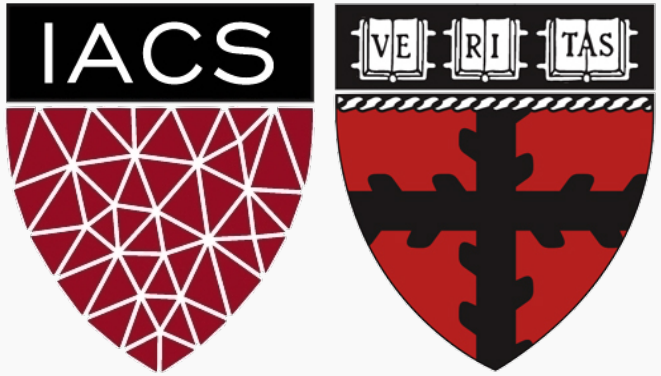
Lecture 10: Classification and Logistic Regression

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



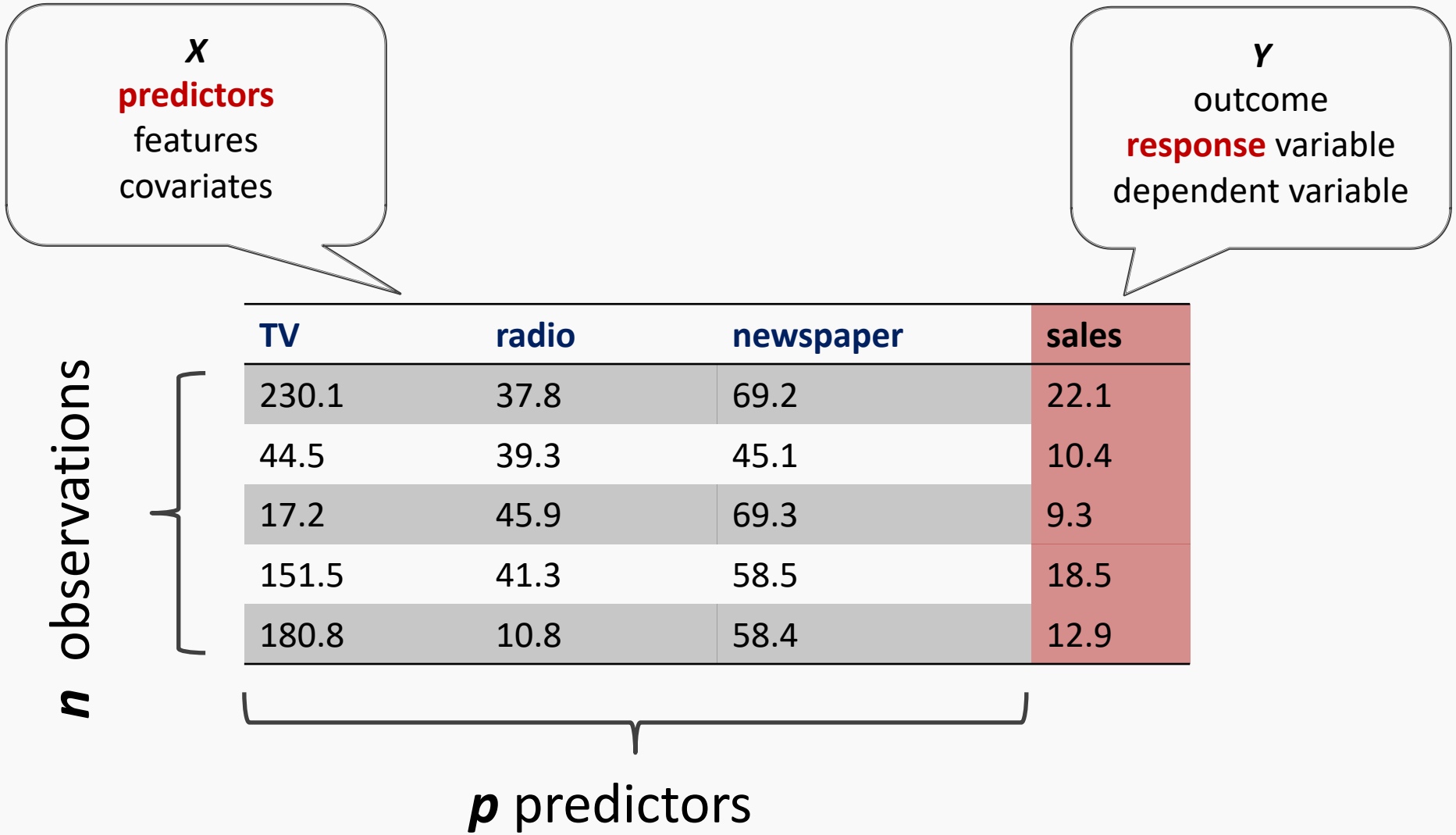
Announcements

- Project assignments coming out Wednesday. Email helpline TODAY if you haven't submitted preferences.
- HW2: grades coming tonight.
- HW3: due Wed @ 11:59pm.
- HW4: individual assignment. No working with other students. Feel free to use Ed, OHs, and Google like normal.

Lecture Outline

- 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)



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

Heart Data

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 HD 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

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), homework, and coming labs 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:

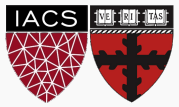
$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$$

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:

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

Linear regression **does not work well**, or is not appropriate at all, in this setting.



Simple Classification Example (cont.)

A linear regression could be used to predict y from $\boldsymbol{x}$. What would be wrong with such a model?

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 (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, etc):

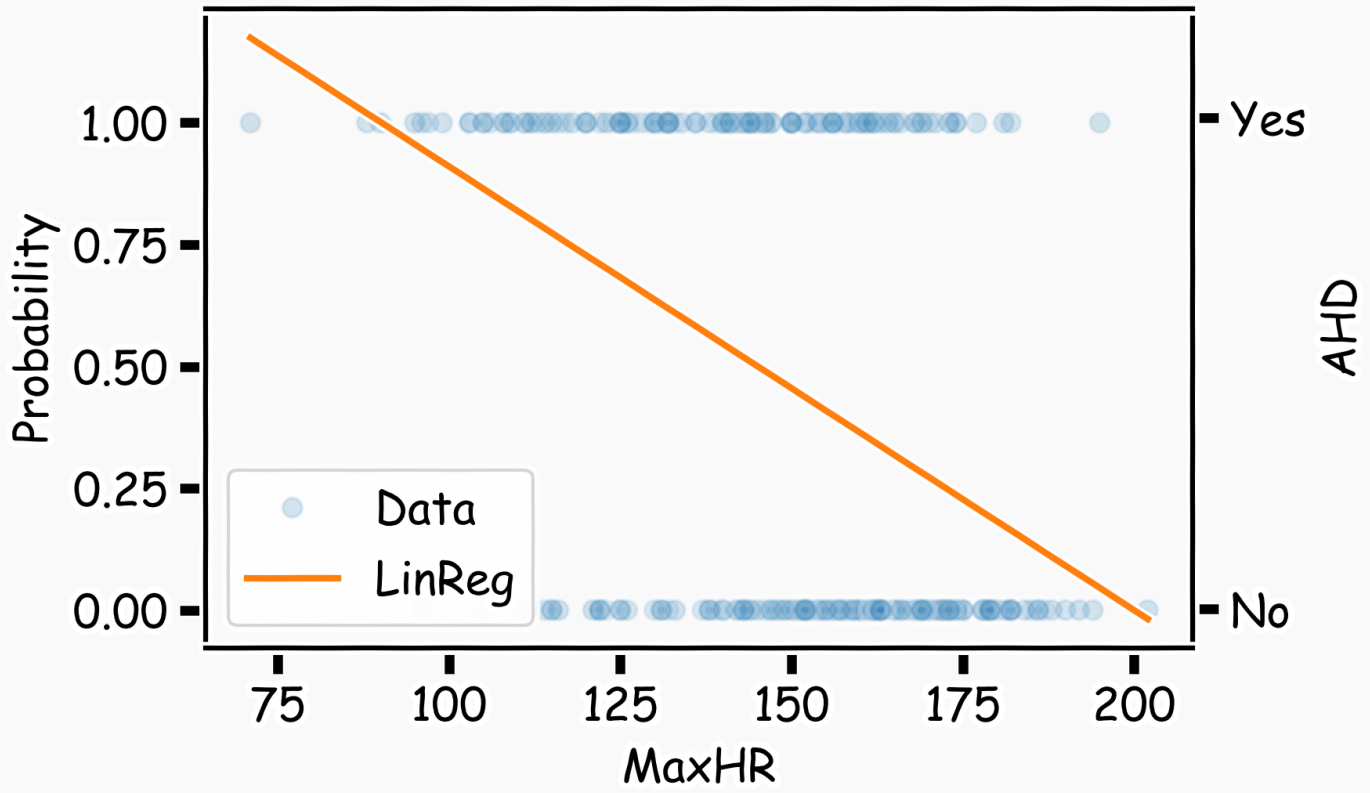
$$y = \begin{cases} 1 & if \text{ lives in the Quad} \\ 0 & \text{otherwise} \end{cases}.$$

