# Classification with Logistic Regression

Pavlos Protopapas



# Lecture Outline

- What is Classification?
- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
- Estimating the Simple Logistic Model
- Classification using the Logistic Model
- Multiple Logistic Regression
- Extending the Logistic Model
- Classification Boundaries



# Lecture Outline

- What is Classification?
- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
- Estimating the Simple Logistic Model
- Classification using the Logistic Model
- Multiple Logistic Regression
- Extending the Logistic Model
- Classification Boundaries



# Advertising Data (from earlier lectures)

X
**predictors**
features
covariates

Y
outcome
**response** variable
dependent variable

*n* observations

| TV | radio | newspaper | sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |

***p*** predictors

The response variable is continuous or quantitative

# Heart Data

These data contain a binary outcome AHD for 303 patients who presented with chest pain.

**response** variable Y is Yes/No

| Age | Sex | ChestPain | RestBP | Chol | MaxHR | ExAng | Thal | AHD |
|---|---|---|---|---|---|---|---|---|
| 63 | 1 | typical | 145 | 233 | 150 | 0 | fixed | No |
| 67 | 1 | asymptomatic | 160 | 286 | 108 | 1 | normal | Yes |
| 67 | 1 | asymptomatic | 120 | 229 | 129 | 1 | reversable | Yes |
| 37 | 1 | nonanginal | 130 | 250 | 187 | 0 | normal | No |
| 41 | 0 | nontypical | 130 | 204 | 172 | 0 | normal | No |

# Heart Data

These data contain a binary (or qualitative) outcome AHD for 303 patients who presented with chest pain. An outcome value of:

- ***Yes*** indicates the presence of heart disease based on an angiographic test,
- ***No*** means no heart disease.

There are 13 predictors including:

- Age
- Sex (0 for women, 1 for men)
- Chol (a cholesterol measurement),
- MaxHR
- RestBP

and other heart and lung function measurements.

# Classification

- Linear Regression performs well on tasks that require the prediction of a quantitative response variable.
    - For example, number of taxi pickups, number of bike rentals.
- When the response variable is **categorical (or** qualitative)**,** regression techniques are no longer applicable.
- Such tasks fall under the umbrella of a **classification problem**.
    - For example, whether the next taxi is going to be yellow or blue or black.
- The goal of a classification problem is to attempt to classify an observation into a category (aka, class or cluster) labeled by Y, based on a set of predictor variables X.



# Typical Classification Examples

Classification problems are ubiquitous in many domains, such as healthcare, finance, sports.

Some examples of classification problems are:

- To determine whether a startup is worth investing in
- To determine the disease type of patients based on various genomic markers
- To determine if a certain candidate is suitable for a particular sports team
- To determine if a given image is a real or a fake one



# Why not Linear Regression?

# Simple Classification Example

Given a dataset:

$$\boxed{\left\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\right\}}$$

where the $y$s are categorical, the aim is to predict which category $y$ takes

$$y = \begin{cases} 1 & if \ \text{Computer Science (CS)} \\ 2 & if \ \text{Statistics} \\ 3 & \text{otherwise} \end{cases}$$

# Simple Classification Example (cont.)

- A linear regression could be used to predict y from **x**.
- This model would imply though a specific ordering of the outcome, and would treat a one-unit change in y equivalently
  - For example;  a change from y=1 to y=2 (Computer Science to Statistics) is the considered the same as a change from y=2 to y=3 (Statistics to everyone else).
- However, this change should not be interpreted as the same.

# Simple Classification Example (cont.)

- Additionally, if the ordering of the response variable is changed, the model estimates and predictions would be fundamentally different.
  - For example, a model trained with $y = 1$ represents Statistics and $y = 2$ represents CS is different from a model trained with the original ordering.

- If a categorical response variable is ***ordinal*** (has a natural ordering, like Freshman, Sophomore, etc.), then a linear regression model would make some sense but is still not ideal.

