Classification with Logistic Regression

Pavlos Protopapas, Ignacio Becker

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

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	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
63	1	typical	145	233	1	2	150	0	2.3	3	0.0	fixed	No
67	1	asymptomatic	160	286	0	2	108	1	1.5	2	3.0	normal	Yes
67	1	asymptomatic	120	229	0	2	129	1	2.6	2	2.0	reversable	Yes
37	1	nonanginal	130	250	0	0	187	0	3.5	3	0.0	normal	No
41	0	nontypical	130	204	0	2	172	0	1.4	1	0.0	normal	No

Heart Data

These data contain a binary 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

Up to this point, the methods we have seen have centered around modeling and the prediction of a **quantitative** response variable (ex, number of taxi pickups, number of bike rentals, etc).

Linear **regression** (and Ridge, LASSO, etc) perform well under these situations

When the response variable is **categorical**, then the problem is no longer called a regression problem but is instead labeled as a **classification problem**.

The goal is to attempt to classify each observation into a category (aka, class or cluster) defined by Y, based on a set of predictor variables X.

Typical Classification Examples

The motivating examples for this lecture(s), are based [mostly] on medical data sets. Classification problems are common in this domain:

- Trying to determine where to set the *cut-off* for some diagnostic test (pregnancy tests, prostate or breast cancer screening tests, etc…)
- Trying to determine if cancer has gone into remission based on treatment and various other indicators
- Trying to classify patients into types or classes of disease based on various genomic markers

Why not Linear Regression?

Simple Classification Example

Given a dataset:

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$$

where the y are categorical (sometimes referred to as qualitative), we would like to be able to predict which category y takes on given x.

A categorical variable y could be encoded to be quantitative. For example, if y represents concentration of Harvard undergrads, then y could take on the values:

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

Simple Classification Example (cont.)

A linear regression could be used to predict y from $\boldsymbol{x}$.

The model would imply a specific ordering of the outcome, and would treat a one-unit change in y equivalent. The jump from $y = 1$ to y = 2 (CS to Statistics) should not be interpreted as the same as a jump from $y = 2$ to $y = 3$ (Statistics to everyone else).

Similarly, the response variable could be reordered such that $y = 1$ represents Statistics and $y = 2$ represents CS, and then the model estimates and predictions would be fundamentally different.

If the categorical response variable was ***ordinal*** (had a natural ordering, like class year, 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, and then an ordering of the categories is natural.

For example, an upperclassmen Harvard student could be categorized as:

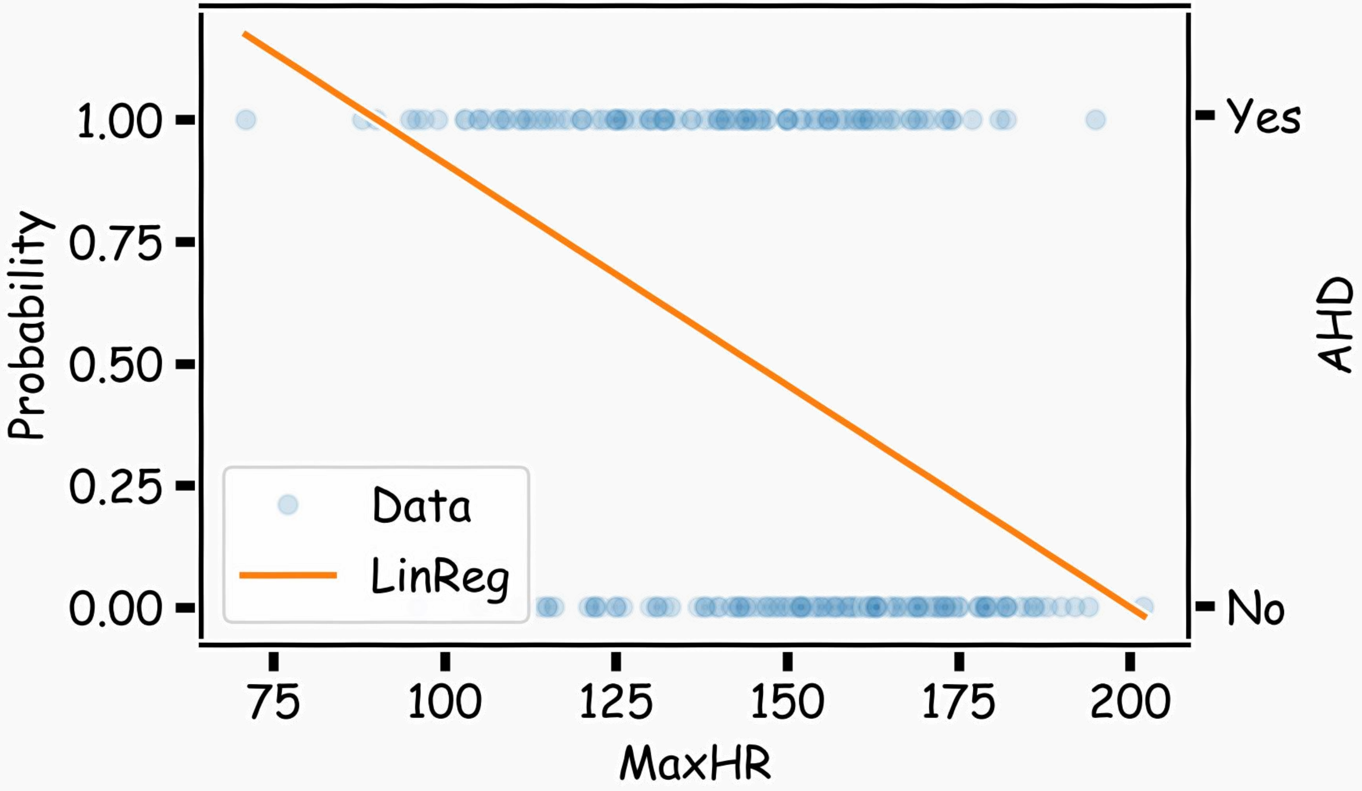
$$y = \begin{cases} 1 & if \text{ lives in the Quad} \\ 0 & \text{otherwise} \end{cases}.$$

Note: the $y = 0$ category is a "catch-all" so it would involve both River House students and those who live in other situations: off campus.

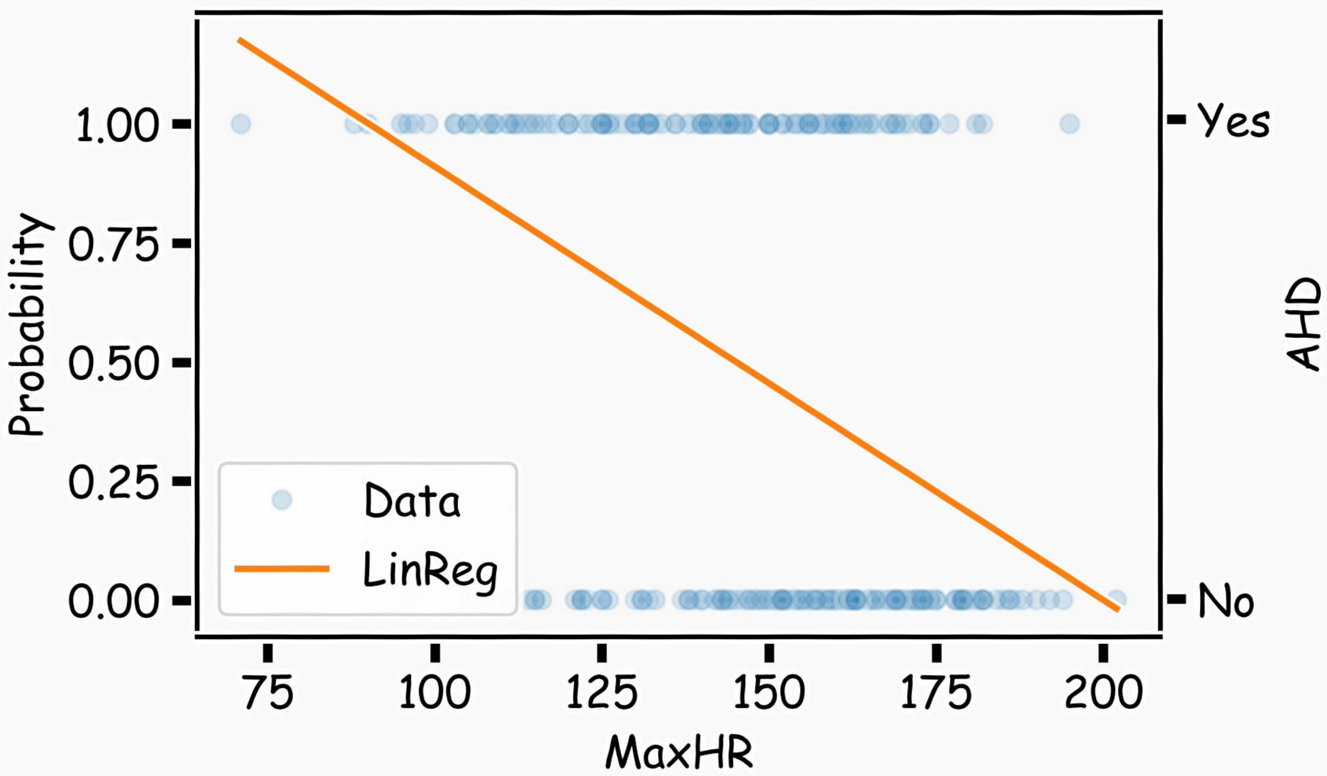
Linear regression could be used to predict the probability $\mathrm{P}(y = 1)$ directly from a set of covariates (like sex, whether an athlete or not, concentration, GPA, etc.), and if $\mathrm{P}(\mathrm{y} = 1) \geq 0.5$, we could predict the student lives in the Quad and predict other houses if $\mathrm{P}(\mathrm{y} = 1) < 0.5$.

