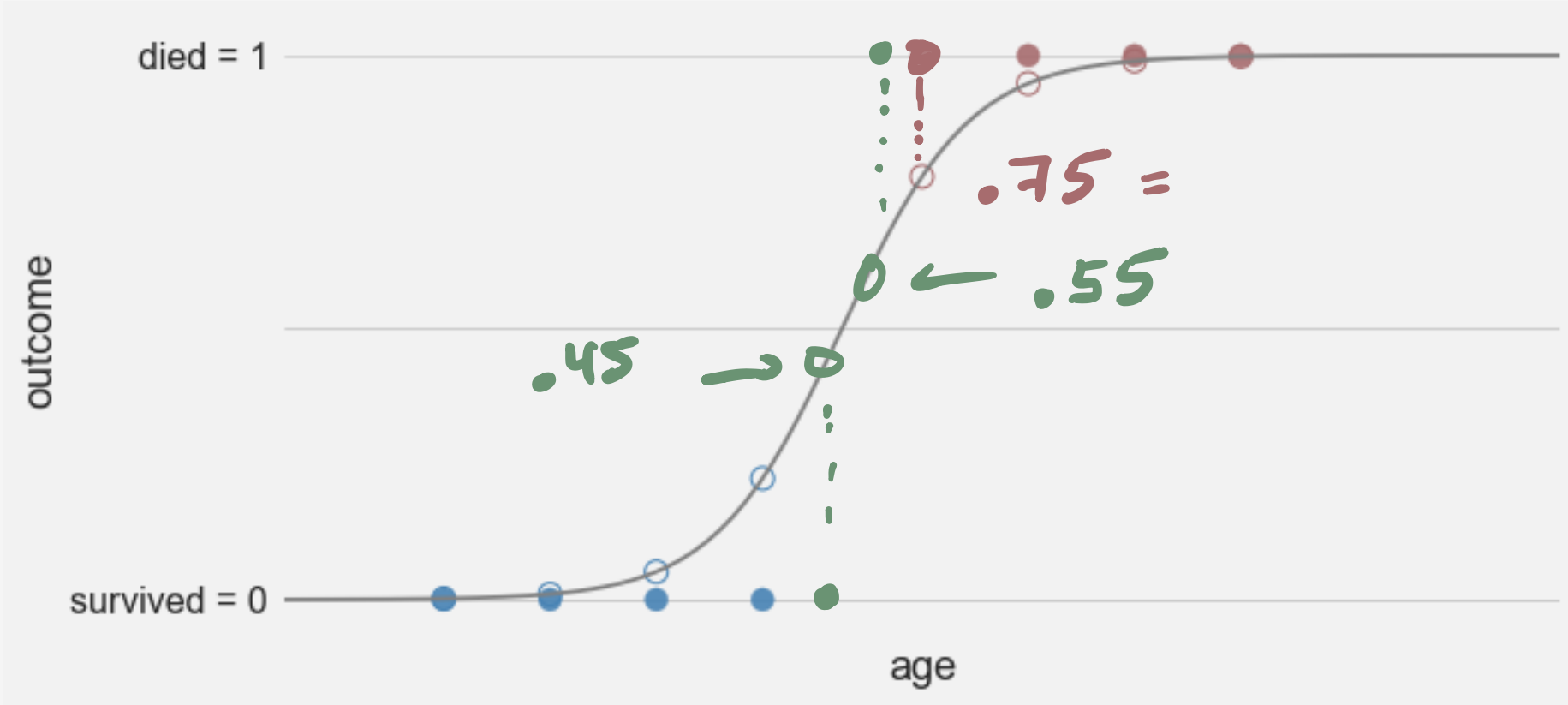
Idea: Need something that behaves more like a probability …