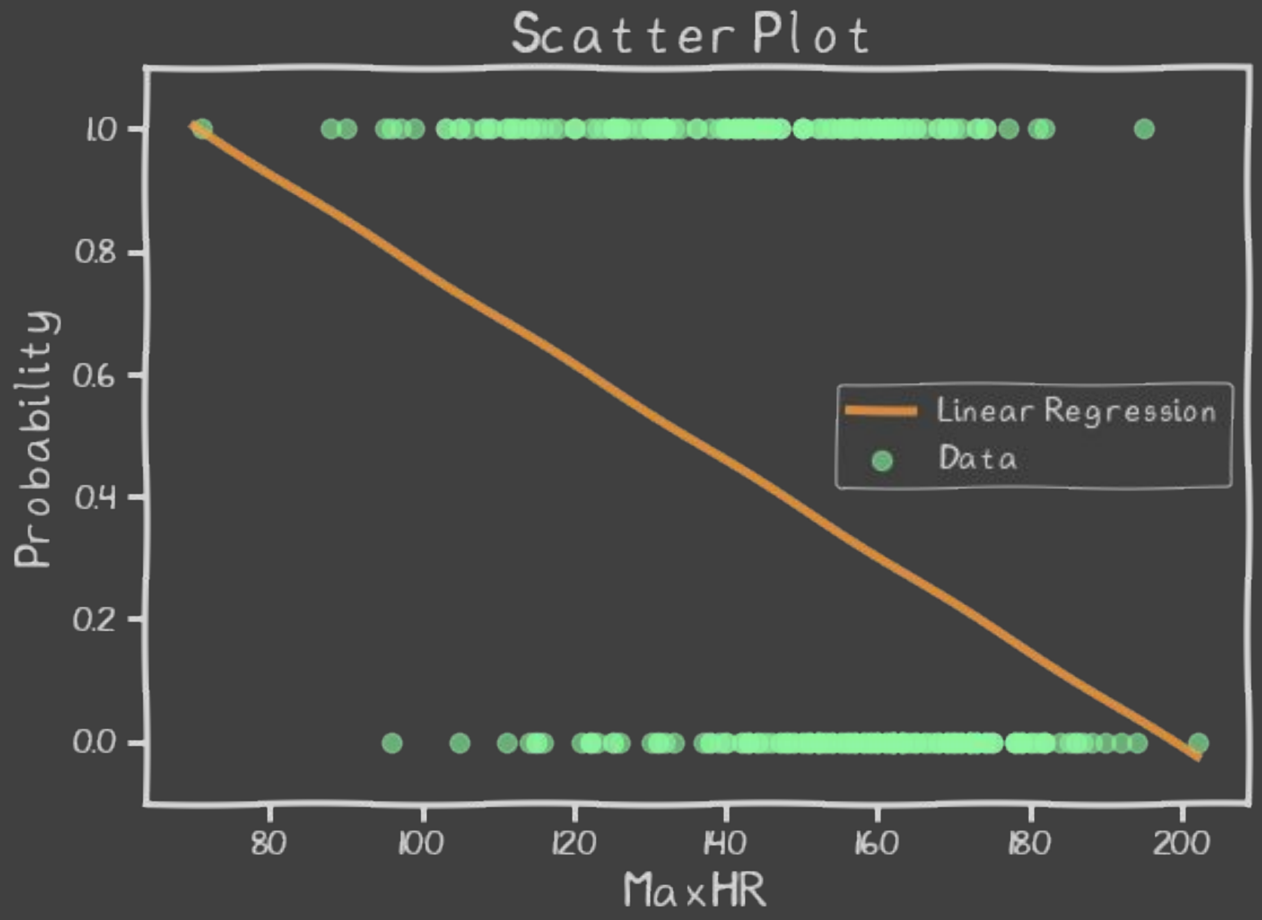
# Even Simpler Classification Problem: Binary Response

- The simplest form of classification is when the response variable $y$ has only two categories.
- There is a natural ordering of the categories.