Even Simpler Classification Problem: Binary Response (cont)

What could go wrong with this linear regression model?



Even Simpler Classification Problem: Binary Response (cont)

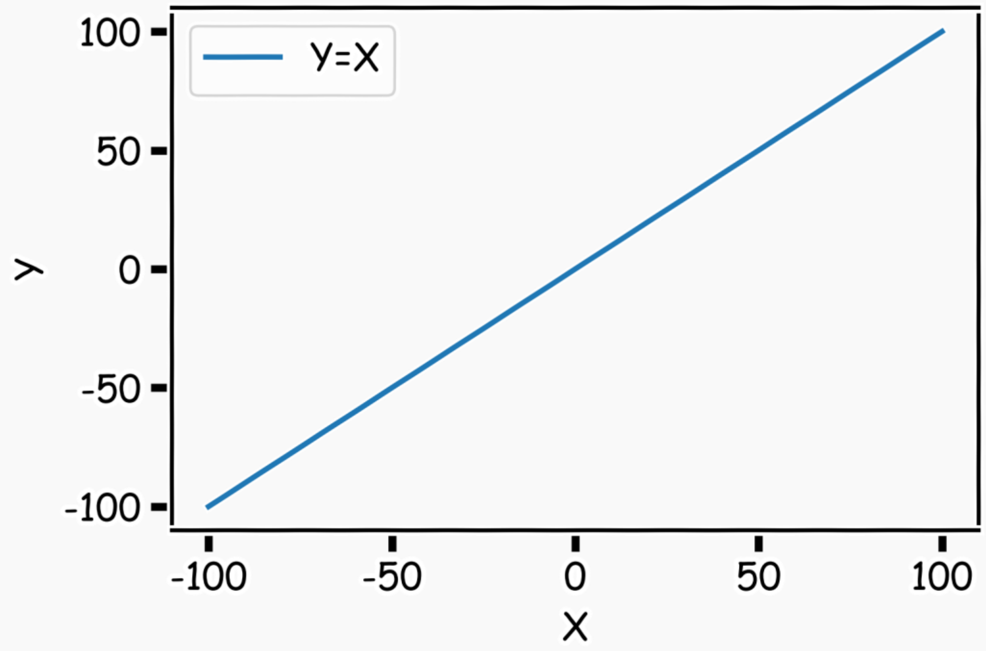


The main issue is you could get nonsensical values for y. Since this is modeling $P(y = 1)$, values for $\hat{y}$ below 0 and above 1 would be at odds with the natural measure for y.

