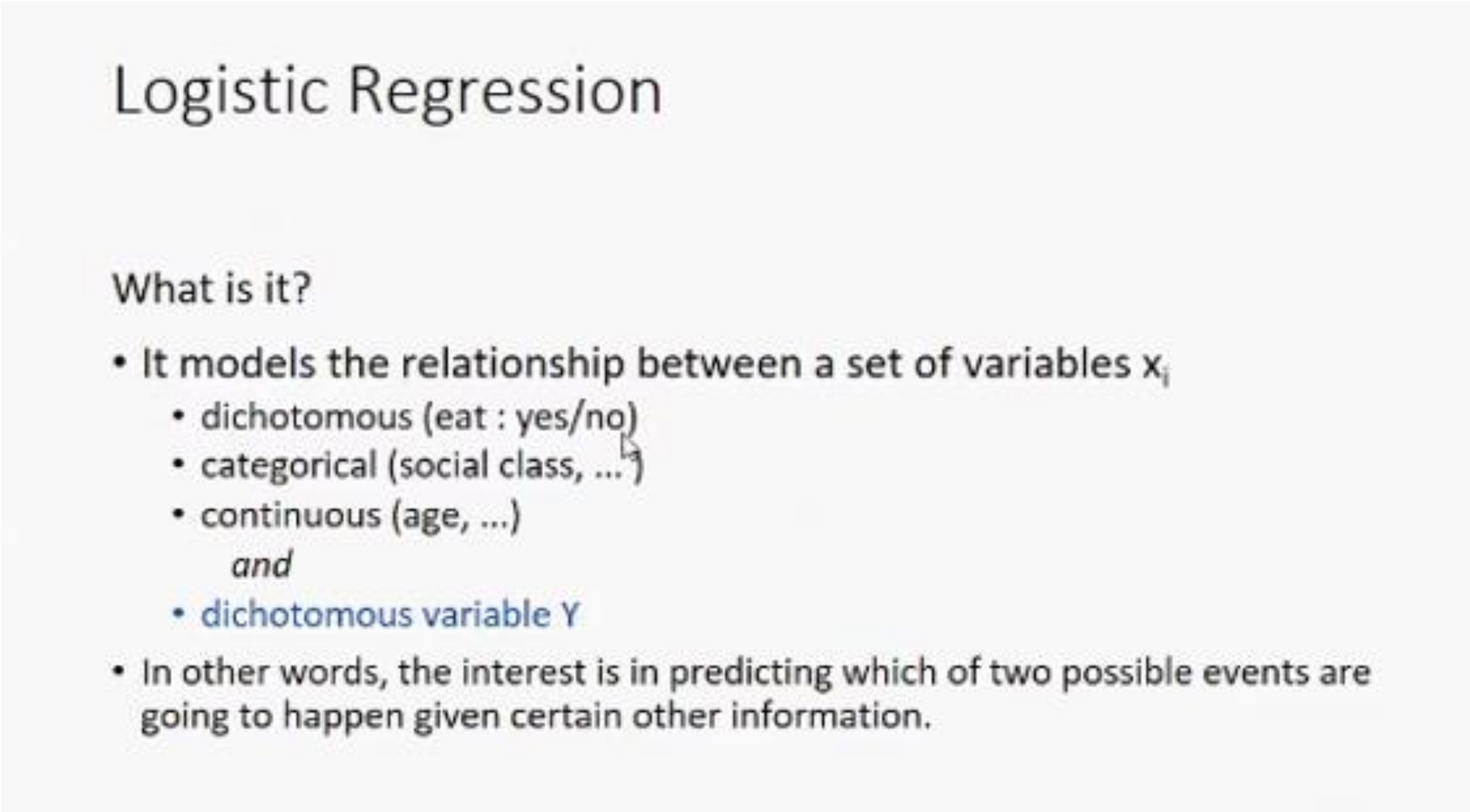
Logistic Regression

While performing the data analysis we face multiple problems like understanding the problem statement, collecting the correct data, selecting the relevant data, which technique I should follow etc.



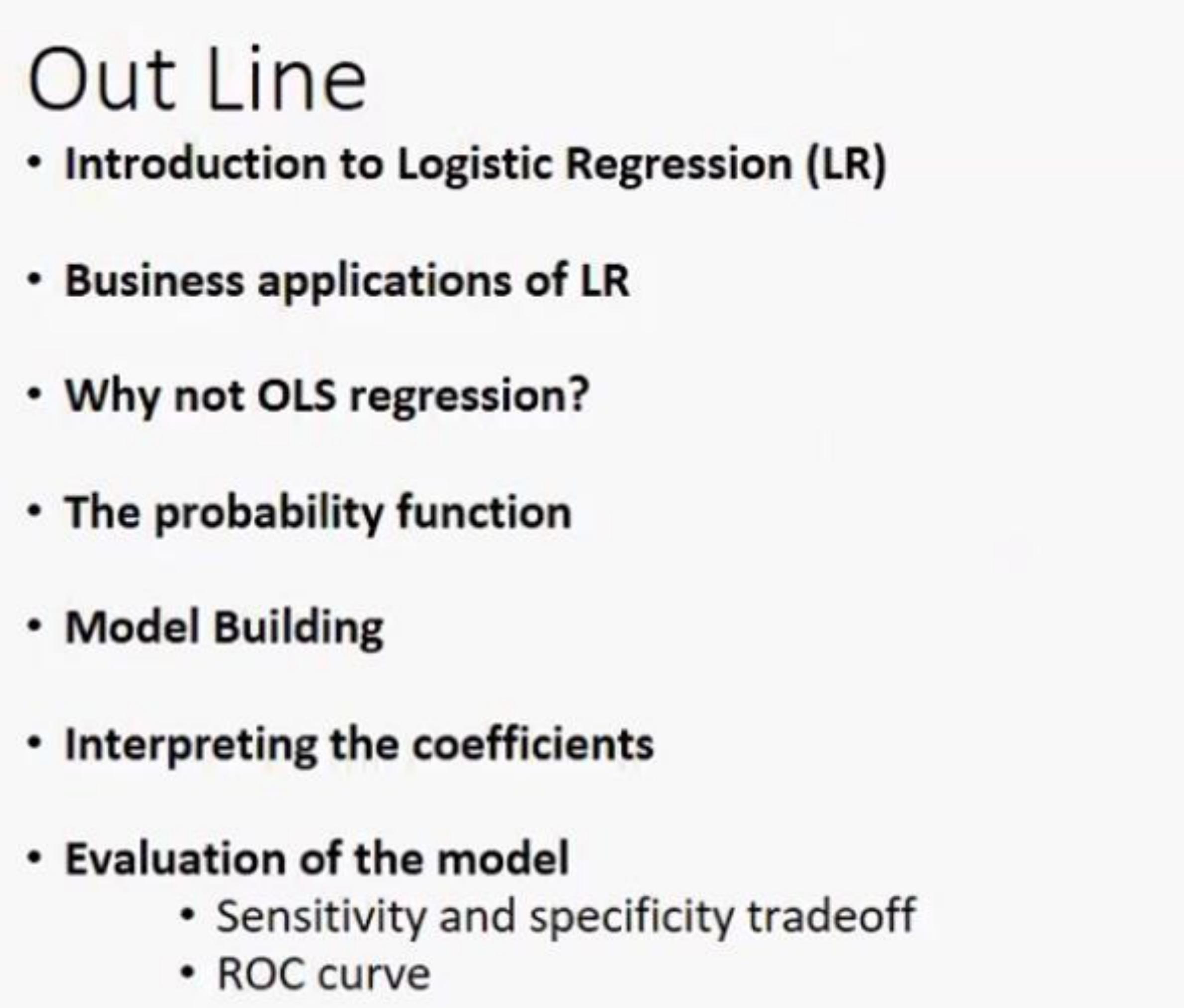
Logistic Regression

What is it?

- It models the relationship between a set of variables x_i
  - dichotomous (eat : yes/no)
  - categorical (social class, ...)
  - continuous (age, ...)

    and
  - dichotomous variable Y
- In other words, the interest is in predicting which of two possible events are going to happen given certain other information.