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

Even Simpler Classification Problem: Binary Response (cont)

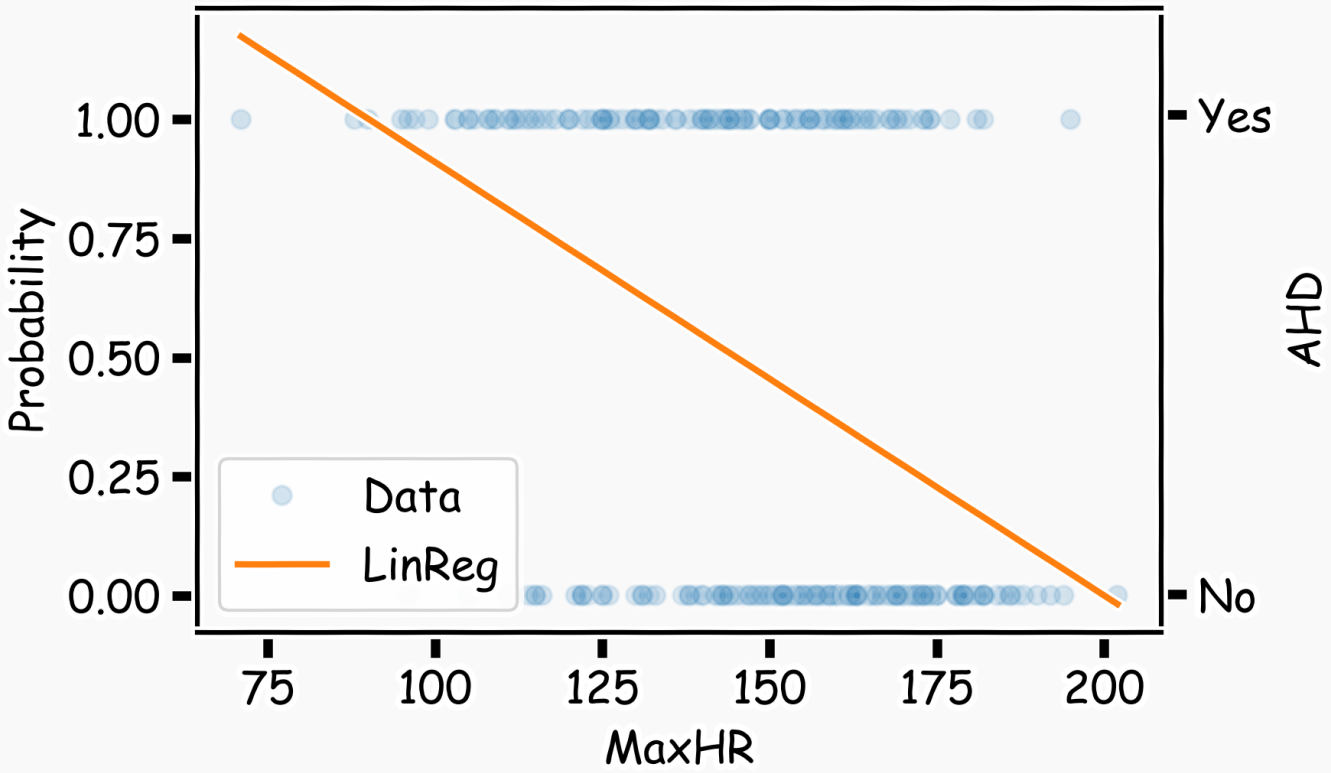
What could go wrong with this linear regression model?

.



Even Simpler Classification Problem: Binary Response (cont)

how do you interpret y-values that are between 0 and 1? Probability of having a heart attack/heart disease maybe!

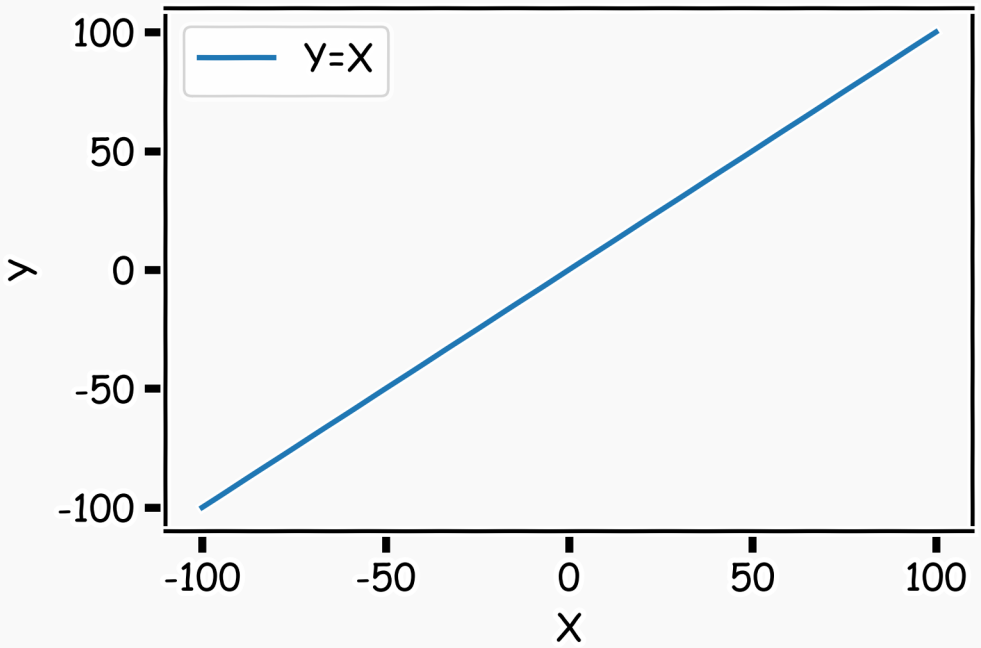


The main issue is you could get non-sensical values for y. Since this is modeling $P(y = 1)$, values for $\hat{y}$ below 0 and above 1 would be at odds with the natural measure for y. Linear regression can lead to this issue.