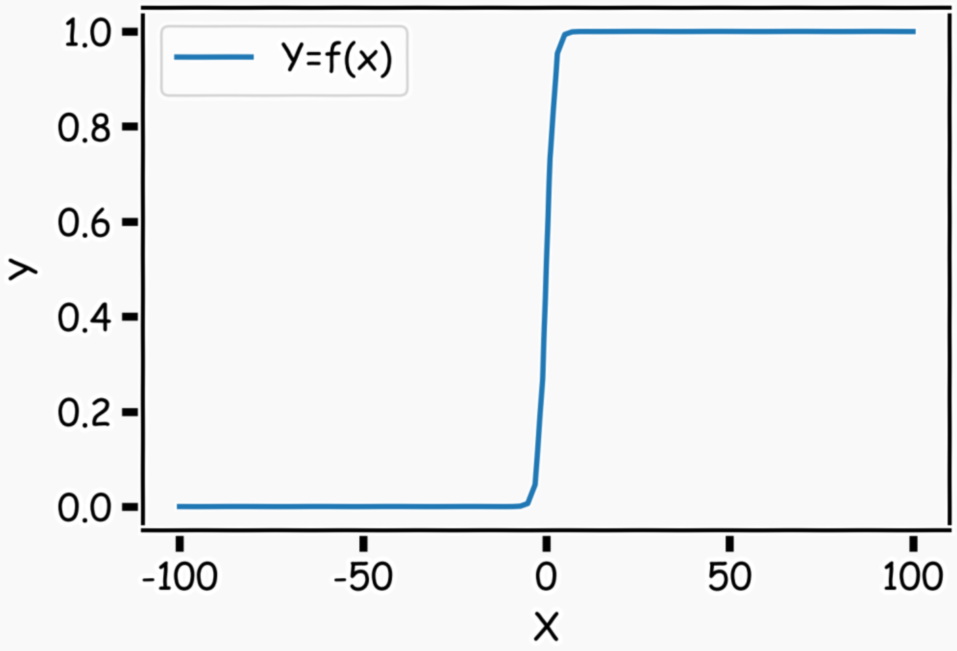
Binary Response & Logistic Regression

Ignacio’s Game #45

Think of a function that would do this for us



$$Y = f(x)$$



Logistic Regression

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

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

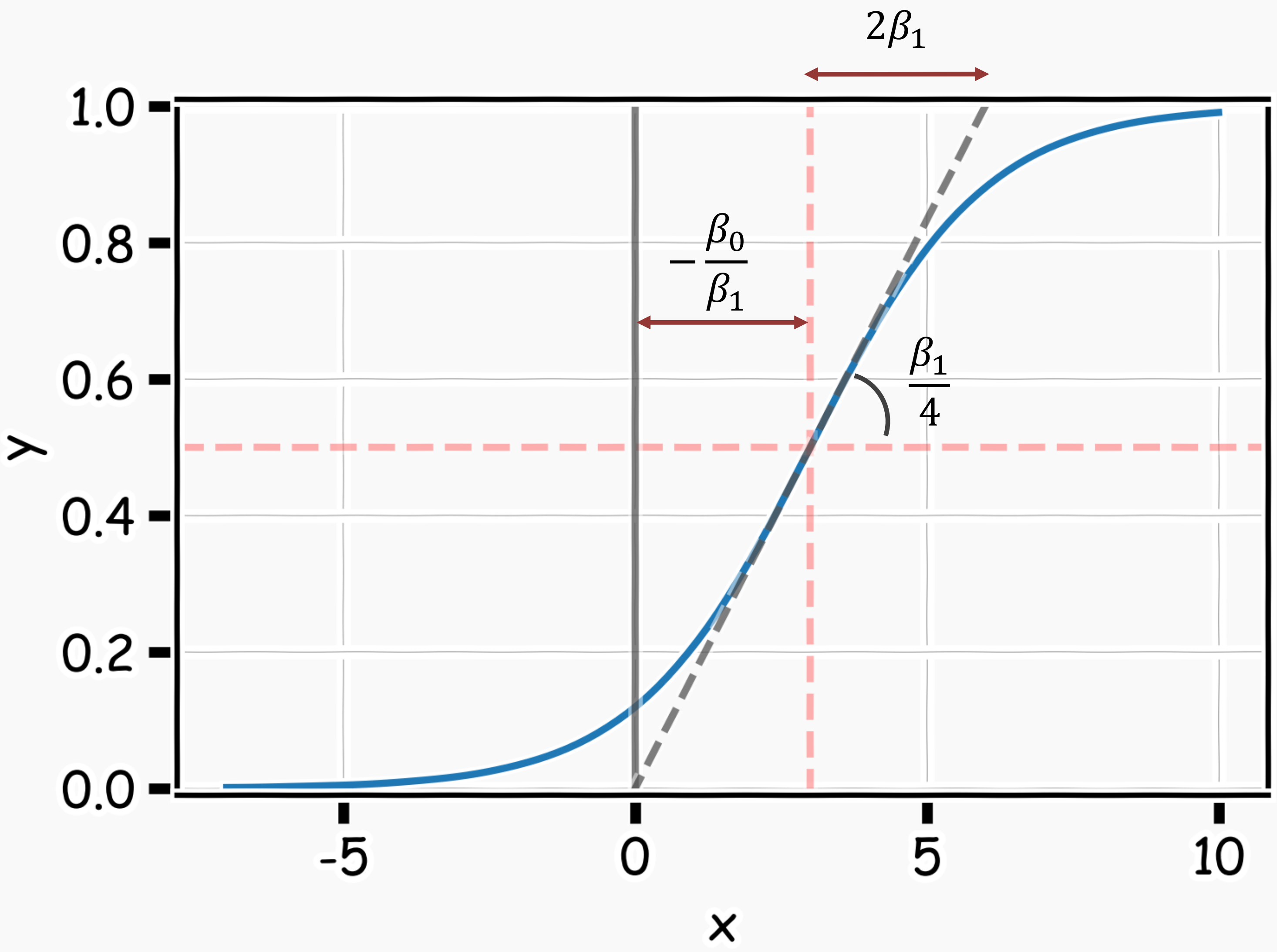
As a result the model will predict $P(y=1)$ with an S-shaped curve, which is the general shape of the logistic function.