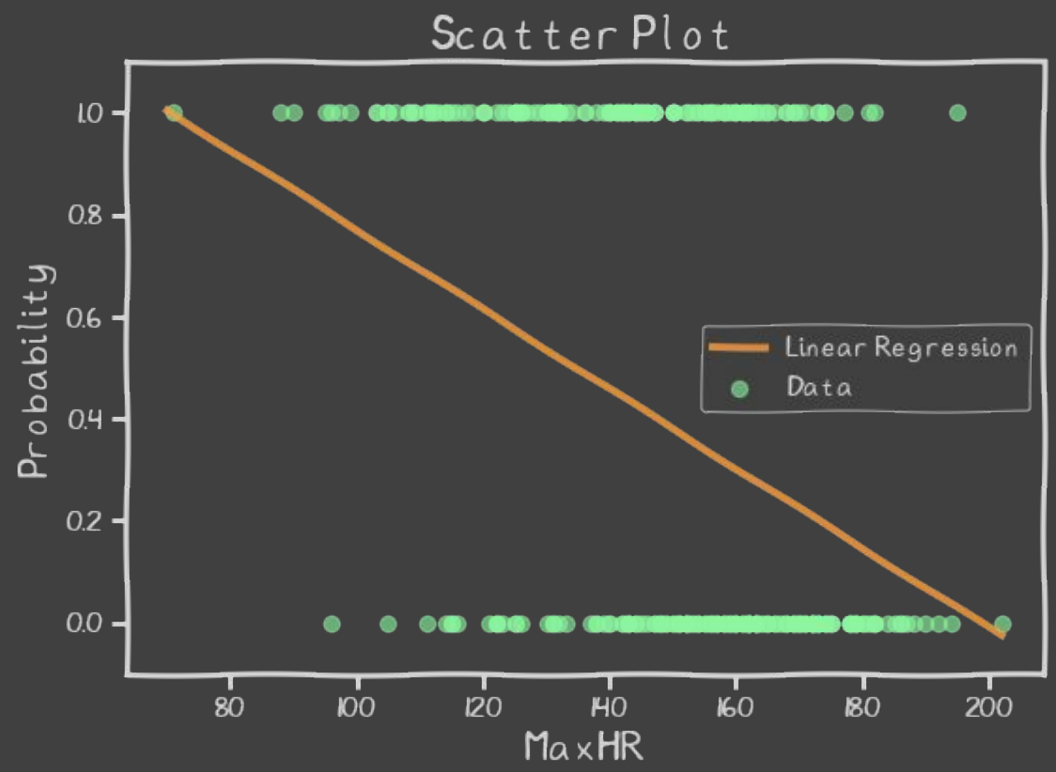
$$\boxed{y = \begin{cases} 1 & if\ \text{lives in the Quad} \\ 0 & \text{otherwise} \end{cases}}$$

# Even Simpler Classification Problem: Binary Response (cont.)

What could go wrong with this linear regression model?



# Even Simpler Classification Problem: Binary Response (cont.)



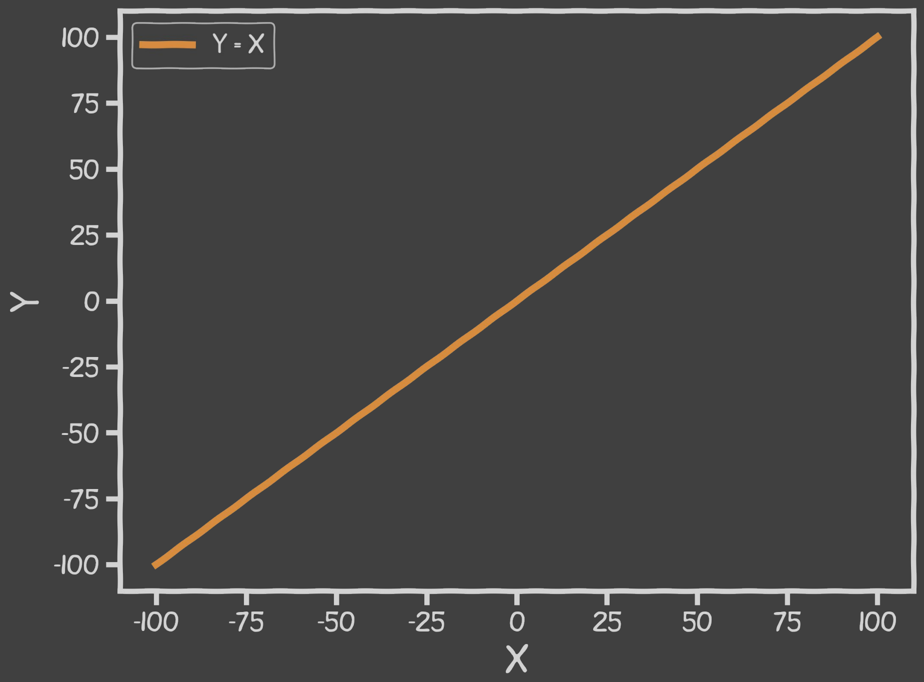

The main issue is you could get nonsensical values for $y$. Since this is modeling $P(y = 1)$ , values for $\hat{y}$ below 0 and above 1 would be at odds with the natural measure for prob. Linear regression can lead to this issue.