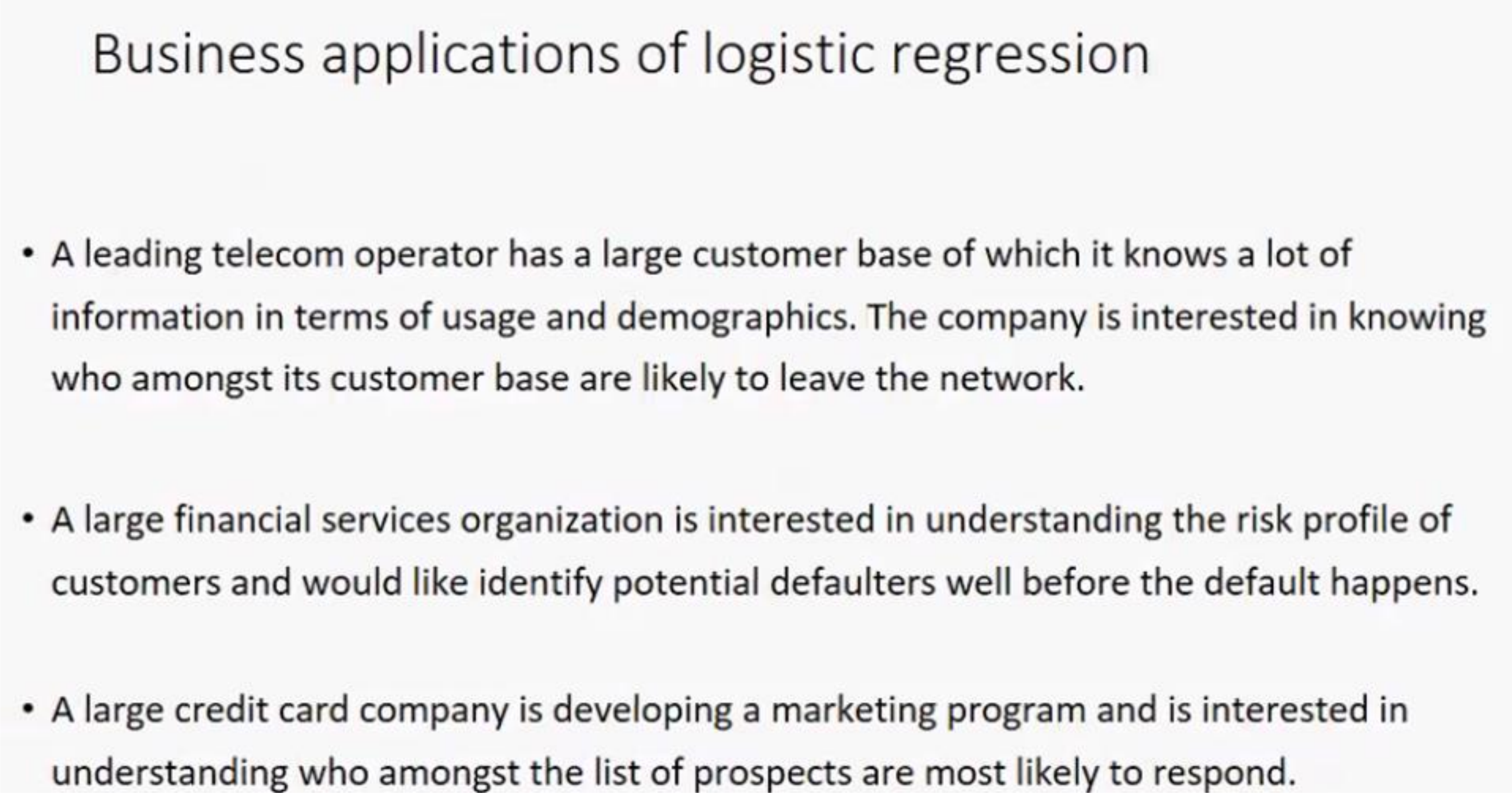
Logistic Regression



Out Line

- **Introduction to Logistic Regression (LR)**
- **Business applications of LR**
- **Why not OLS regression?**
- **The probability function**
- **Model Building**
- **Interpreting the coefficients**
- **Evaluation of the model**
  - Sensitivity and specificity tradeoff
  - ROC curve

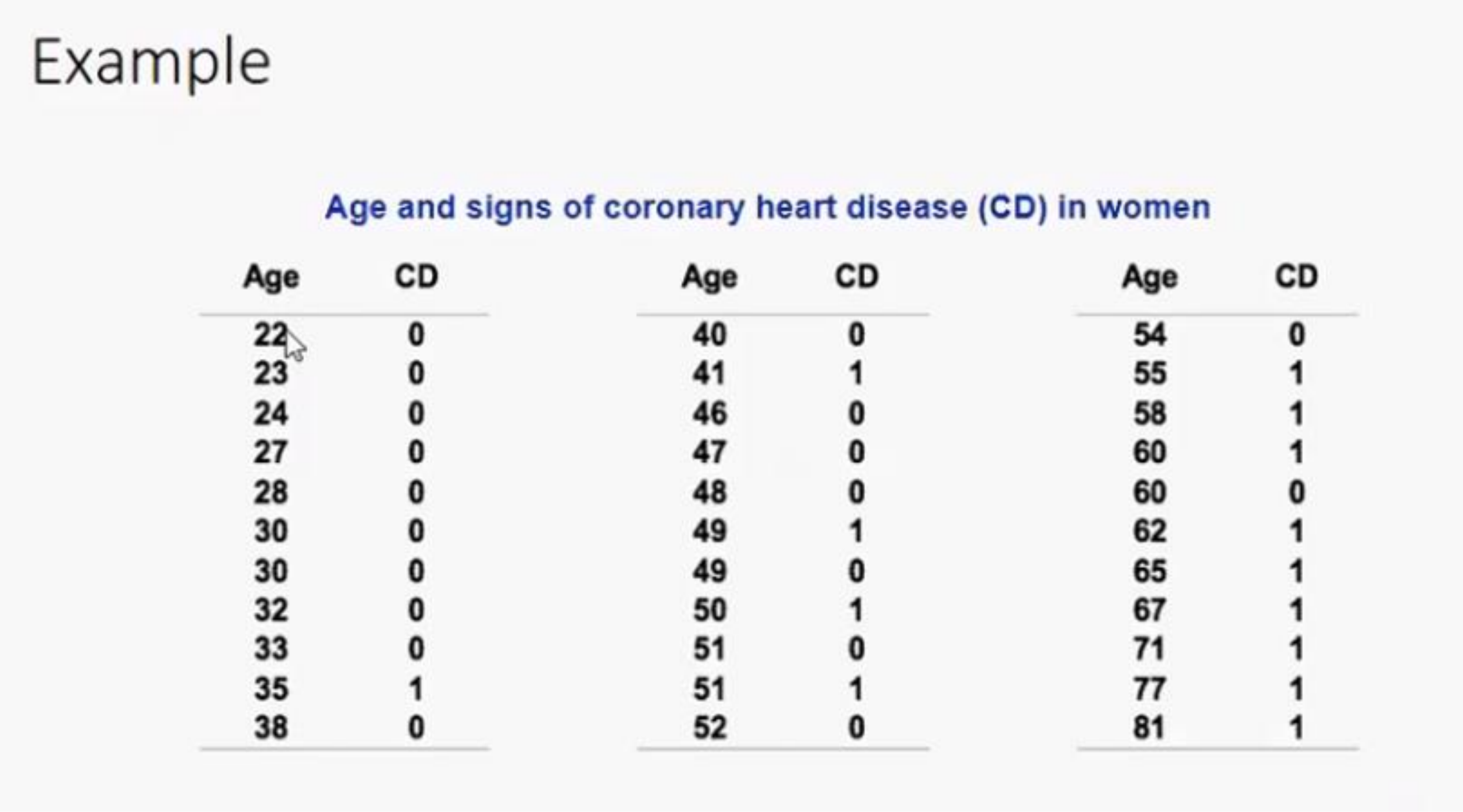
Logistic Regression



Business applications of logistic regression

- A leading telecom operator has a large customer base of which it knows a lot of information in terms of usage and demographics. The company is interested in knowing who amongst its customer base are likely to leave the network.
- A large financial services organization is interested in understanding the risk profile of customers and would like identify potential defaulters well before the default happens.
- A large credit card company is developing a marketing program and is interested in understanding who amongst the list of prospects are most likely to respond.