β_0 shifts the curve right or left by $\mathrm{c} = -\frac{\beta_0}{\beta_1}$.

β_1 controls how steep the S-shaped curve is. Distance from ½ to almost 1 or ½ to almost 0 to ½ is $\frac{2}{\beta_1}$

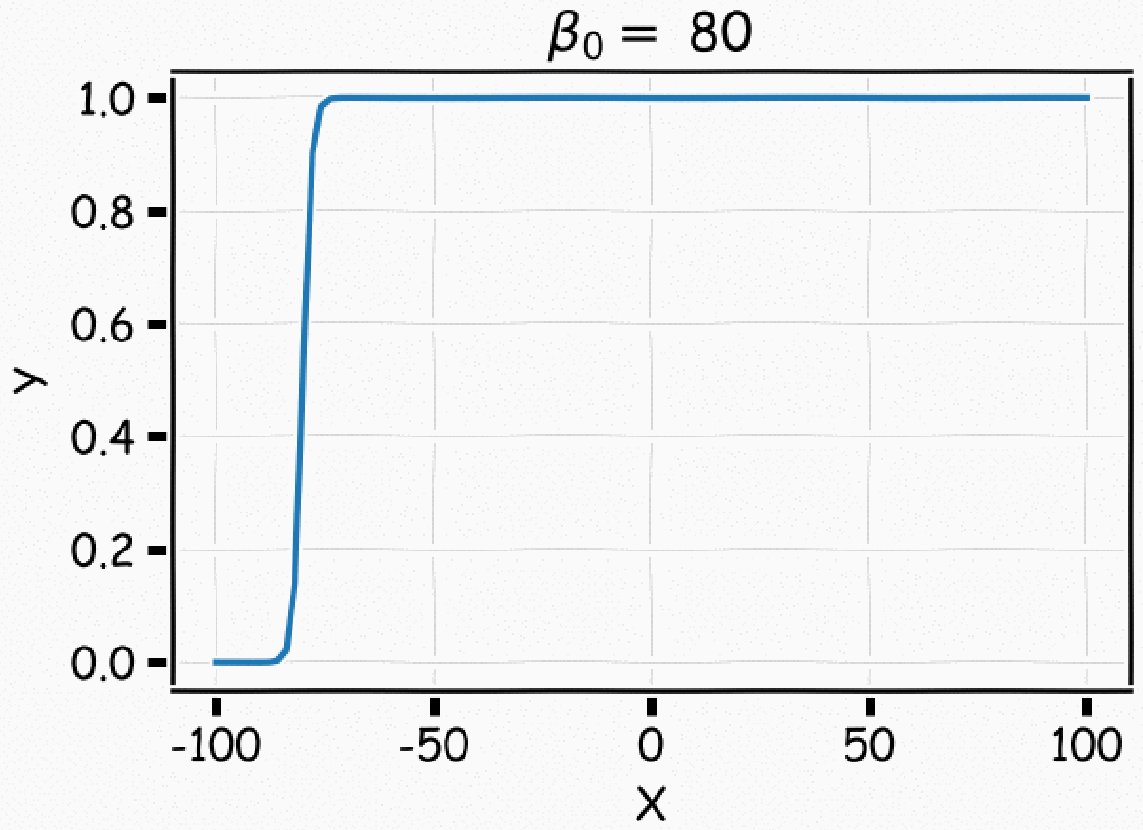
Note: if β_1 is positive, then the predicted $P(y=1)$ goes from zero for small values of X to one for large values of X and if β_1 is negative, then the $P(y = \ 1)$ has opposite association.