Binary Response & Logistic Regression

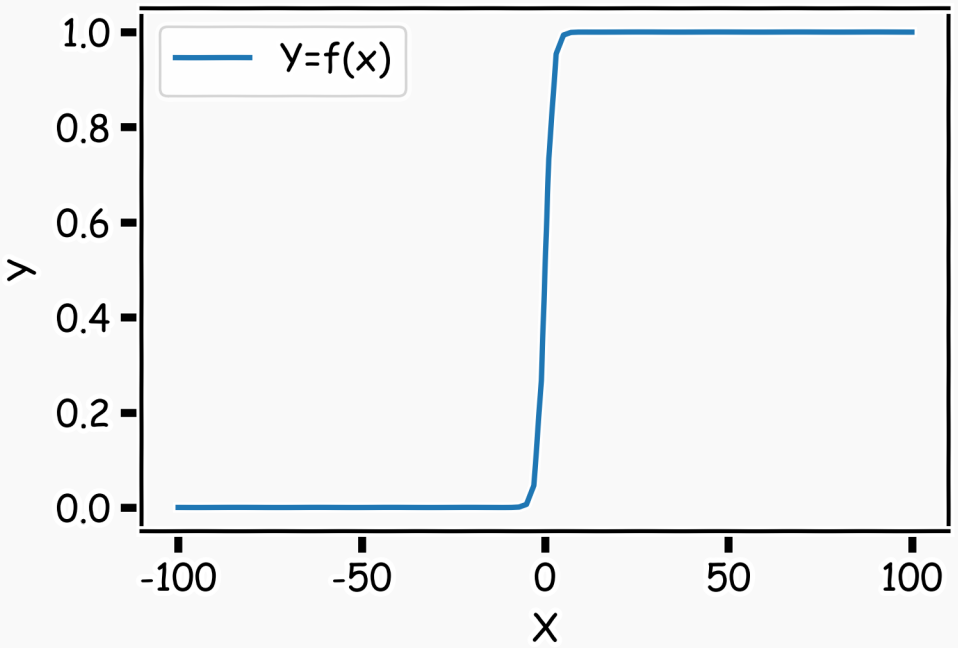
Pavlos Game #45

Think of a function that would do this for us



$$Y = f(x)$$

predictors can take on any value, but response is bounded between 0-1



These functions are CDFs!

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 $P(y = 1)$:

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

the 'junk' is the linear relationship
B1 is the steepness of the curve
B1 positive - upward slope (higher prob)
B1 negative - negative slope (lower prob)

Logistic Regression

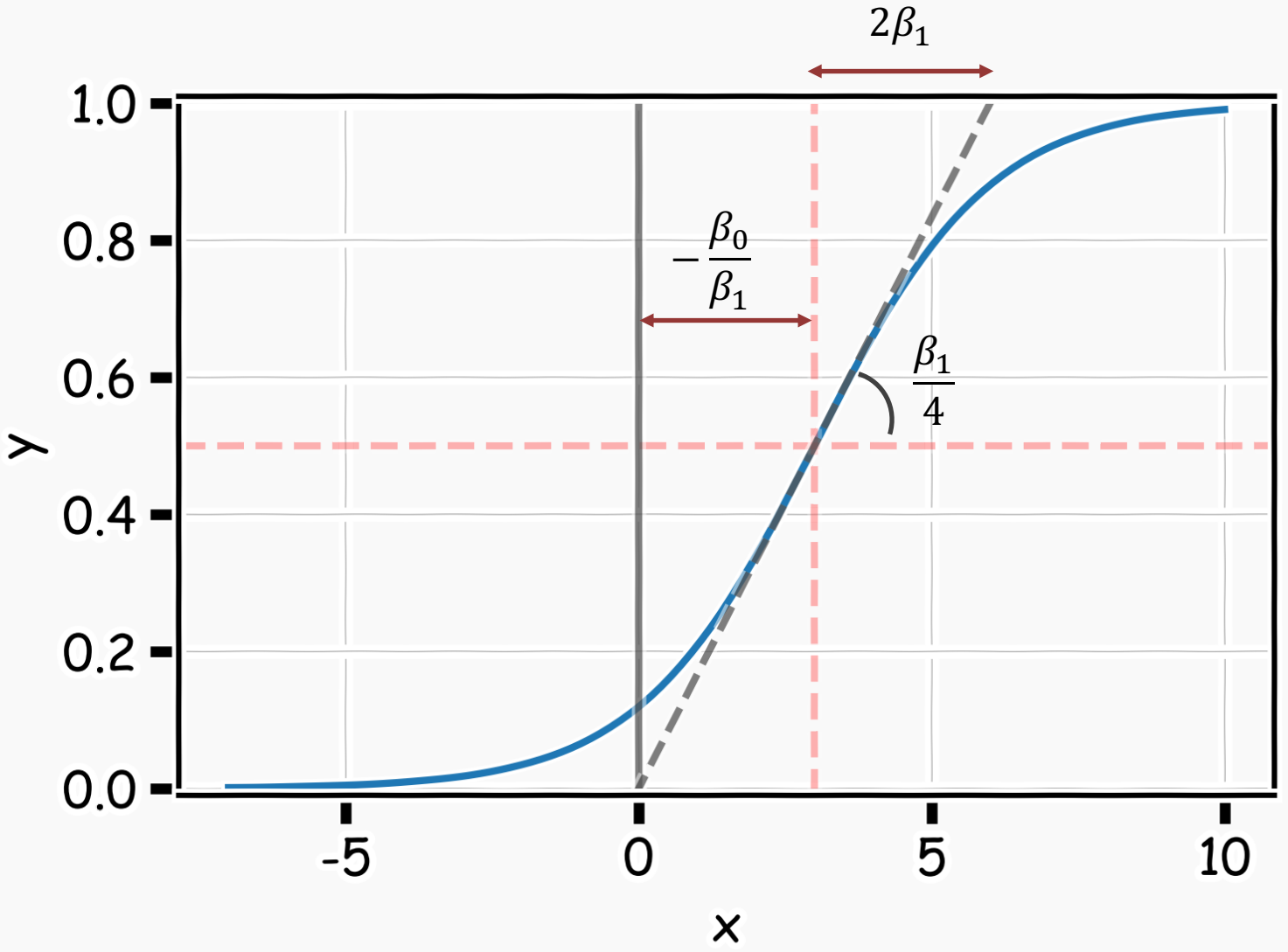
As a result the model will predict $P(y = 1)$ with an S-shaped curve, which is the general shape of the logistic function.

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

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

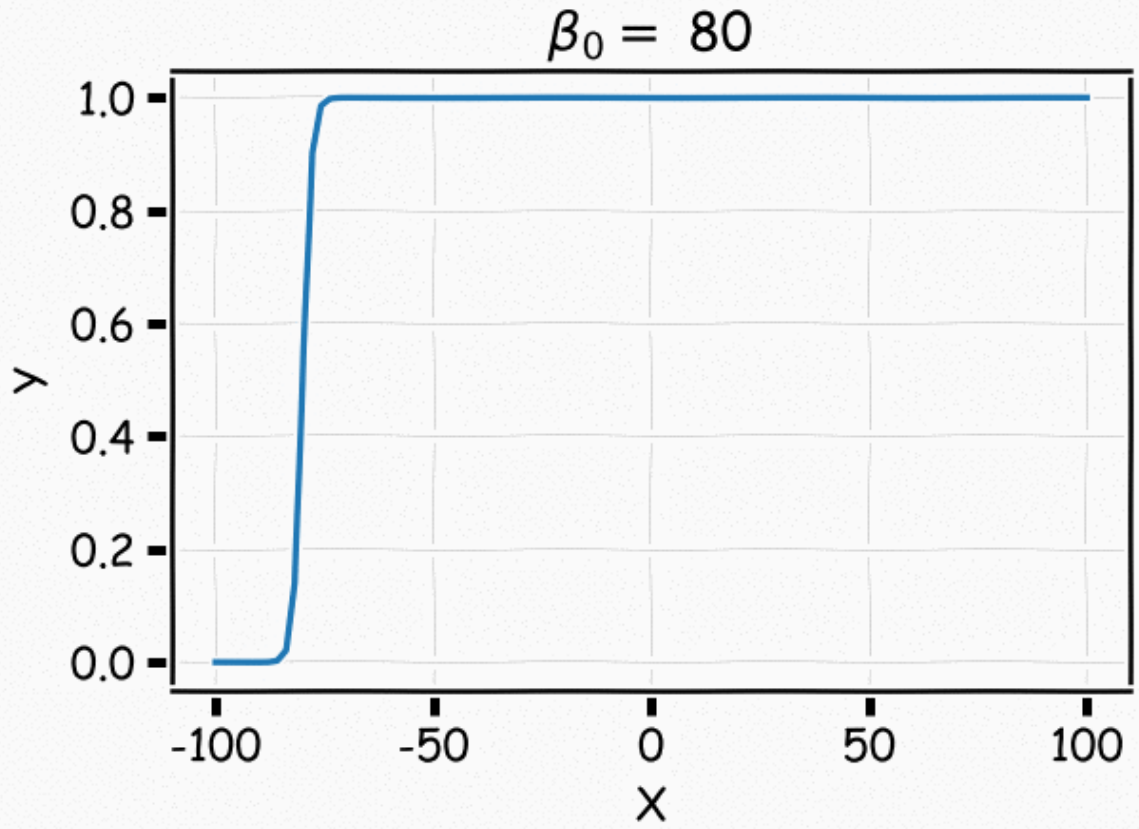
Note: if β_1 is positive, then the predicted $P(y = 1)$ goes from zero for small values of X to one for large values of X and if β_1 is negative, then the $P(y = 1)$ has opposite association.