Logistic Regression

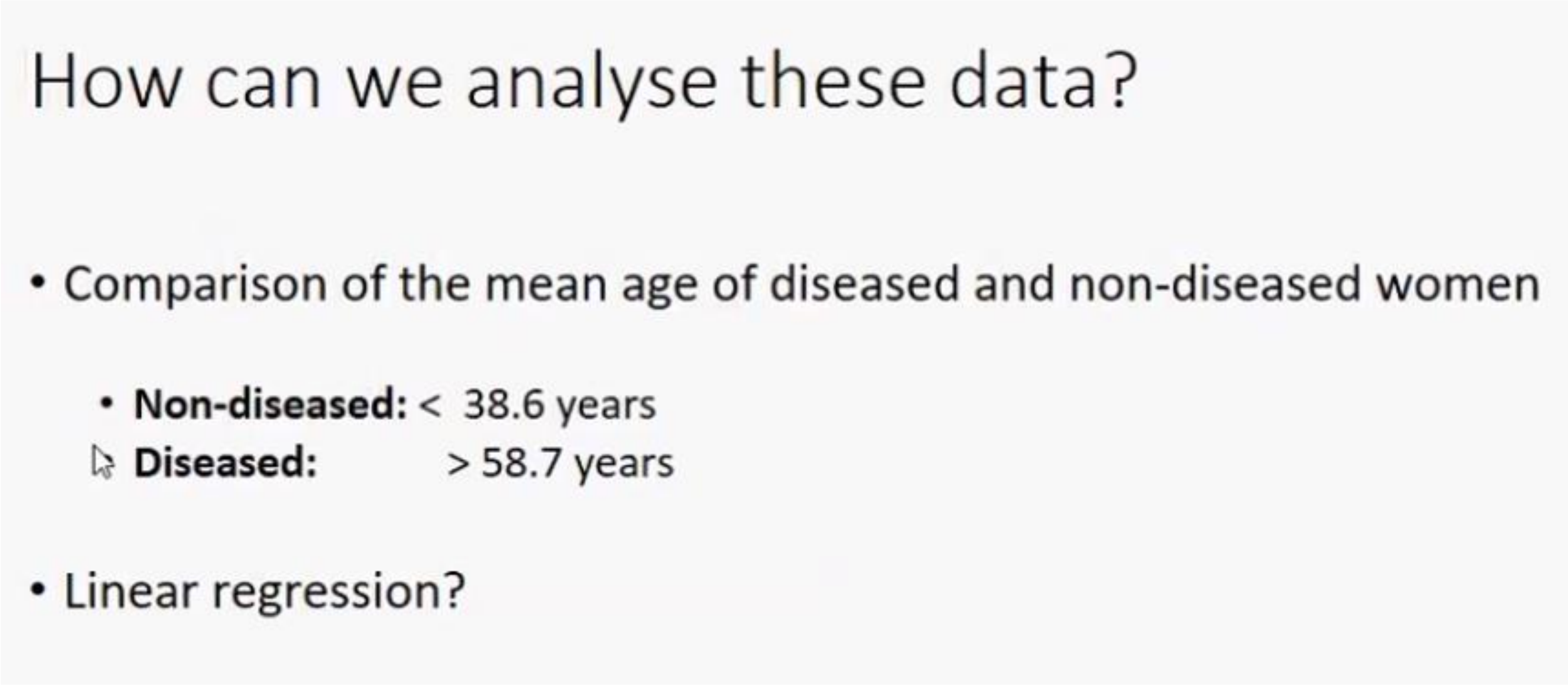


Example

Age and signs of coronary heart disease (CD) in women

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

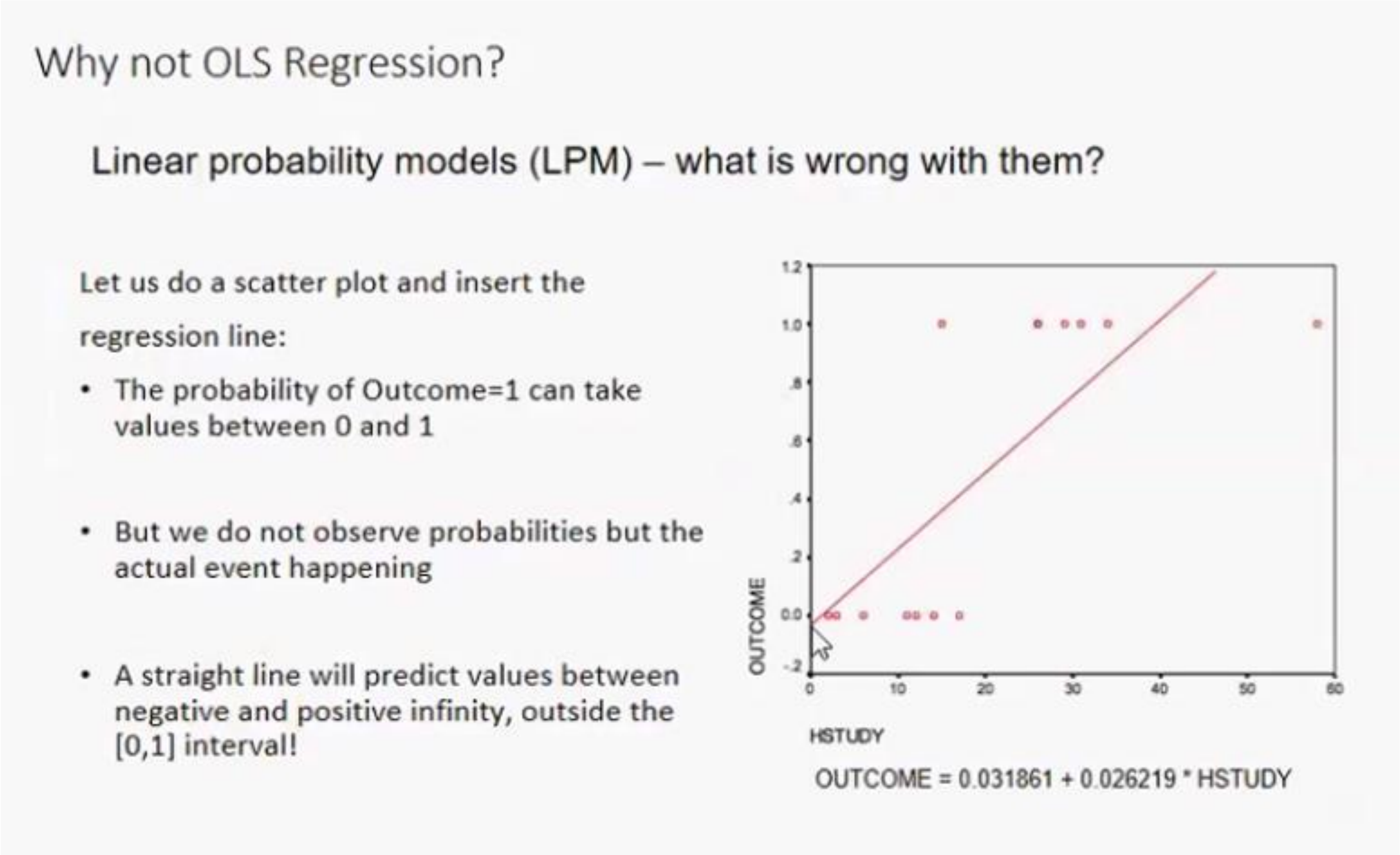
Logistic Regression



How can we analyse these data?