# Binary Response & Logistic Regression

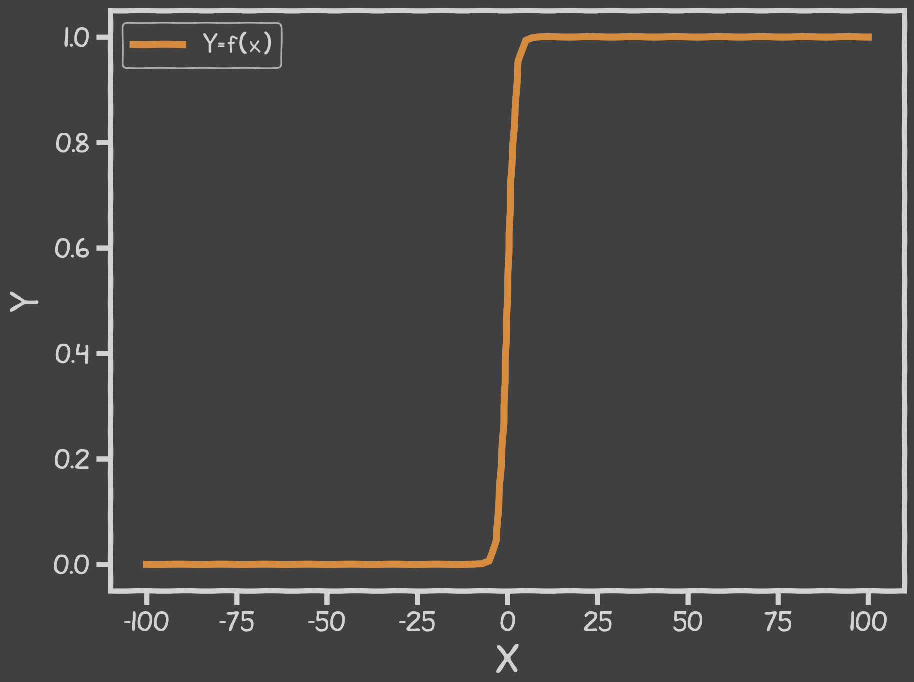
# Pavlos Game #45

Since we know, linear regression yields values for prob that are larger than 1 or smaller than zero what can we do to fix this?
We could transform the output of the linear regression to something that goes from 0 to 1.

Think of a function that would do this for us




$$Y' = f(y)$$



# Logistic Regression

- Logistic Regression addresses the problem of estimating a probability, $P(y = 1)$, to be outside the range of [0,1].

- In particular The logistic regression model uses a function, called the ***logistic*** function, to model $P(y = 1)$:

$$\boxed{P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}}$$

# Logistic Regression

As a result, the model will predict $P(y = 1)$ with an $S$-shaped curve, which is the general shape of the logistic function.

$\beta_0$ shifts the curve right or left by $\mathrm{c} = -\dfrac{\beta_0}{\beta_1}$.