Logistic Regression



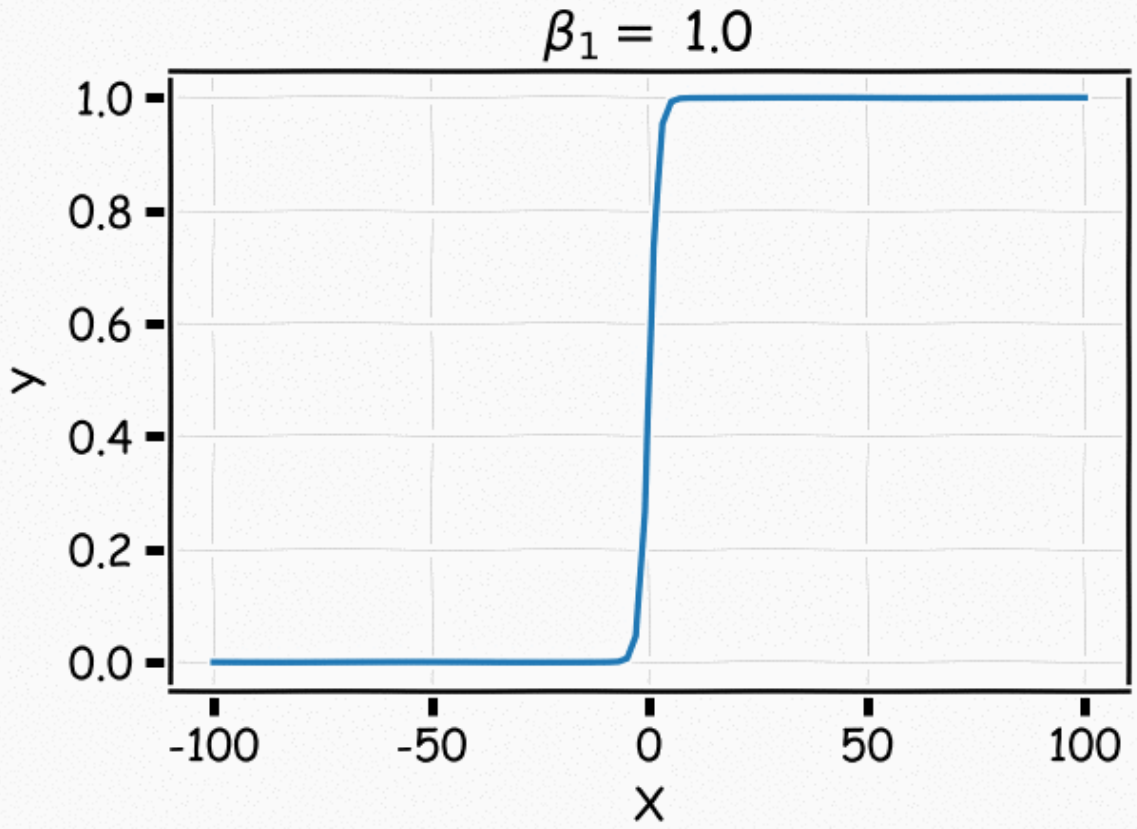
Logistic Regression

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



Logistic Regression

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



Logistic Regression

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

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

probability of success over the probability of failure = odds

The value inside the natural log function $\frac{P(Y=1)}{1-P(Y=1)}$, is called the ***odds***, thus logistic regression is said to model the ***log-odds*** with a linear function of the predictors or features, X. This gives us the natural interpretation of the estimates similar to linear regression: a one unit change in X is associated with a β_1 change in the log-odds of $Y = 1$; or better yet, a one unit change in X is associated with an e^{β_1} change in the odds that $Y = 1$.

Estimating the Simple Logistic Model

Estimation in Logistic Regression

Unlike in linear regression where there exists a closed-form solution to finding the estimates, $\hat{\beta}_j$'s, for the true parameters, logistic regression estimates cannot be calculated through simple matrix multiplication.

Questions:

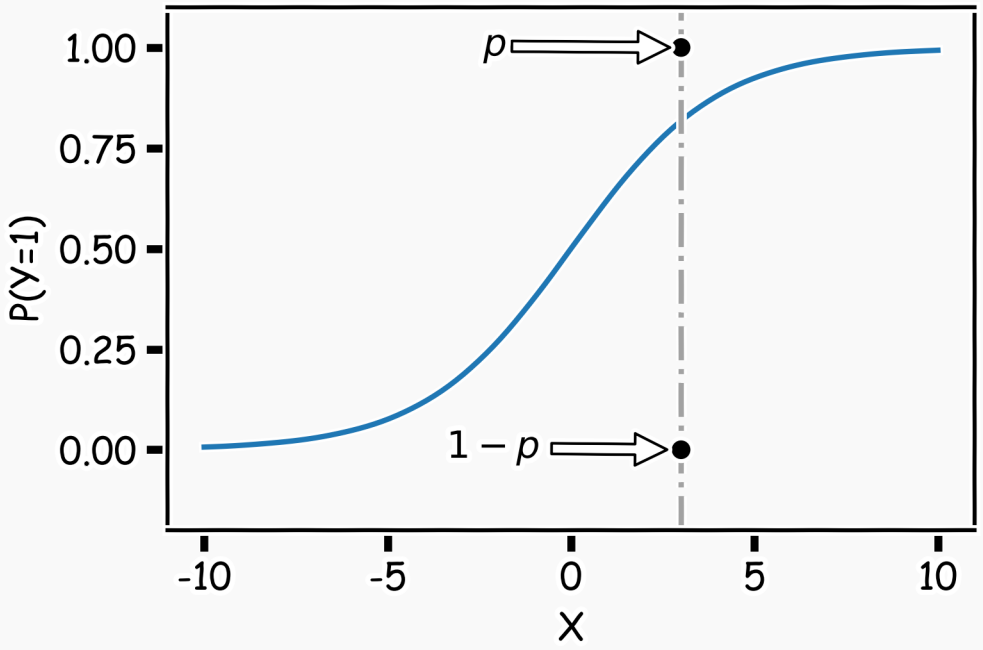
- In linear regression what loss function was used to determine the parameter estimates? MSE

- What was the probabilistic perspective on linear regression?

Normally distributed residuals - we can assume normal distribution of our response. Maximizing joint probability of observations is the same as minimizing MSE

- Logistic Regression also has a likelihood based approach to estimating parameter coefficients.

Estimation in Logistic Regression



Probability $Y = 1$: p

Probability $Y = 0$: $1 - p$

Bernoulli dist

$$P(Y = y) = p^y (1 - p)^{(1-y)}$$

where:

$p = P(Y = 1|X = x)$ and therefore p depends on X.

Thus not every p is the same for each individual measurement.