- Comparison of the mean age of diseased and non-diseased women
  - **Non-diseased:** < 38.6 years
  - **Diseased:** > 58.7 years
- Linear regression?

Logistic Regression



Why not OLS Regression?

Linear probability models (LPM) – what is wrong with them?

Let us do a scatter plot and insert the

regression line:

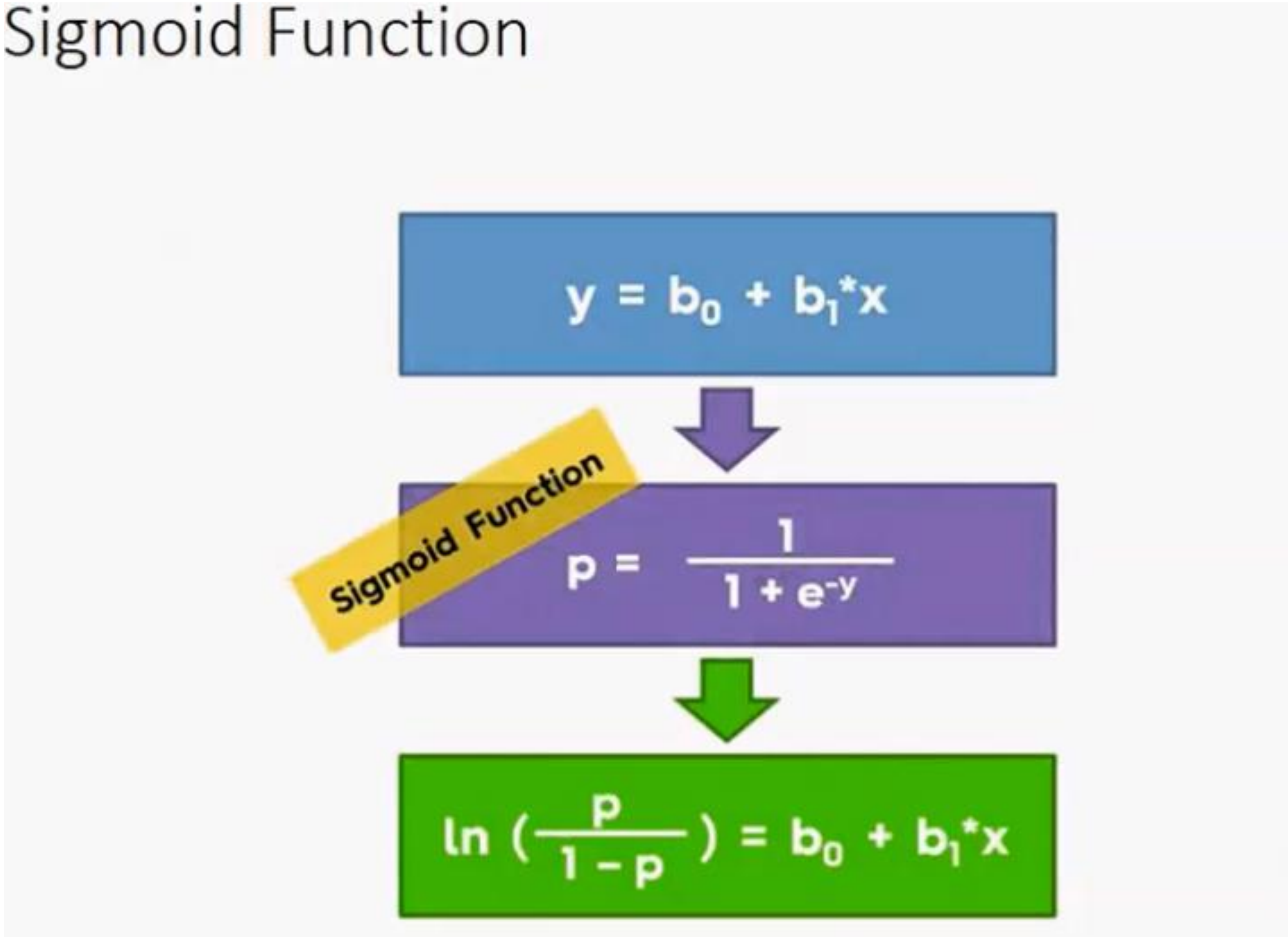
- The probability of Outcome=1 can take values between 0 and 1

- But we do not observe probabilities but the actual event happening

- A straight line will predict values between negative and positive infinity, outside the [0,1] interval!

OUTCOME = 0.031861 + 0.026219 * HSTUDY

Logistic Regression



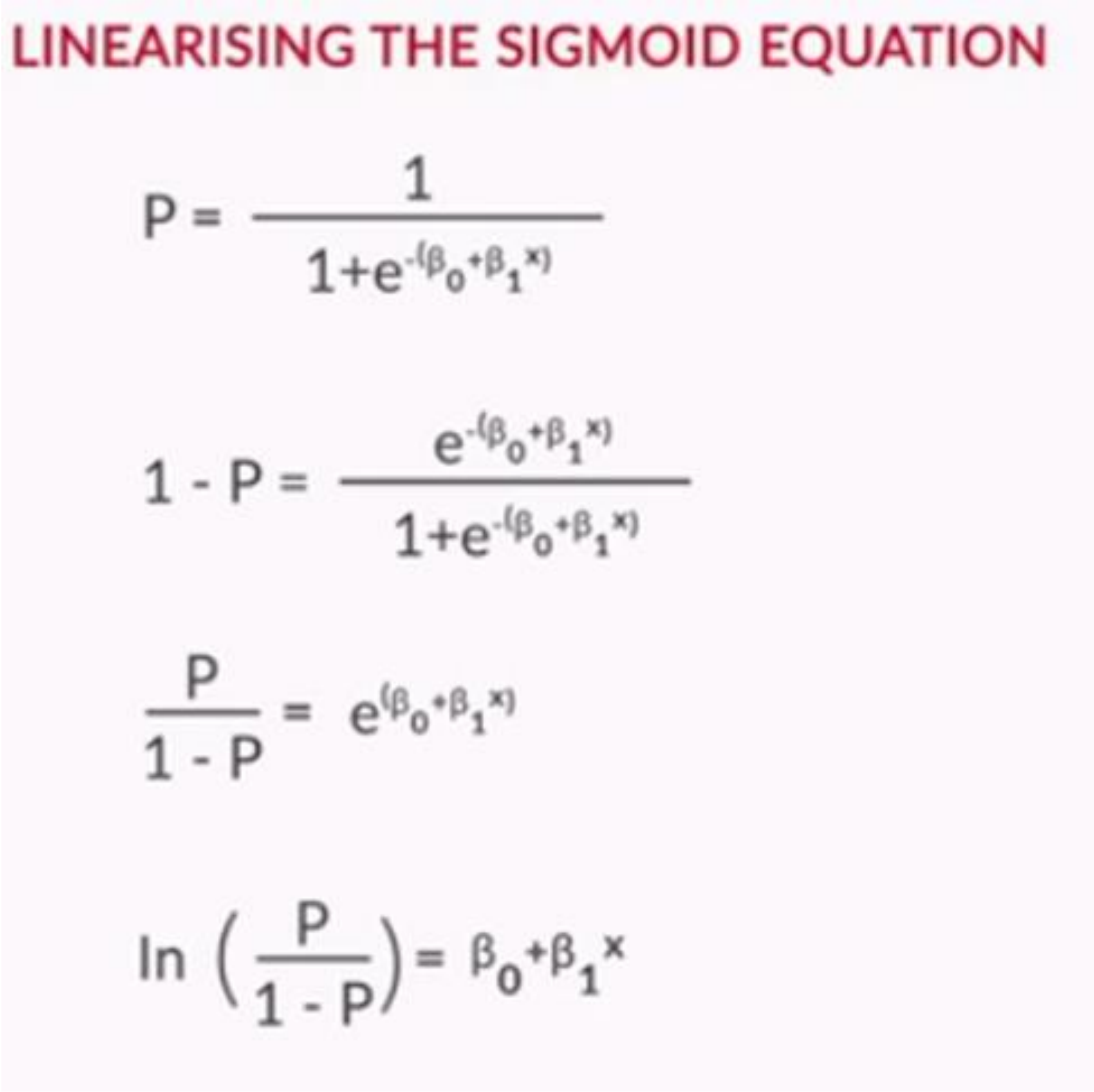
Sigmoid Function

$$y = b_0 + b_1 {}^* x$$

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1 - p}\right) = b_0 + b_1 {}^* x$$