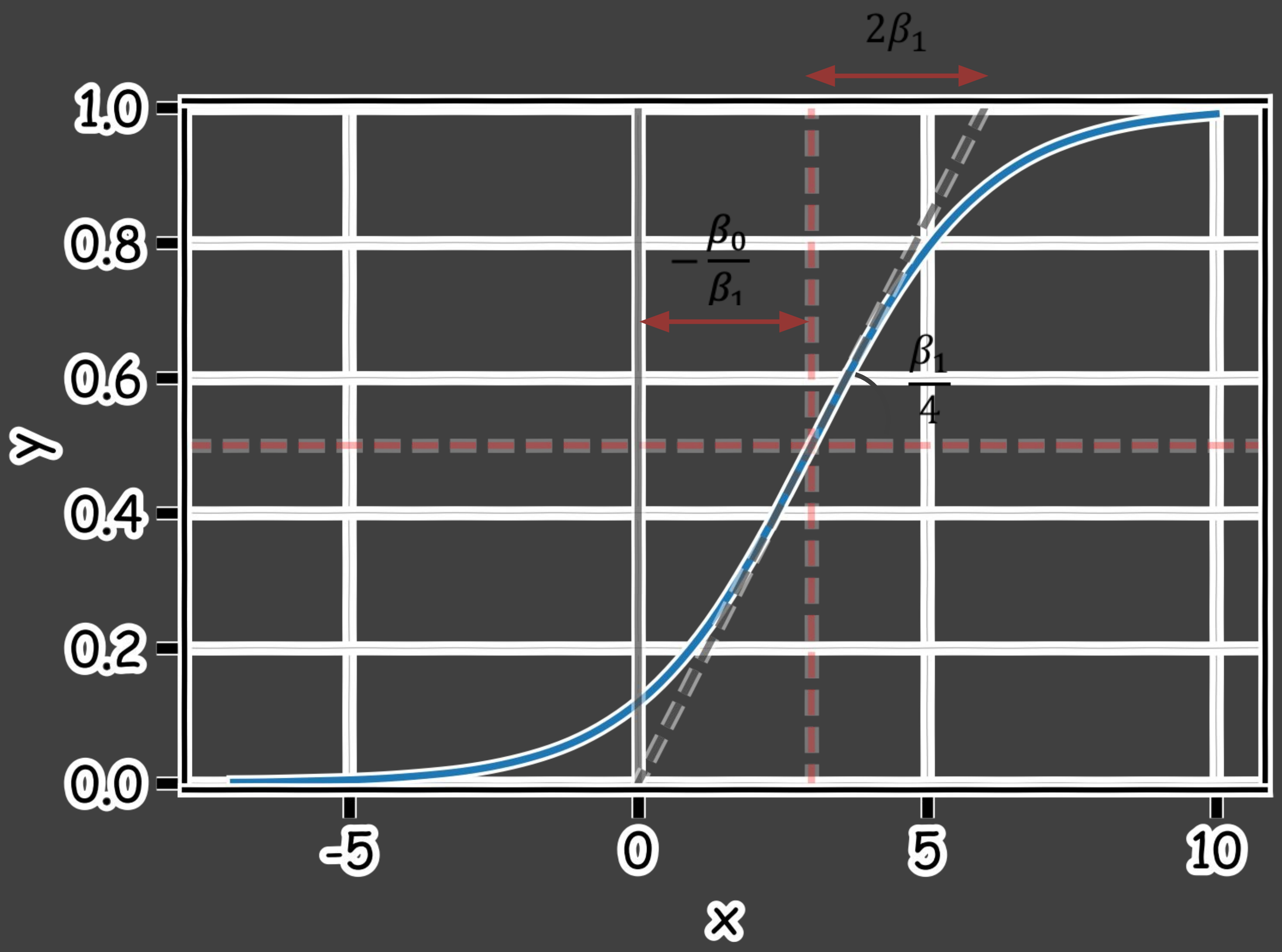
$\beta_1$ controls how steep the $S$-shaped curve is. Distance from ½ to almost 1 or ½ to almost 0 to ½ is $\dfrac{2}{\beta_1}$

Note: if $\beta_1$ is positive, then the predicted $P(y = 1)$ goes from zero for small values of $X$ to one for large values of $X$ and if $\beta_1$ is negative, then the $P(y = \ 1)$ has opposite association.