Likelihood

The likelihood of a single observation for *p* given *x and y* is:

$$L(p_i|Y_i) = P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

subscripts - not every y and p are the same

Given the observations are independent, what is the likelihood function for *p*?

$$L(p|Y) = \prod_i P(Y_i = y_i) = \prod_i p_i^{y_i}(1-p_i)^{1-y_i}$$

we should choose p that maximizes the likelihood function

$$l(p|Y) = -\log L(p|Y) = -\sum_i y_i \log p_i + (1-y_i)\log(1-p_i)$$

take log of product to make things easier to deal with

Loss Function

p is a function of B (B0,B1), X based on the logistic function relationship

$$l(p|Y) = -\sum_i \left[y_i \log \frac{1}{1+e^{-\beta X_i}} + (1-y_i) \log \left(1 - \frac{1}{1+e^{-\beta X_i}}\right) \right]_i$$

How do we minimize this? negative of log likelihood - we need to minimize now

Differentiate, equate to zero and solve for it!

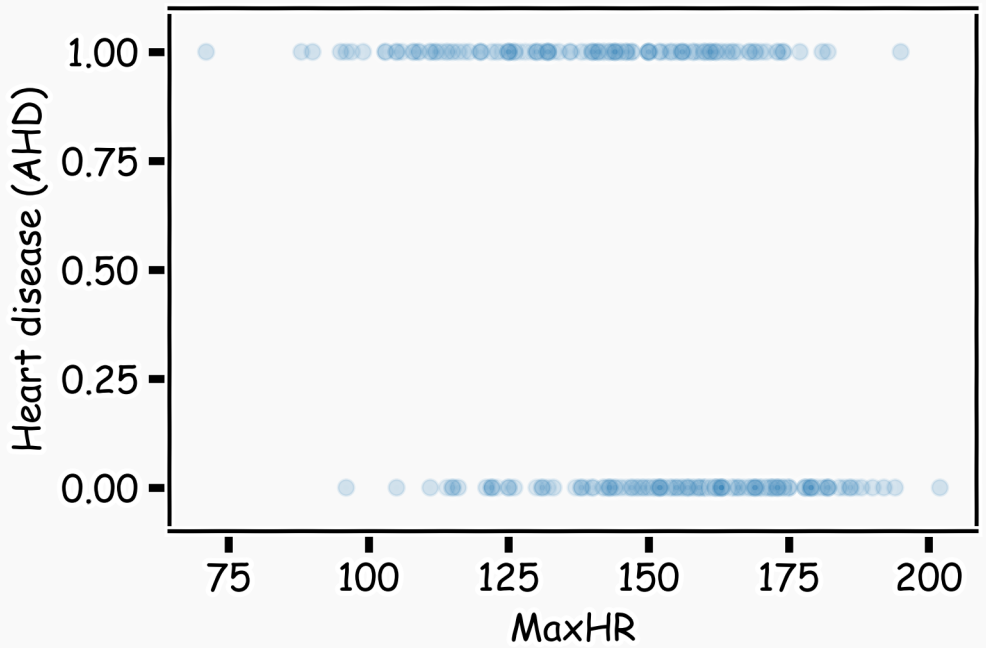
But jeeze does this look messy?! It will not necessarily have a closed form solution.

So how do we determine the parameter estimates? Through an iterative approach (we will talk about this *at length* in future lectures). This is a convex function so we good

Heart Data: logistic estimation

We'd like to predict whether or not a person has a heart disease. And we'd like to make this prediction, for now, just based on the MaxHR.

How should we visualize these data?



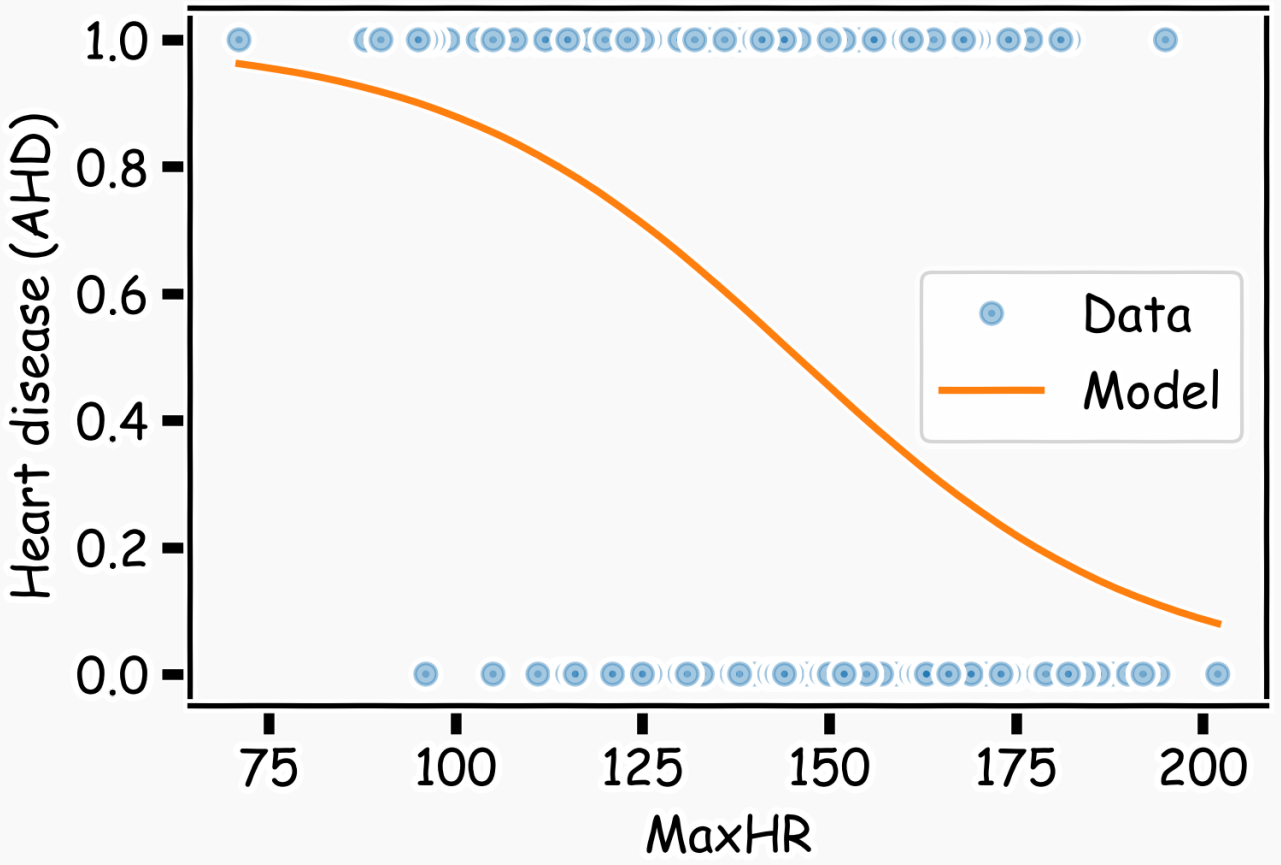
Higher MaxHR, lower risk of AHD

Heart Data: logistic estimation

To classify:
below 0.5, you are a 0
above 0.5, you are a 1

You don't have to use 0.5 for your cutoff. If you choose 0.1 as your cutoff, your misclassification rate goes up. BUT you might not treat misclassification one way or the other equally

e.g. classifiying a heart attack when not vs missing a heart attack when its happening



Heart Data: logistic estimation

There are various ways to fit a logistic model to this data set in Python. The most straightforward in `sklearn` is via `linear_model.LogisticRegression`.

```
from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression(C=100000, fit_intercept=True)
logreg.fit(data_x.values.reshape(-1,1), data_y);


print('Estimated beta1: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

C is the inverse of a penalty term - by default, logistic reg is regularized, so if you want to avoid it, you need to set a really large number

Text
Text
Text