Logistic Regression
- Bivariate Analysis



LINEARISING THE SIGMOID EQUATION

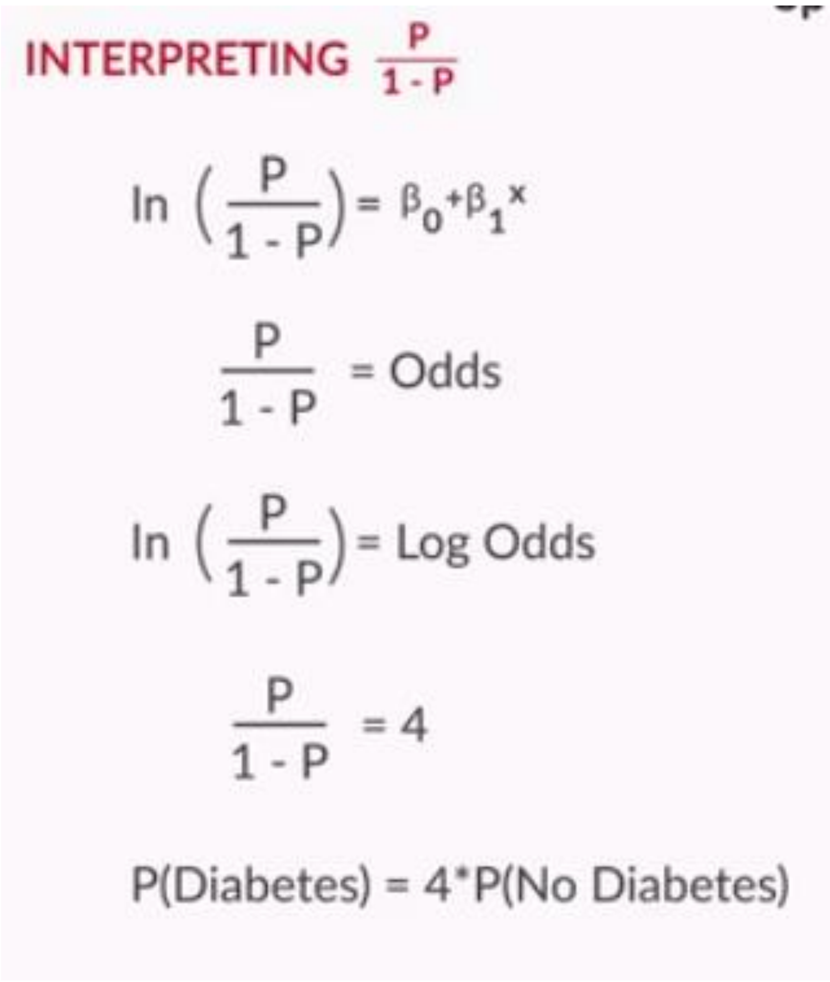
$$P = \frac{1}{1+e^{-(\beta_0+\beta_1x)}}$$