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

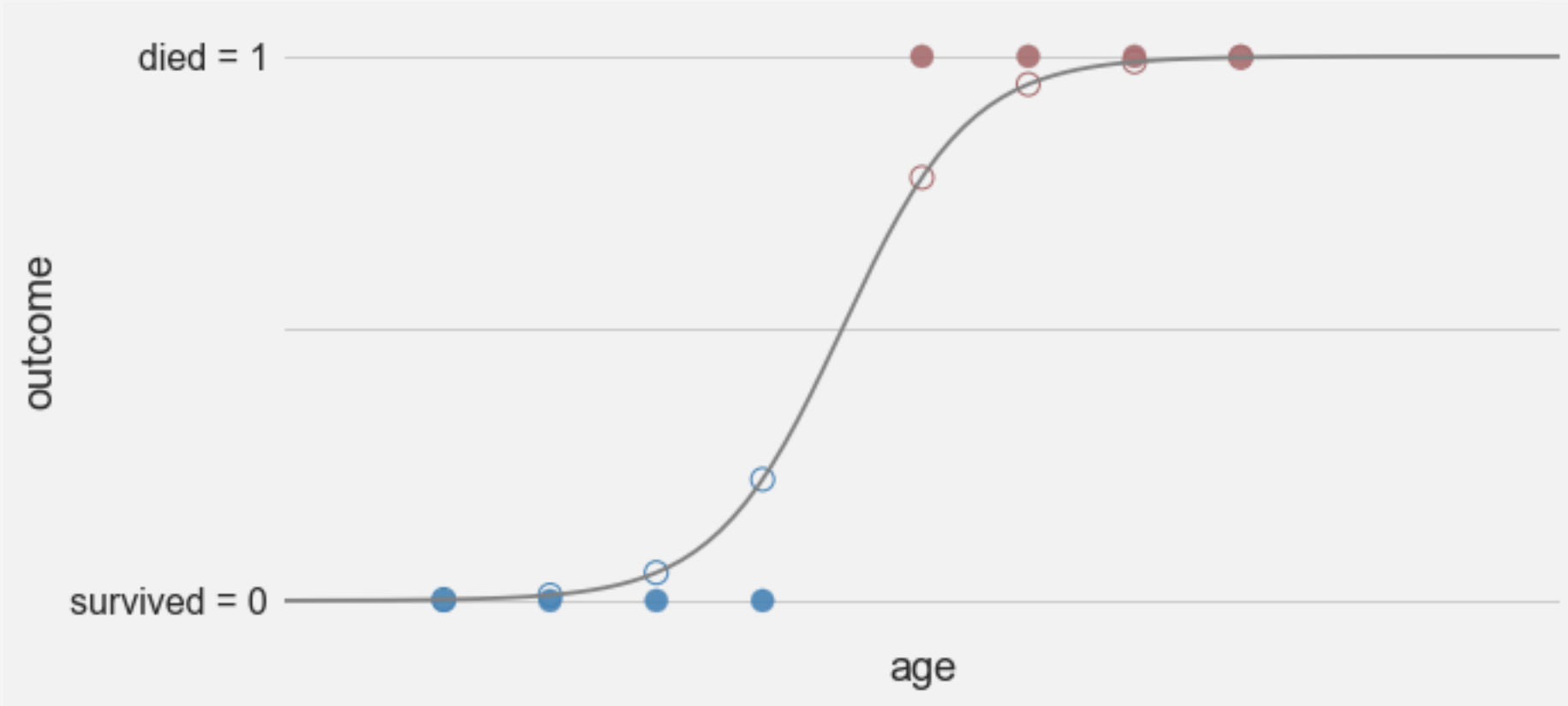
Idea: Need something that behaves more like a probability …



Predicting Survival

The input to the model is a single feature: $x_1 =$ age

Question: What kind of function does this?



The Sigmoid Function

It has Everything!