```
Estimated beta1:
 [[-0.04326016]]
Estimated beta0:
 [ 6.30193148]
```

When you don't have a heart rate,
your predicted log odds of having a heart is 6.3

probability = 0.9999999999

e^b0+b1X/(1+ e^b0+b1X) = probability of having a heart attack

Heart Data: logistic estimation

Answer some questions:

- Write down the logistic regression model.
- Interpret $\hat{\beta}_1$.
- Estimate the probability of heart decease for someone (like Pavlos) with MaxHR$\approx$200? Plug X=200 into the equation
- If we were to use this model purely for classification, how would we do so? See any issues?

Categorical Predictors

Just like in linear regression, when the predictor, X, is binary, the interpretation of the model simplifies (and there is a quick closed form solution).

In this case, what are the interpretations of $\hat{\beta}_0$ and $\hat{\beta}_1$?

For the heart data, let X be the indicator that the individual is a male or female. What is the interpretation of the coefficient estimates in this case?

The observed percentage of HD for women is 26% while it is 55% for men.

Calculate the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ if the indicator for HD was predicted from the gender indicator.

with a single binary predictor, there is a closed form solution since you can just plug in to solve for B0, B1

Statistical Inference in Logistic Regression

most naive model - predict everyone to be the most common category

The **uncertainty of the estimates** $\hat{\beta}_0$ and $\hat{\beta}_1$ can be quantified and used to calculate both confidence intervals and hypothesis tests.

The estimate for the standard errors of these estimates, likelihood-based, is based on a quantity called Fisher's Information (beyond the scope of this class), which is related to the curvature of the log-likelihood function.

Due to the nature of the underlying Bernoulli distribution, if you estimate the underlying proportion p_i, you get the variance for free! Because of this, the inferences will be based on the normal approximation (and not t-distribution based).

Of course, you could always **bootstrap** the results to perform these inferences as well.

hypothesis testing to determine significance of predictors (we need uncertainty of estimates!)
you can use bootstrapping or closed form solution

Classification using the Logistic Model

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 mentioned before, 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 $Y = 0$.

Using such an approach is called the standard **Bayes classifier**.

minimizes misclassification

The Bayes classifier takes the approach that assigns each observation to the most likely class, given its predictor values.

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:

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

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

You can't differentiate this loss function - bad for choosing parameters

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

problem if you have 99% no heart disease, 1% heart disease - if you predict everyone as no heart disease, you already have 99% accuracy. Bayes classifier can't help you with this

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?

Multiple Logistic Regression

Multiple Logistic Regression

It is simple to illustrate examples in logistic regression when there is just one predictors variable.

But the approach 'easily' generalizes to the situation where there are multiple predictors.

A lot of the same details as linear regression apply to logistic regression. Interactions can be considered. Multicollinearity is a concern. So is overfitting. Etc...

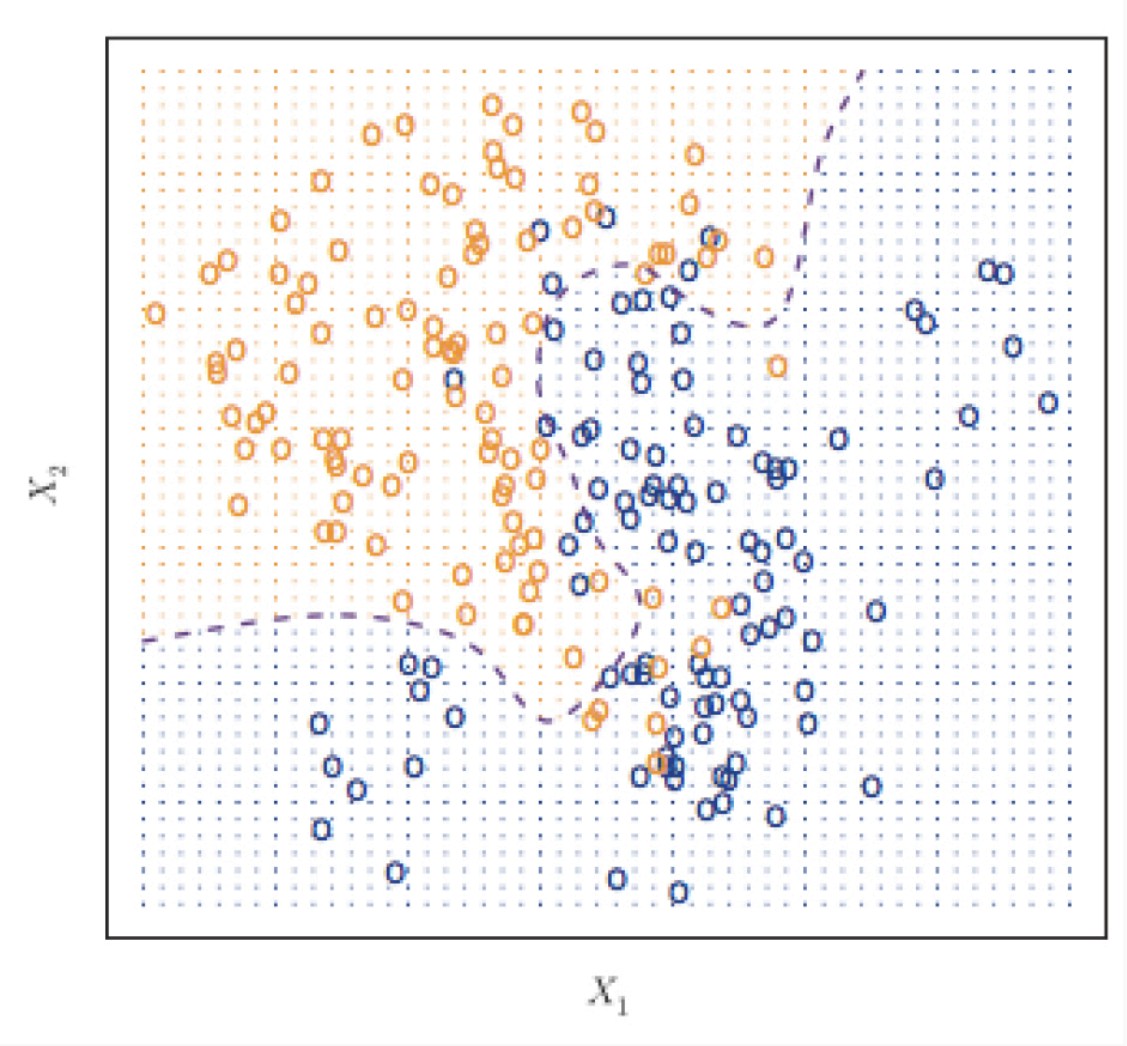
So how do we correct for such problems?

Regularization and checking though train, test, and cross-validation!

We will get into the details of this, along with other extensions of logistic regression, in the next lecture.

Classifier with two predictors

How can we estimate a classifier, based on logistic regression, for the following plot?



Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model. What was the model statement (in terms of linear predictors)?

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

Multiple logistic regression is a generalization to multiple predictors. More specifically we can define a multiple logistic regression model to predict $P(Y=1)$ as such:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Fitting Multiple Logistic Regression

The estimation procedure is identical to that as before for simple logistic regression:

- a likelihood approach is taken, and the function is maximized across all parameters $\beta_0, \beta_1, \ldots, \beta_p$ using an iterative method like Newton-Raphson.

The actual fitting of a Multiple Logistic Regression is easy using software (of course there's a python package for that) as the iterative maximization of the likelihood has already been hard coded.

In the `sklearn.linear_model` package, you just have to create your multidimensional design matrix X to be used as predictors in the `LogisticRegression` function.

Interpretation of Multiple Logistic Regression

Interpreting the coefficients in a multiple logistic regression is similar to that of linear regression.

Key: since there are other predictors in the model, the coefficient $\hat{\beta}_j$ is the association between the j^{th} predictor and the response (on log odds scale). But do we have to say? Controlling for the other predictors in the model.

We are trying to attribute the partial effects of each model controlling for the others (aka, controlling for possible *confounders*}.

Interpreting Multiple Logistic Regression: an Example

Let's get back to the Heart Data. We are attempting to predict whether someone has HD based on MaxHR and whether the person is female or male. The simultaneous effect of these two predictors can be brought into one model.

Recall from earlier we had the following estimated models:

$$\log\left(\frac{\widehat{P(Y=1)}}{1-\widehat{P(Y=1)}}\right) = 6.30 - 0.043 \cdot X_{MaxHR}$$

$$\log\left(\frac{\widehat{P(Y=1)}}{1-\widehat{P(Y=1)}}\right) = -1.06 + 1.27 \cdot X_{gender}$$

Interpreting Multiple Logistic Regression: an Example

The results for the multiple logistic regression model are:

```
data_x = df_heart[['MaxHR','Sex']]
data_y = df_heart['AHD']

logreg = LogisticRegression(C=100000, fit_intercept=True)
logreg.fit(data_x, data_y);



print('Estimated beta1: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1:
 [[-0.04496354  1.40079047]]
Estimated beta0:
 [ 5.58662464]
```

no collinearity if predictor coeffs don't change when going from a single predictor model to multiple predictor model

Some questions

1. Estimate the odds ratio of HD comparing men to women using this model.
2. Is there any evidence of multicollinearity in this model?
3. Is there any confounding in this problem?

Siomilar idea as collinearity

Interactions in Multiple Logistic Regression

Just like in linear regression, interaction terms can be considered in logistic regression. An **interaction terms** is incorporated into the model the same way, and the interpretation is very similar (on the log-odds scale of the response of course).

Write down the model for the Heart data for the 2 predictors plus the interactions term.

Interpreting Multiple Logistic Regression: an Example

The results for the multiple logistic regression model are:

Slide Type -

```
df_heart['Interaction'] = df_heart.MaxHR * df_heart.Sex

data_x = df_heart[['MaxHR','Sex', 'Interaction']]
data_y = df_heart['AHD']

logreg = LogisticRegression(C=100000, fit_intercept=True)
logreg.fit(data_x, data_y);


print('Estimated beta1, beta2, beta3: \n', logreg.coef_)
print('Estimated beta0: \n', logreg.intercept_)
```