$$1 - P = \frac{e^{-(\beta_0+\beta_1x)}}{1+e^{-(\beta_0+\beta_1x)}}$$

$$\frac{P}{1 - P} = e^{(\beta_0+\beta_1x)}$$

$$\text{In}\left(\frac{P}{1 - P}\right) = \beta_0+\beta_1x$$

Logistic Regression
- Bivariate Analysis



INTERPRETING $\frac{P}{1 - P}$

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

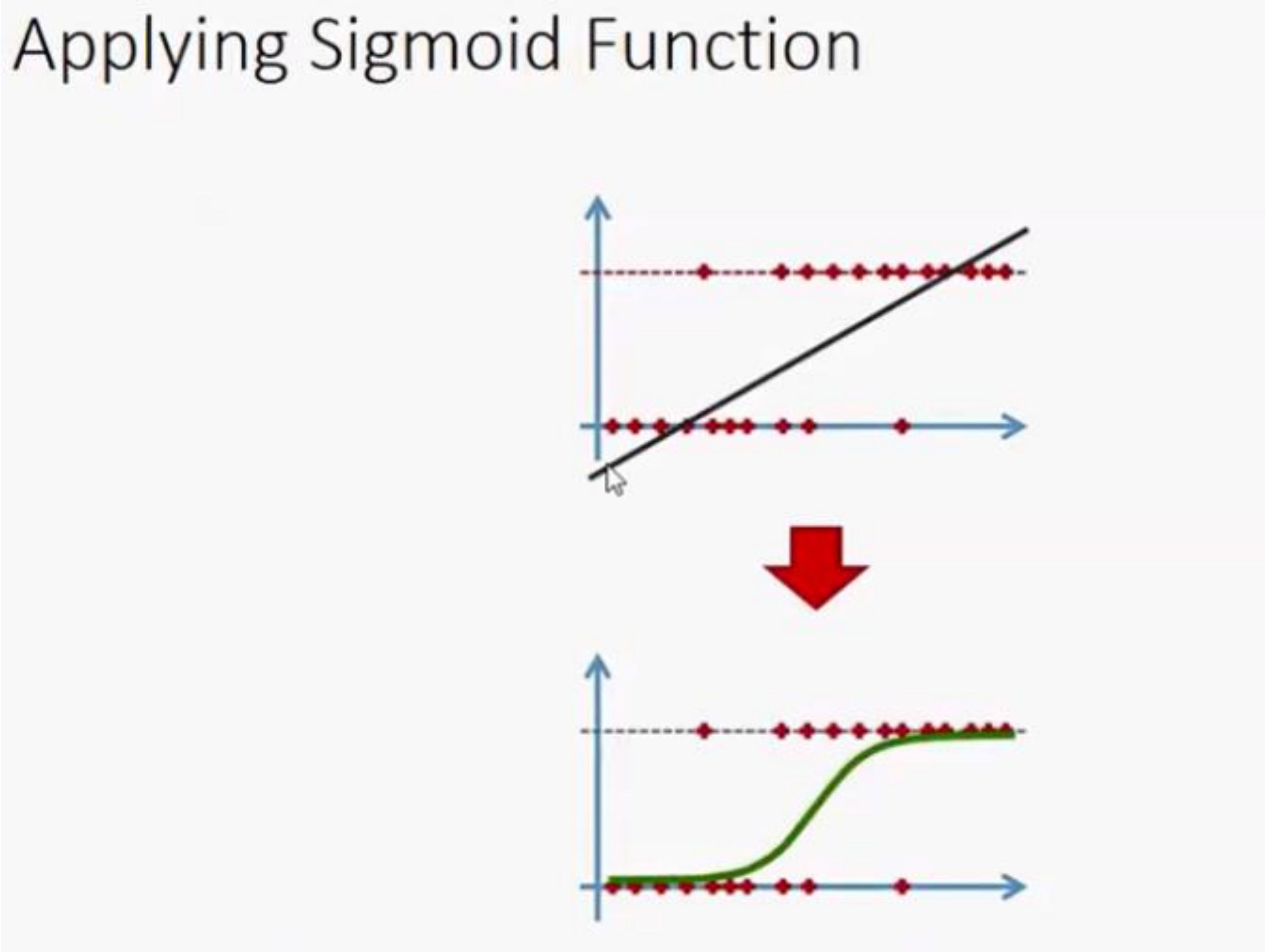
$$\frac{P}{1 - P} = \text{Odds}$$

$$\ln\left(\frac{P}{1 - P}\right) = \text{Log Odds}$$

$$\frac{P}{1 - P} = 4$$

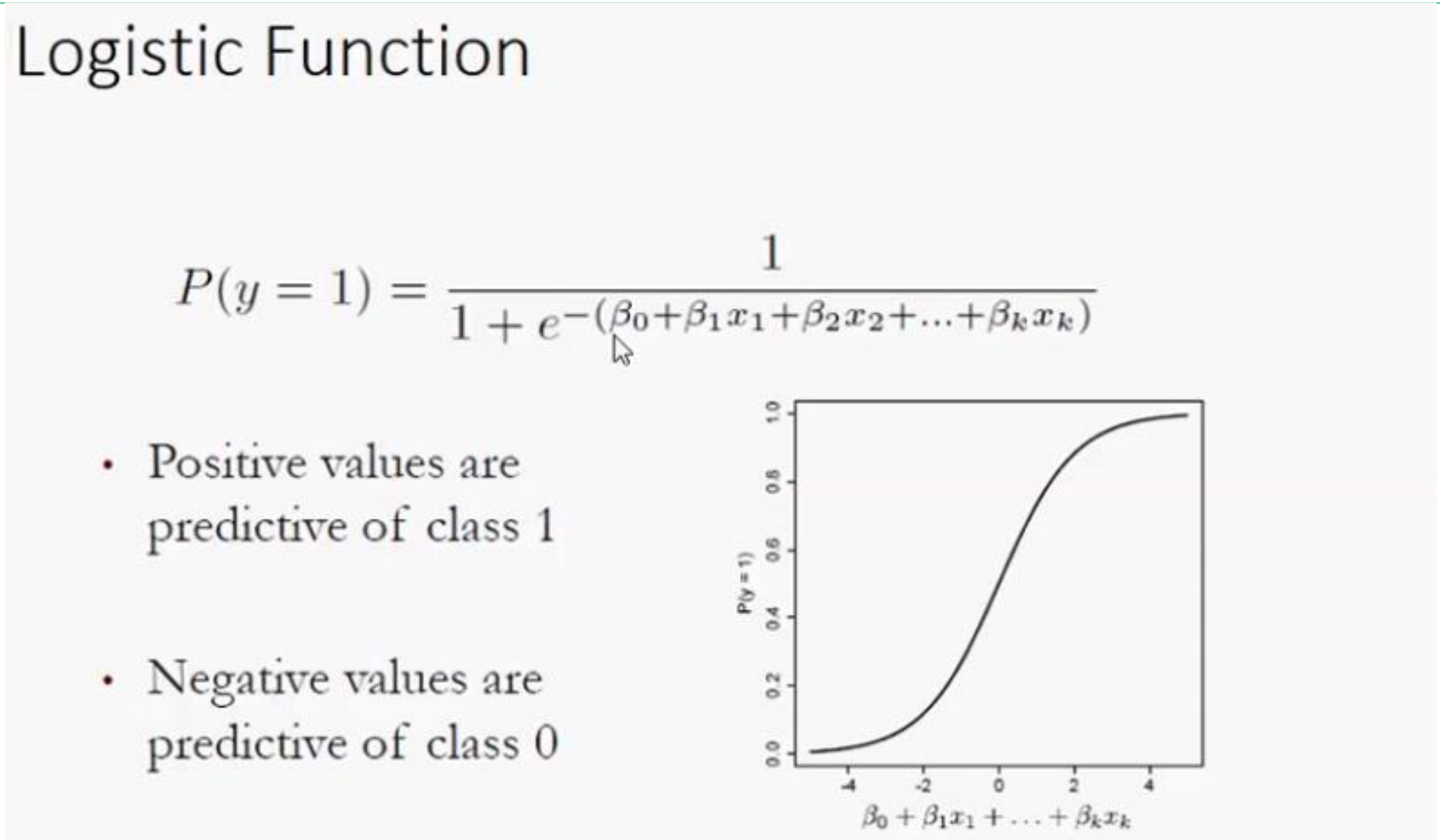
P(Diabetes) = 4*P(No Diabetes)

Logistic Regression



Applying Sigmoid Function

Logistic Regression

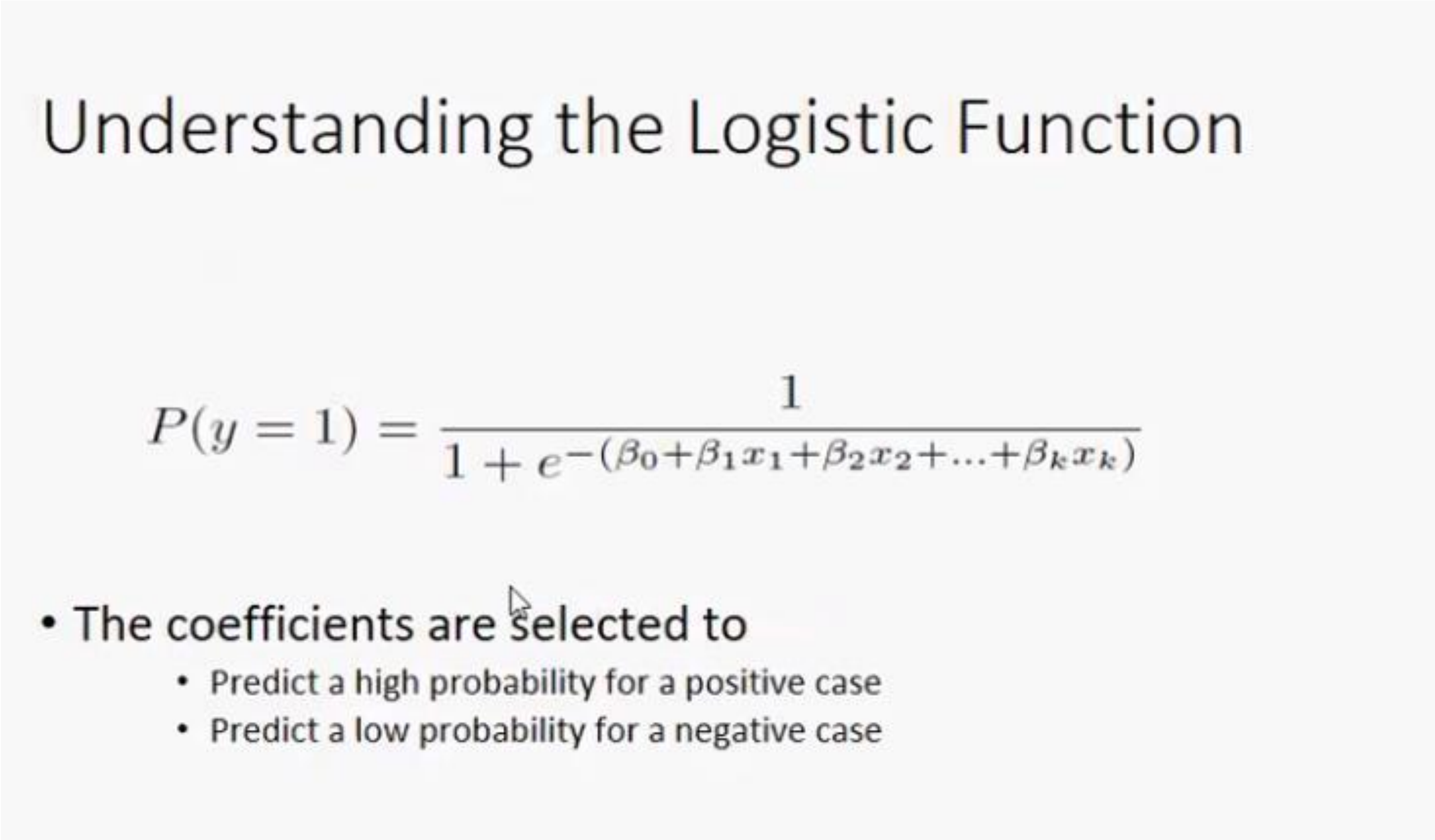


Logistic Function

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}$$

- Positive values are predictive of class 1
- Negative values are predictive of class 0