Logistic Regression



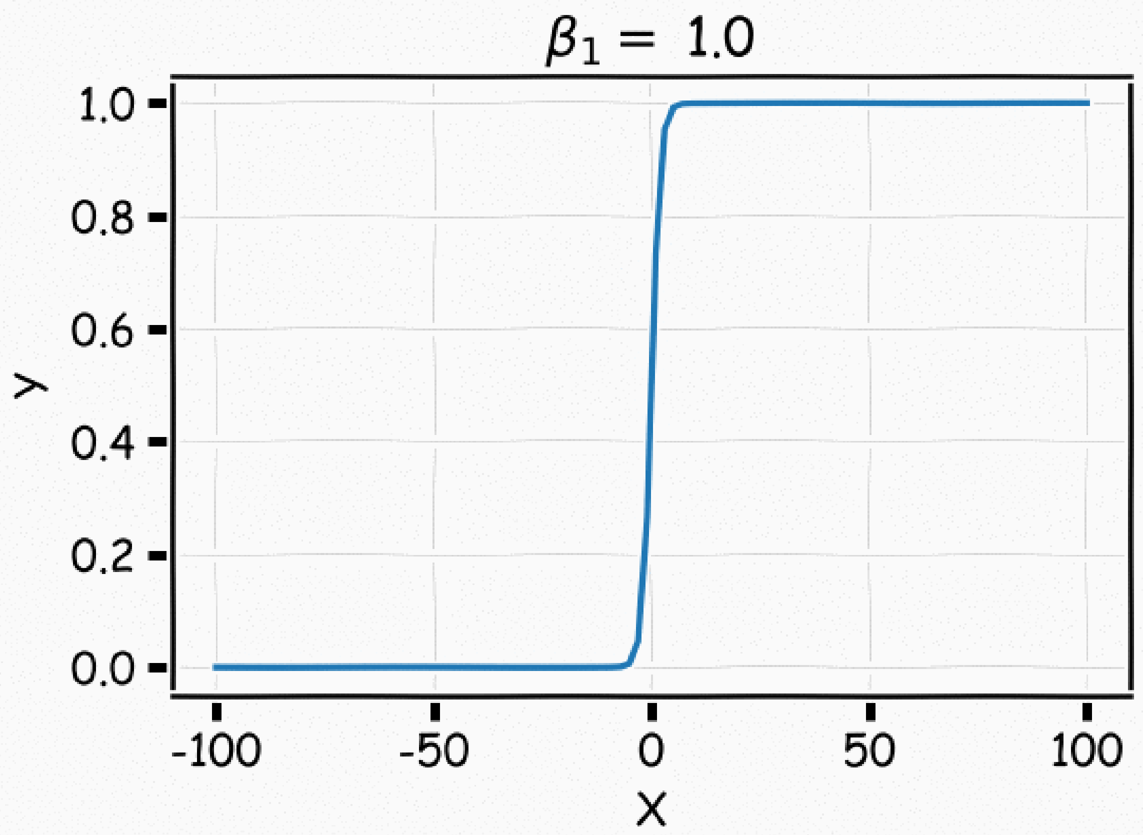
Logistic Regression

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



Logistic Regression

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



Interpretation of β's

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

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

logistic regression is said to model the ***log-odds*** with a linear function of the predictors or features, X.