```
Estimated beta1, beta2, beta3:
 [[-0.02645985  5.38749287 -0.02689767]]
Estimated beta0:
 [ 2.88218441]
```

interaction between male*maxHR - relationship between maxHR and gender more negative for me

Some questions

1. Write down the complete model. Break this down into the model to predict log-odds of heart disease (HD) based on MaxHR for women and the same model for men. How is this different from the previous model (without interaction)?

2. Interpret the results of this model. What does the coefficient for the interaction term represent?

3. Estimate the odds ratio of HD comparing men to women using this model [trick question].

4. Is there any evidence of multicollinearity in this model?

if you have an interaction, you will have collinearity by construction

5. Is there any confounding in this problem?

Extending the Logistic Model

Model Diagnostics in Logistic Regression

In linear regression, when is the model appropriate (aka, what are the assumptions)?

In logistic regression, when is the model appropriate?

We don't have to worry about the distribution of the residuals (we get that for free).

What we do have to worry about is how Y 'links' to X in its relationship. More specifically, we assume the 'S'-shaped (aka, sigmoidal) curve follows the logistic function. How could we check this?

Are there other 'S' shaped functions?

Alternatives to logistic regression

Why was the logistic function chosen to model how a binary response variable can be predicted from a quantitative predictor?

Because it it takes as inputs values in $(0,1)$ and outputs values $(-\infty, \infty)$ so that the estimation of β is unbounded.

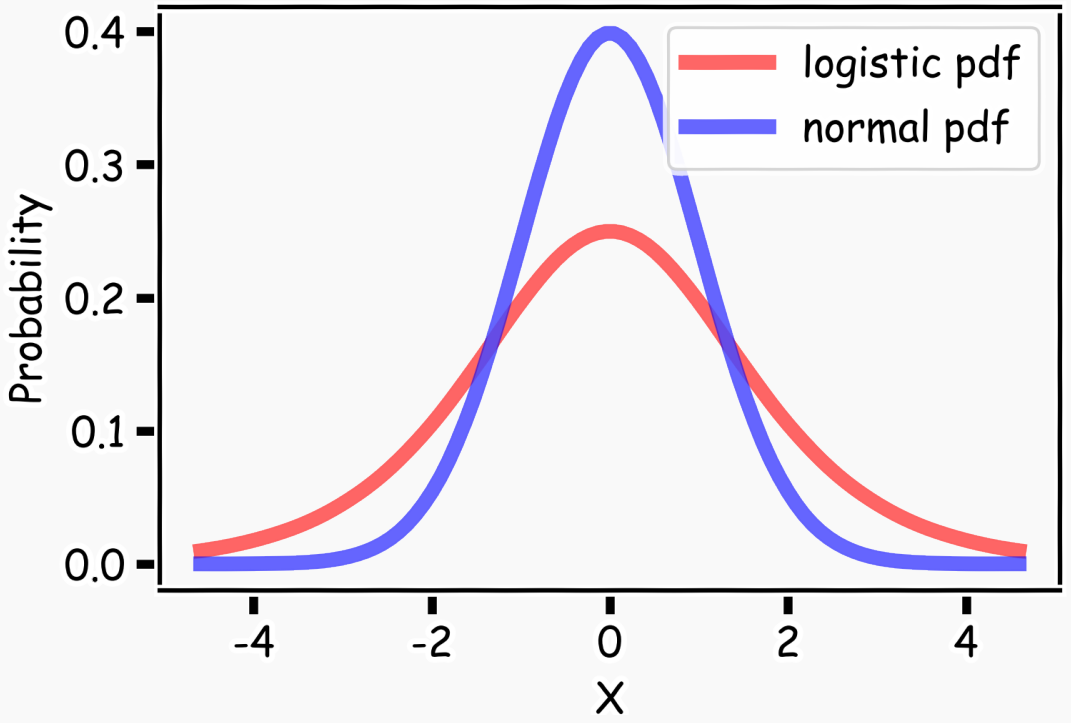
This is not the only function that does this. Any suggestions?

Any *inverse CDF* function for an unbounded continuous distribution can work as the `link' between the observed values for $Y$ and how it relates `linearly' to the predictors.

So what are possible other choices? What differences do they have? Why is logistic regression preferred?

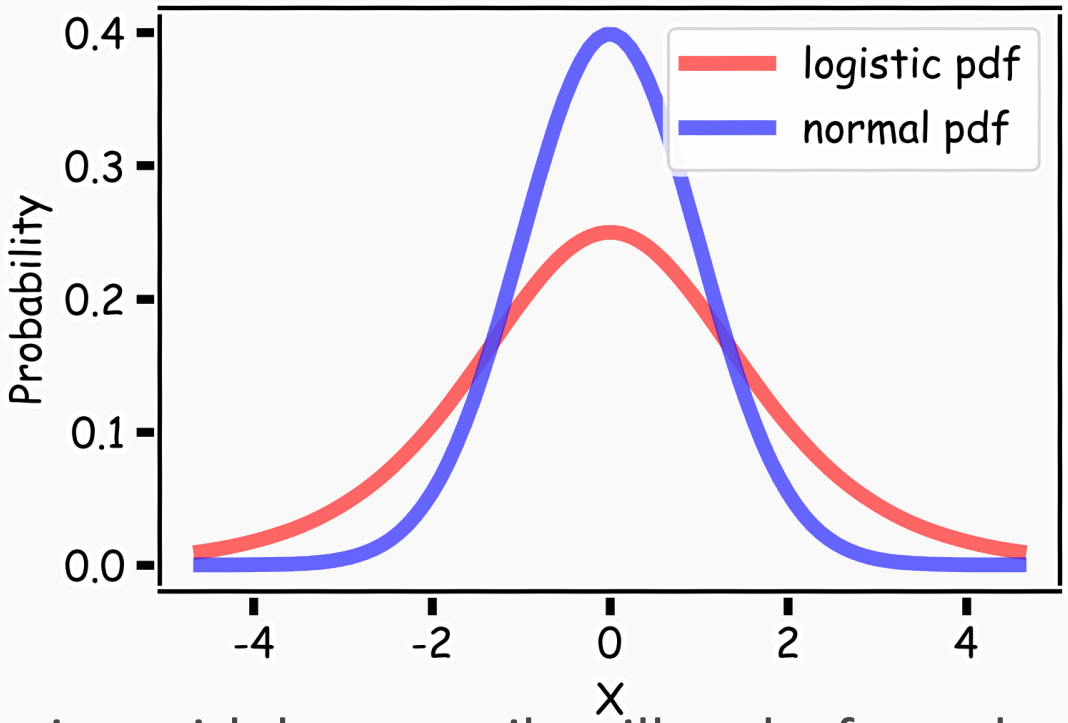
Logistic vs Normal pdf

The choice of link function determines the shape of the S' shape. Let's compare the pdf's for the Logistic and Normal distributions (called a 'probit' model, econometricians love these):



So what?

Logistic vs Normal pdf



Choosing a distribution with longer tails will make for a shape that asymptotes more slowly (likely a good thing for model fitting).

Classification Boundaries

Classification boundaries

Recall that we could attempt to purely classify each observation based on whether the estimated $P(Y = 1)$ from the model was greater than 0.5.

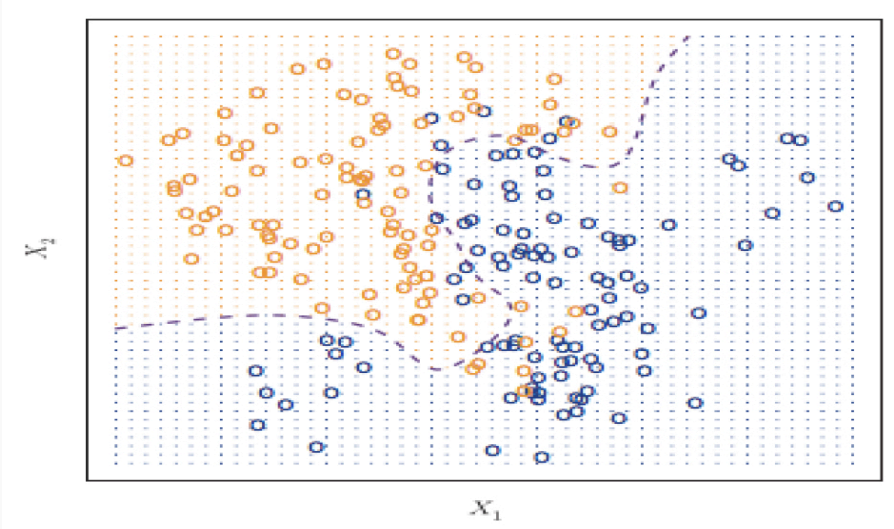
When dealing with 'well-separated' data, logistic regression can work well in performing classification.

We saw a 2-D plot last time which had two predictors, X_1, X_2 and depicted the classes as different colors. A similar one is shown on the next slide.

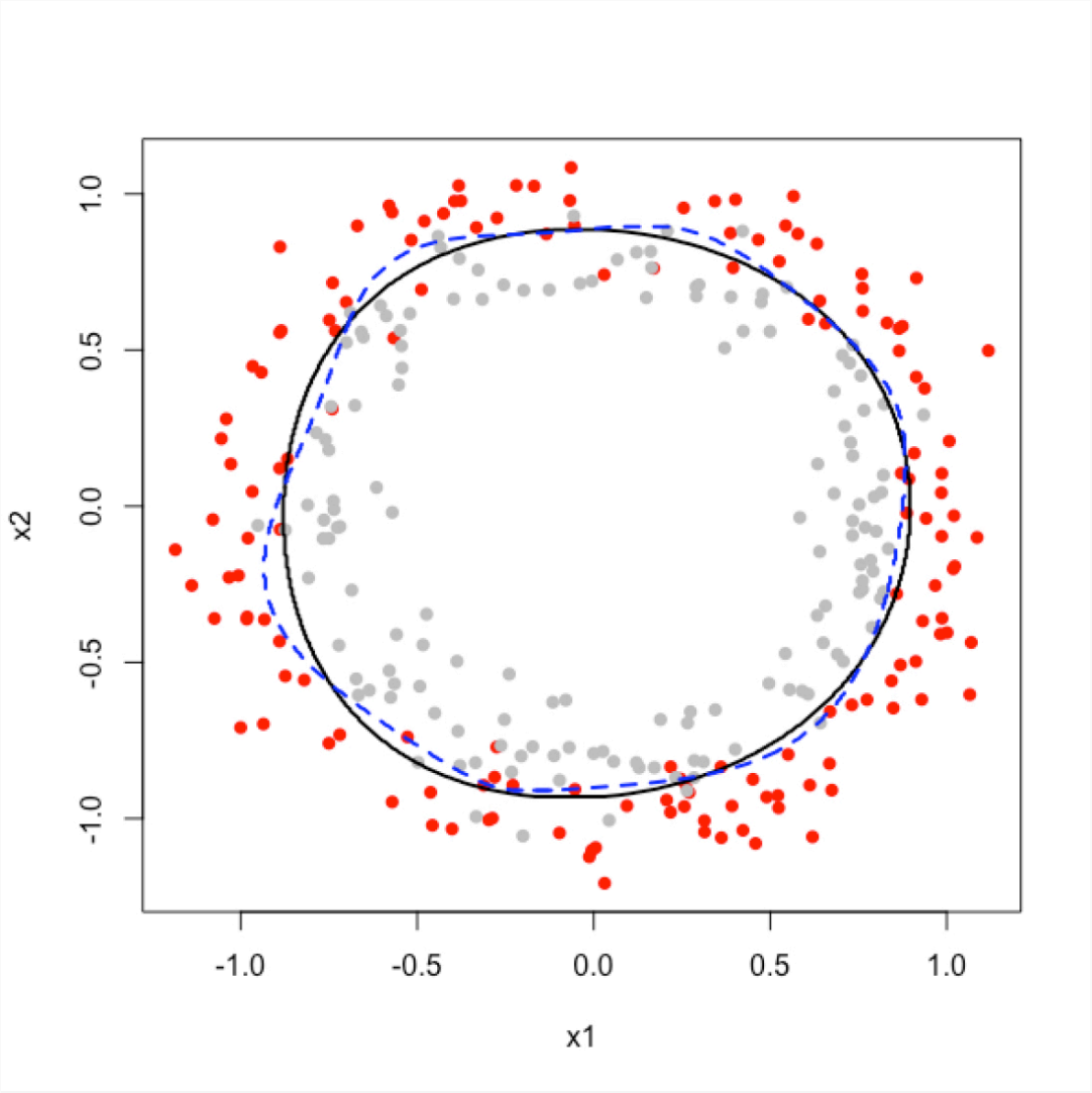
logistic regression - maps a straight line to separate the orange and the blue

How to improve using logistic regression machinery:
Add polynomials and regularize?

Other ways to do it?
kNN regression



2D Classification in Logistic Regression: an Example



2D Classification in Logistic Regression: an Example

Would a logistic regression model perform well in classifying the observations in this example?

What would be a good logistic regression model to classify these points?

Based on these predictors, two separate logistic regression model were considered that were based on different ordered polynomials of X_1, X_2 and their interactions. The 'circles' represent the boundary for classification.

How can the classification boundary be calculated for a logistic regression?

2D Classification in Logistic Regression: an Example

In the previous plot, which classification boundary performs better? How can you tell? How would you make this determination in an actual data example?

We could determine the misclassification rates in left out validation or test set(s)