Logistic Regression

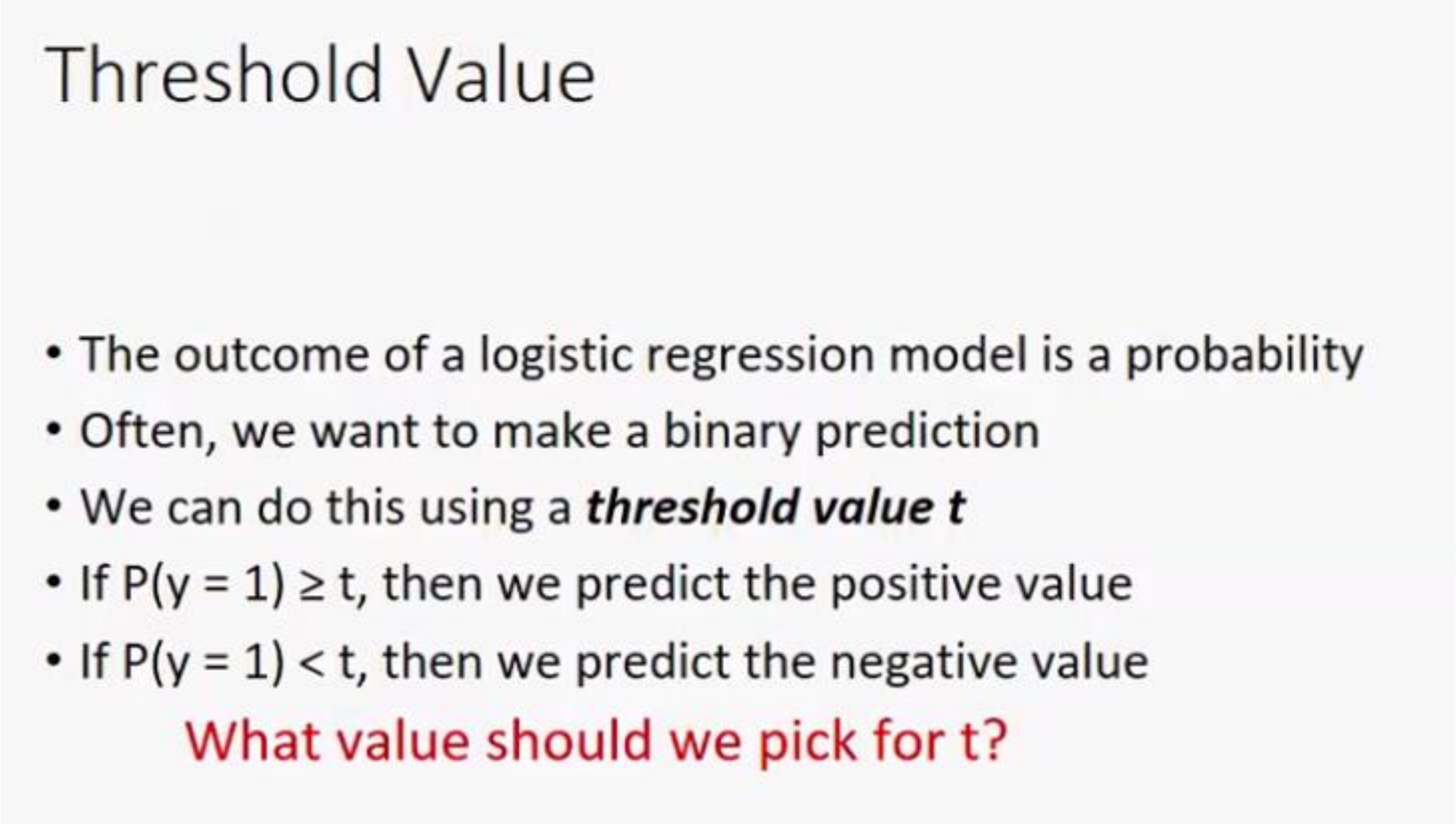


Understanding the Logistic Function

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

- The coefficients are selected to
  - Predict a high probability for a positive case
  - Predict a low probability for a negative case

Logistic Regression

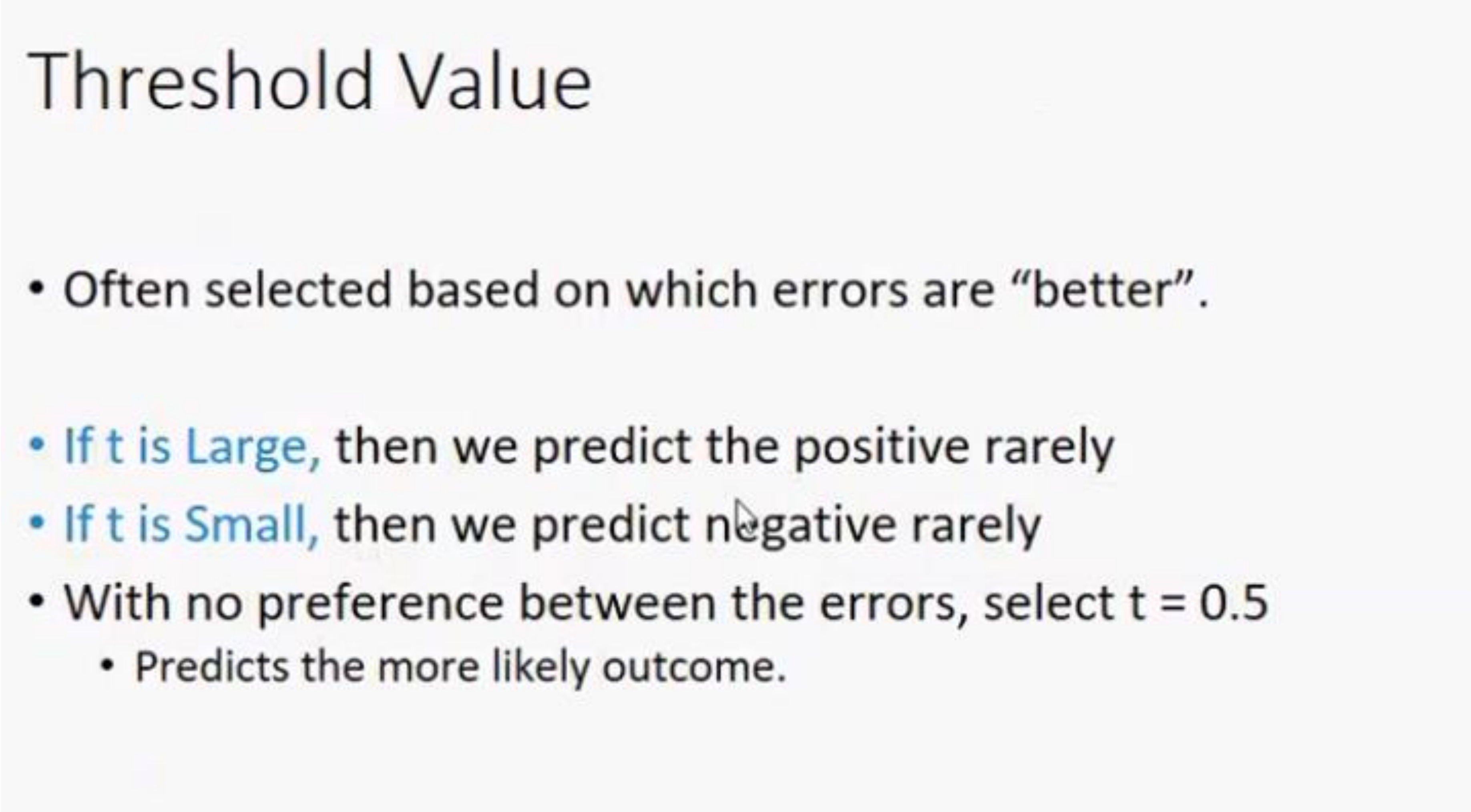


Threshold Value

- The outcome of a logistic regression model is a probability
- Often, we want to make a binary prediction
- We can do this using a ***threshold value t***
- If P(y = 1) ≥ t, then we predict the positive value
- If P(y = 1) < t, then we predict the negative value

What value should we pick for t?