$$\text{sigm}(z) = \frac{1}{1 + \text{Exp}[-z]}$$

$z \to \infty \Rightarrow \frac{1}{1+0} = 1$

$z \to -\infty \Rightarrow \frac{1}{1+\infty} = 0$

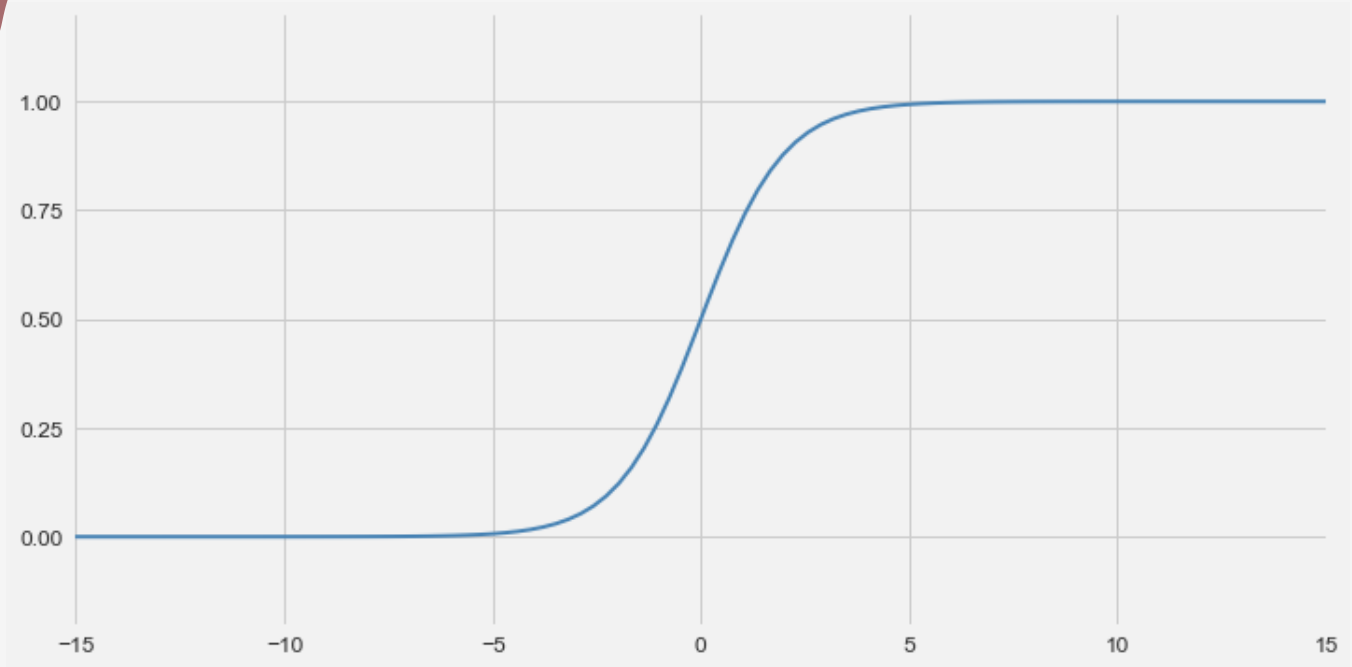
Behaves like a Probability $0 < \text{sigm}(z) < 1$