Natural interpretation: a one unit change in X is associated with a β_1 change in the log-odds of $\mathrm{P}(Y=1)$; or better yet, a one unit change in X is associated with an e^{β_1} change in the odds that $Y = 1$.

Using Logistic Regression for Classification

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

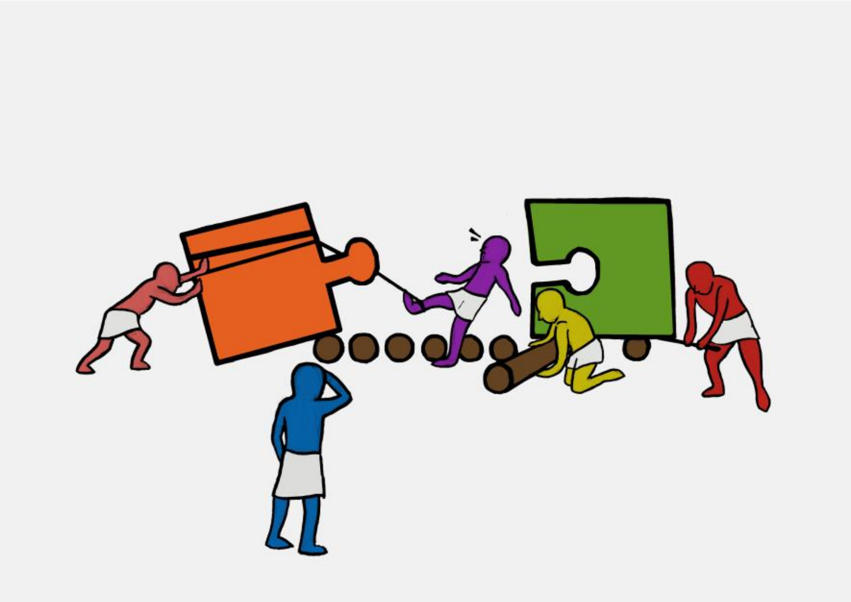
That is, how can we predict when $Y = 1$ vs. when $Y = 0$?

We can classify all observations for which:

$\hat{P}(Y = 1) \geq 0.5$ to be in the group associated with $Y = 1$

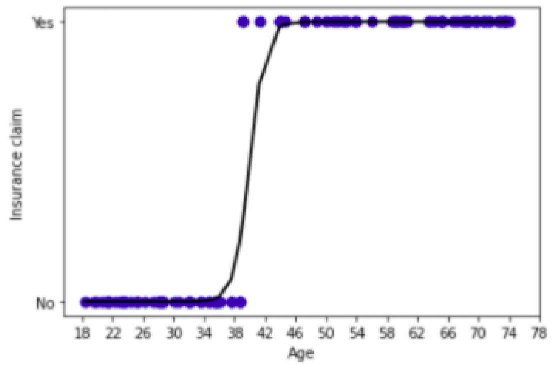
and then classify all observations for which

$\hat{P}(Y = 0) < 0.5$ to be in the group associated with $Y = 0$



Exercise: A.1 - Guesstimating Beta values for Logistic Regression

The goal of the exercise is to produce a plot similar to the one given below, by guesstimating the values of the coefficients β0 and β1.



Instructions:

We are trying to predict who will claim insurance as a function of age using the data. To do so we need :

- Read the `insurance_claim.csv` as a dataframe.
- Assign the predictor and response variables.
- Guesstimate the values of the coefficients β0 and β1.
- Predict the response variable using the formula of a simple logistic regression given below (no package allowed)
- Compute the accuracy of the model.
- Repeat the above steps by changing the values of the coefficients β0 and β1, until you get "good" accuracy.
- Plot the `Age` vs `Insurance Claim` graph with the fit of the model.