# Logistic Regression

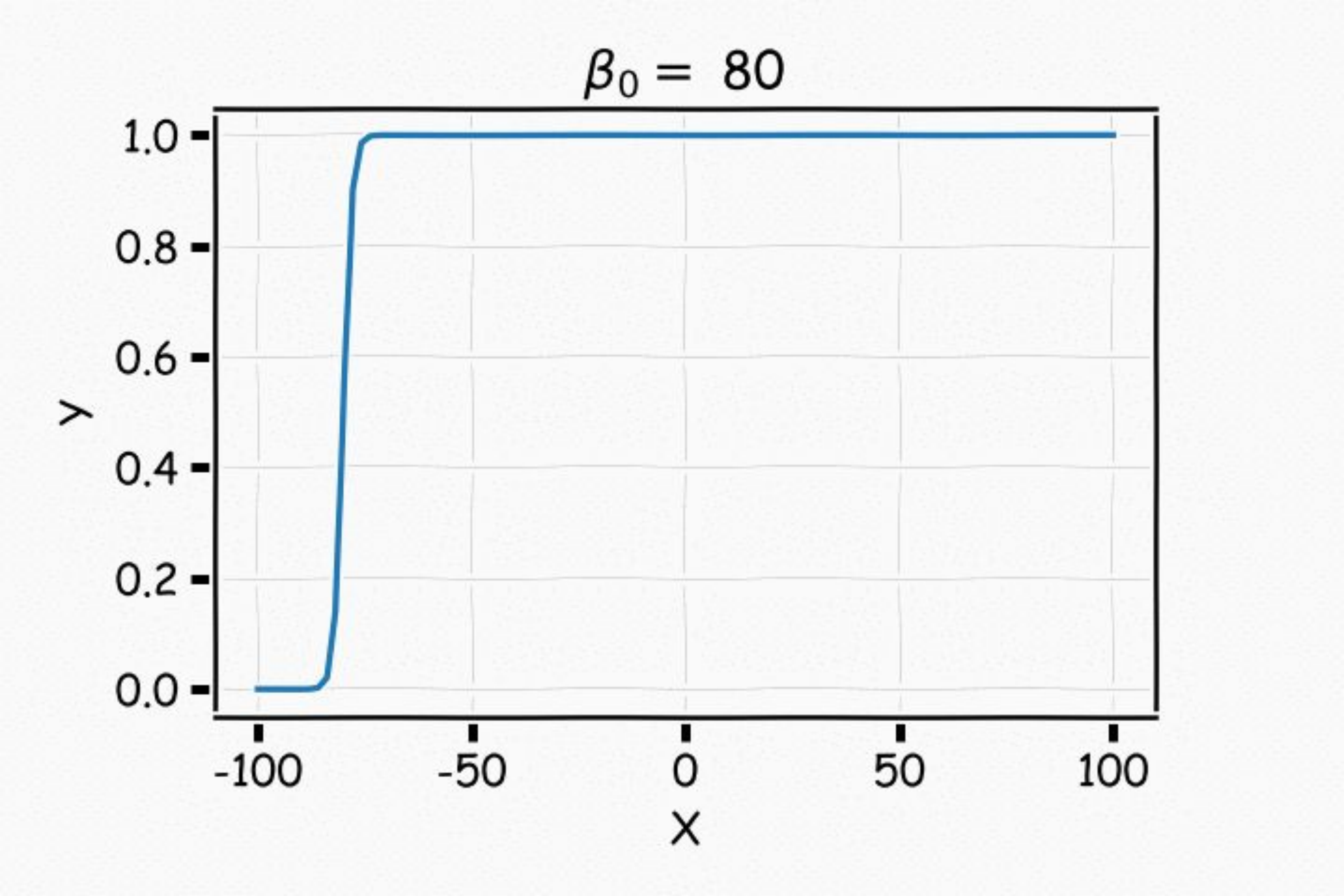
The coefficients beta0 and beta1 now control the shape of this s-shape curve