Distinguishes Between Points

Really Smooth

$z = 0$

$$\frac{1}{1+\text{Exp}[-0]} = \frac{1}{1+1} = \frac{1}{2}$$

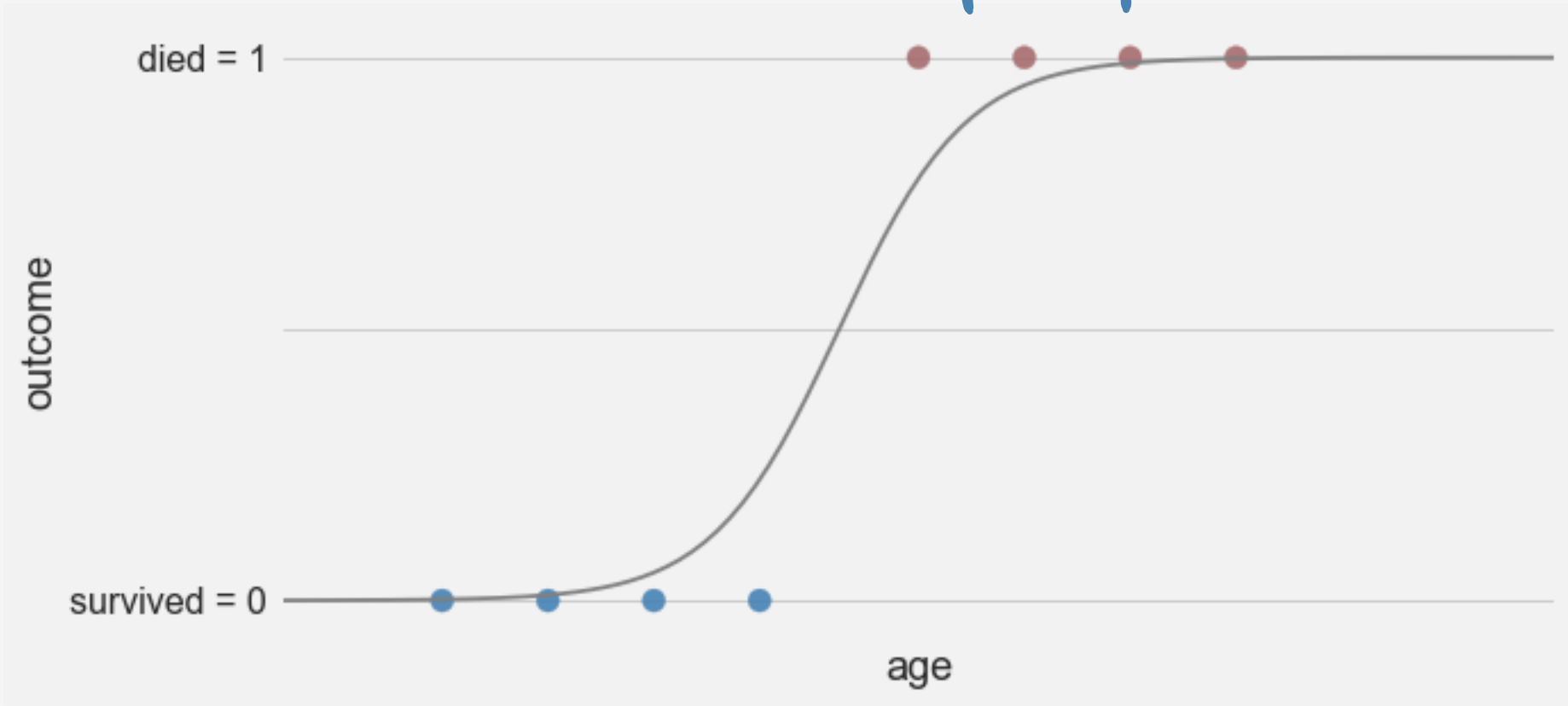


Logistic Regression

The Model:

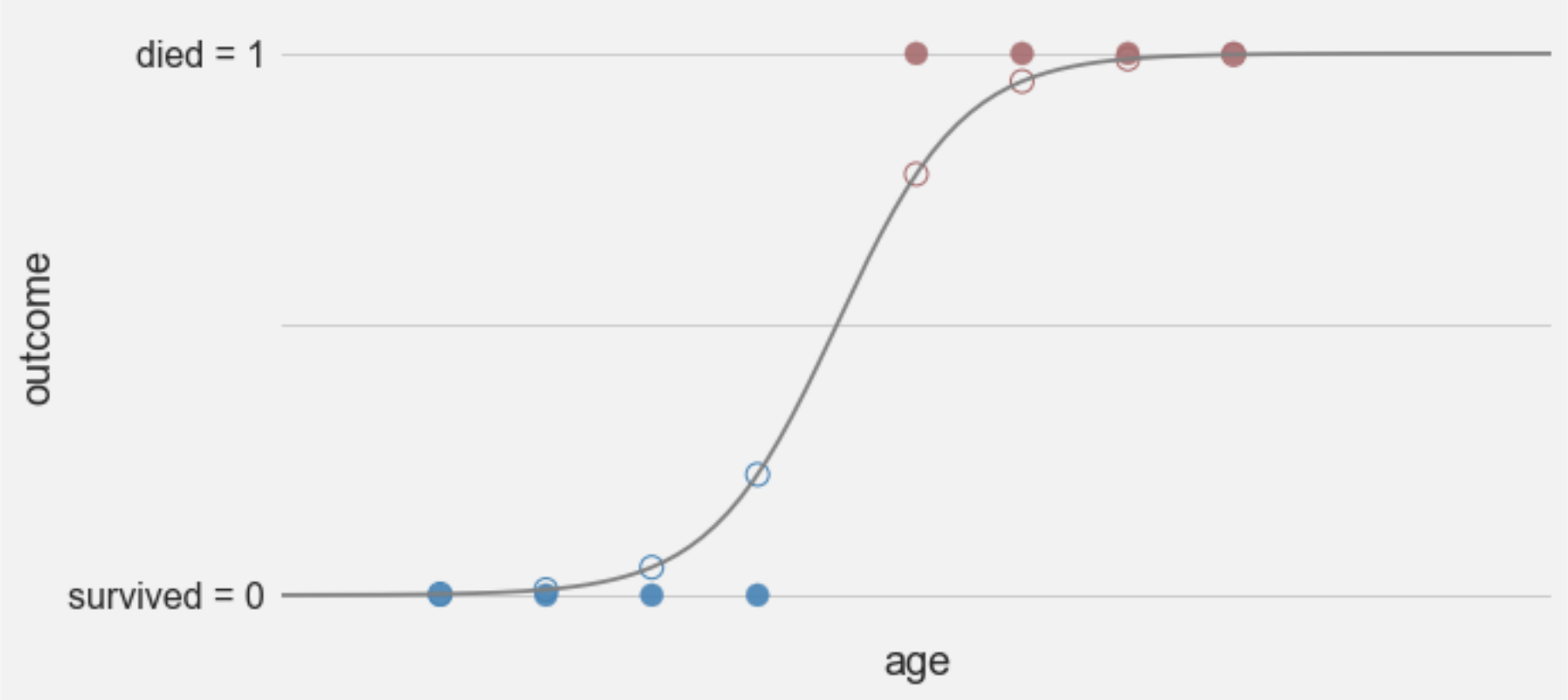
$$p(y=1|x) = \text{sigm}(\beta_0 + \beta_1 x)$$

Learn the weights from the data



Logistic Regression

Classify data point x according to $\hat{y} = \begin{cases} 1 & \text{if } \text{sigm}(\hat{\beta}_0 + \hat{\beta}_1 x) \geq 0.5 \\ 0 & \text{if } \text{sigm}(\hat{\beta}_0 + \hat{\beta}_1 x) < 0.5 \end{cases}$



An Odd(s) View of Logistic Regression

Our inevitable path to Logistic Regression and the sigmoid function began with our insistence on modeling on modeling the relationship between the features and the response as a bonafide probability.

It turns out that through some basic algebra we can arrive at an interpretation of Logistic Regression that is very regression-like.

But first we have to put on our gambling hats and talk about **odds**.

An Odd(s) View of Logistic Regression

In statistics, the odds of an event occurring is the ratio of the probability that the event occurs divided by the probability that the event does not occur, and then generally flipped to get a value bigger than 1

$$\text{odds} = \frac{p}{1-p}$$

$\frac{.75}{.25} = 3$

Example 1: If $p = 0.75$ then $\text{odds} =$ 3

We would say the odds are 3 to 1 IN FAVOR

$\frac{0.1}{0.9} = \frac{1}{9}$

Example 2: If $p = 0.1$ then $\text{odds} =$ 1/9

We would say the odds are 1 to 9 against

An Odd(s) View of Logistic Regression

In Logistic Regression we model $p = p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x)$

If instead we compute the odds that y = 1 given the data, we have

$$\text{ODDS} = \frac{\text{sigm}(\beta_0 + \beta_1 x)}{1 - \text{sigm}(\beta_0 + \beta_1 x)} = \text{Exp}[\beta_0 + \beta_1 x]$$

$$\ln(\text{ODDS}) = \beta_0 + \beta_1 x$$

An Odd(s) View of Logistic Regression

Taking the natural log of both sides, gives

$$\log(\text{odds}) = \beta_0 + \beta_1 x$$

So it turns out we **have** been doing linear regression all along, but for the **log-odds** instead of the probability!

Backing up a step, we had $\text{odds} = \exp(\beta_0 + \beta_1 x)$

$\exp(\beta_0 + \beta_1(x+1))$

This gives us a new interpretation of the Logistic Regression weight β_1

$= \exp((\beta_0 + \beta_1 x) + \beta_1) = \exp(\beta_0 + \beta_1 x)\exp(\beta_1)$

ODDS

An Odd(s) View of Logistic Regression

Logistic Regression with Many Features

The LogReg model with a single feature looks like: $p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x)$

But in real life we typically have many features:

- **Predict** which candidate a person will vote for
- **Features**: education, household income, zip code, religion, etc