The point that P=1/2



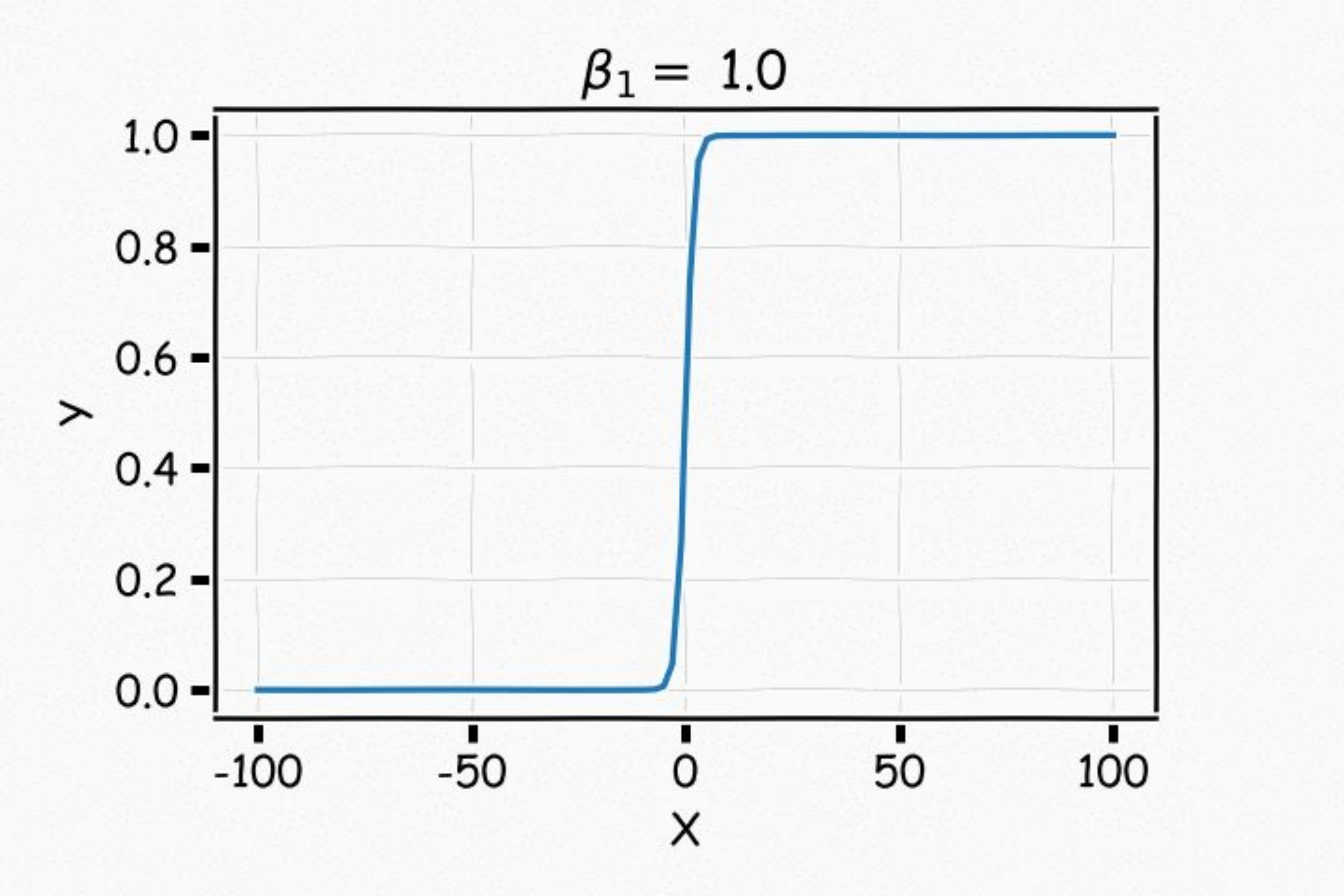
# Logistic Regression

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$



# Logistic Regression

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$



# Interpretation of $\beta$'s

With a little bit of algebraic work, the logistic model can be rewritten as:

$$\ln\left(\underbrace{\frac{P(Y=1)}{1-P(Y=1)}}_{\textit{odds}}\right) = \beta_0 + \beta_1 X$$

# Using Logistic Regression for Classification

**How can we use a logistic regression model to perform classification?**

That is, how can we predict when $Y = 1$ vs. when $Y = 0$?

We can classify all observations for which:

# Using Logistic Regression for Classification

When will this Bayes classifier be a good one? When will it be a poor one?

The Bayes classifier is the one that minimizes the overall classification error rate. That is, it minimizes:

$$\frac{1}{n}\sum_{i}^{n} I\left(y_i = \hat{y}_i\right)$$

Is this a good Loss function to minimize? Why or why not?

The Bayes classifier may be a poor indicator within a group. Think about the Heart Data scatter plot.

# Using Logistic Regression for Classification

This has potential to be a good classifier if the predicted probabilities are on both sides of 0 and 1.

How do we extend this classifier if $Y$ has more than two categories?