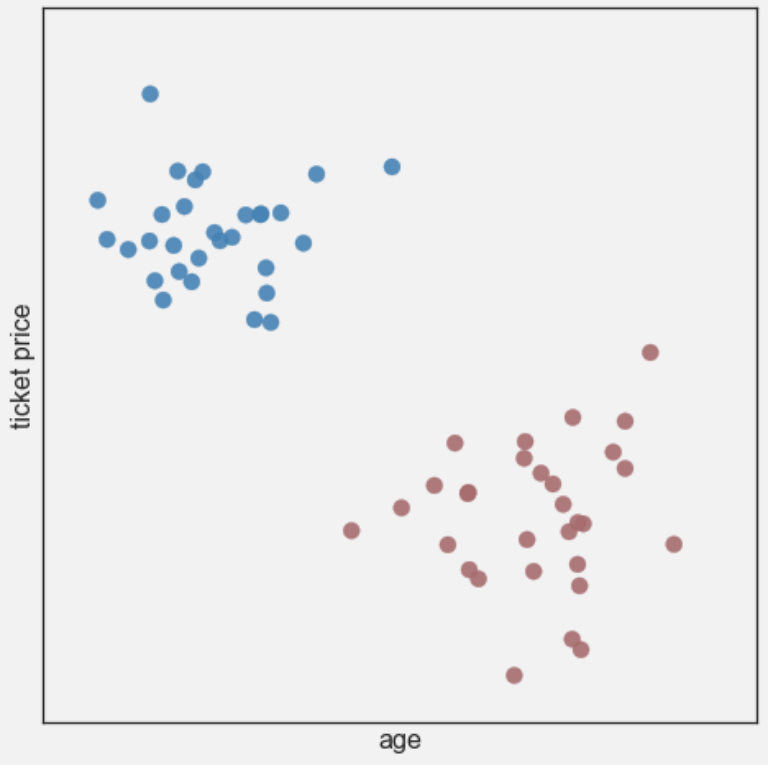
Multiple Feature Logistic Regression Model:

$$p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

Logistic Regression with Many Features

Multiple Feature Logistic Regression Model:

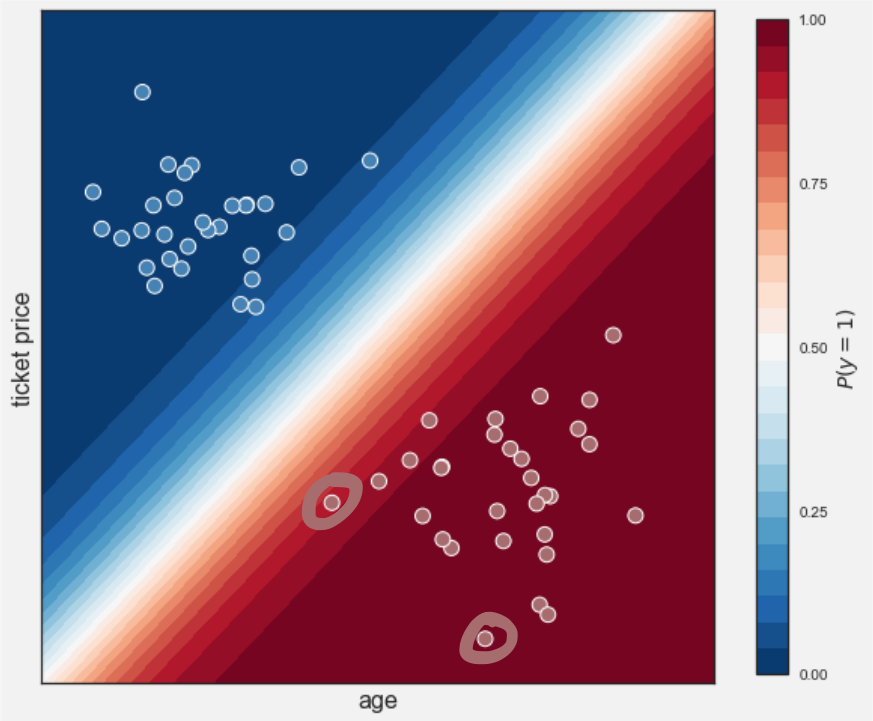
$$p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$$



Logistic Regression with Many Features

Multiple Feature Logistic Regression Model:

$$p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$



OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 25 In-Class Notebook

Let's:

- Explore the sigmoid function
- Figure out how to learn a Logistic Regression model in Python
- Explore Logistic Regression with multiple features

The Decision Boundary

The Decision Boundary is the line/surface that separates predictions into Class 0 and Class 1

For a 2-feature model, it's described by

$$\text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = 0.5$$

Which is just a line in 2D space:

$$\Rightarrow \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

$$\Rightarrow \quad X_2 = -\frac{\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} X_1$$

Properties of the Sigmoid Function

The Sigmoid function has some nice differential properties that we'll explore next time

The most important of which, is that

$$\text{If } f(z) = \text{sigm}(z) \text{ then } f'(z) = \text{sigm}(z)(1 - \text{sigm}(z))$$