Logistic Regression

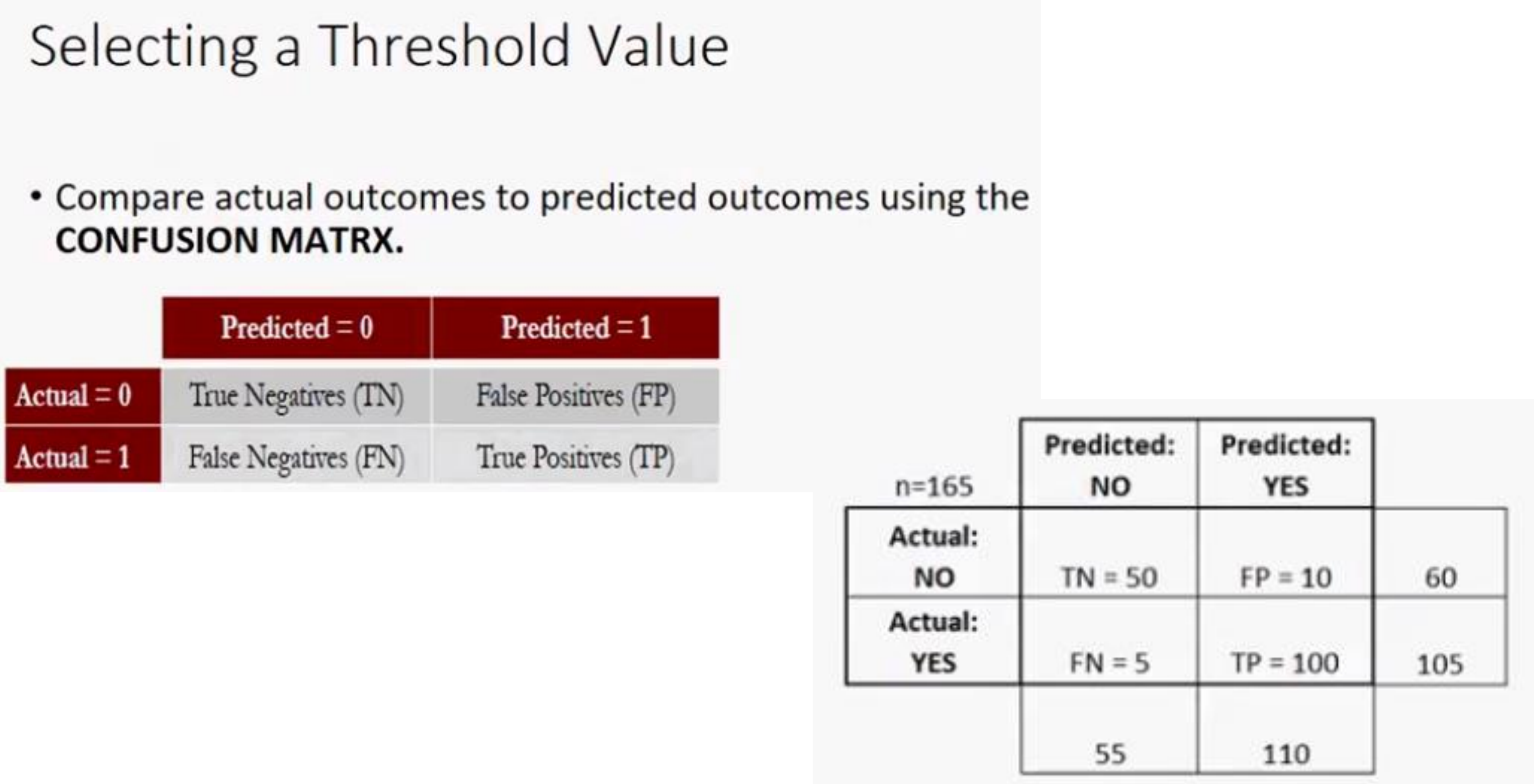


Threshold Value

- Often selected based on which errors are “better”.

- If t is Large, then we predict the positive rarely
- If t is Small, then we predict negative rarely
- With no preference between the errors, select t = 0.5
  - Predicts the more likely outcome.

Logistic Regression



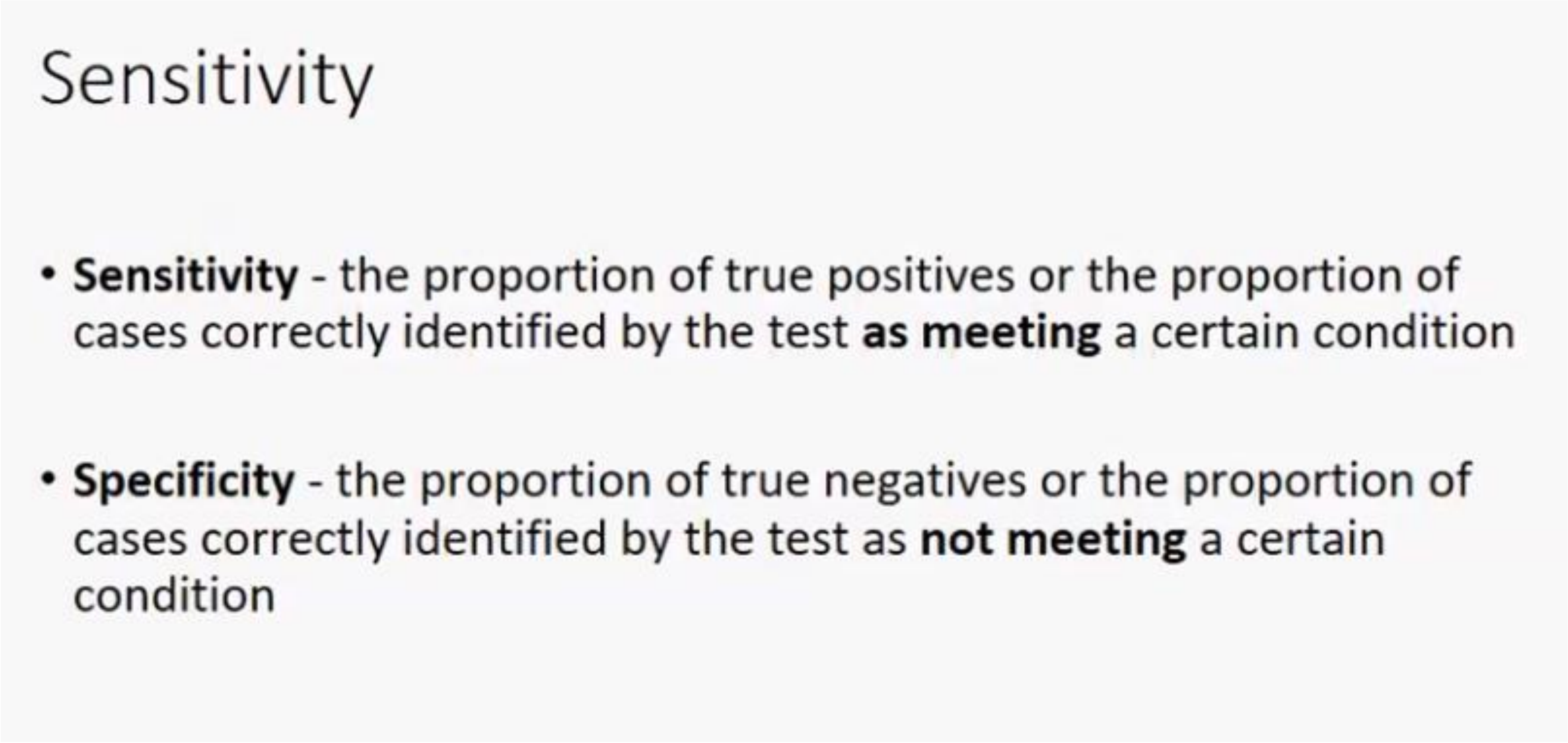
Selecting a Threshold Value

- Compare actual outcomes to predicted outcomes using the **CONFUSION MATRX.**

	Predicted = 0	Predicted = 1
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

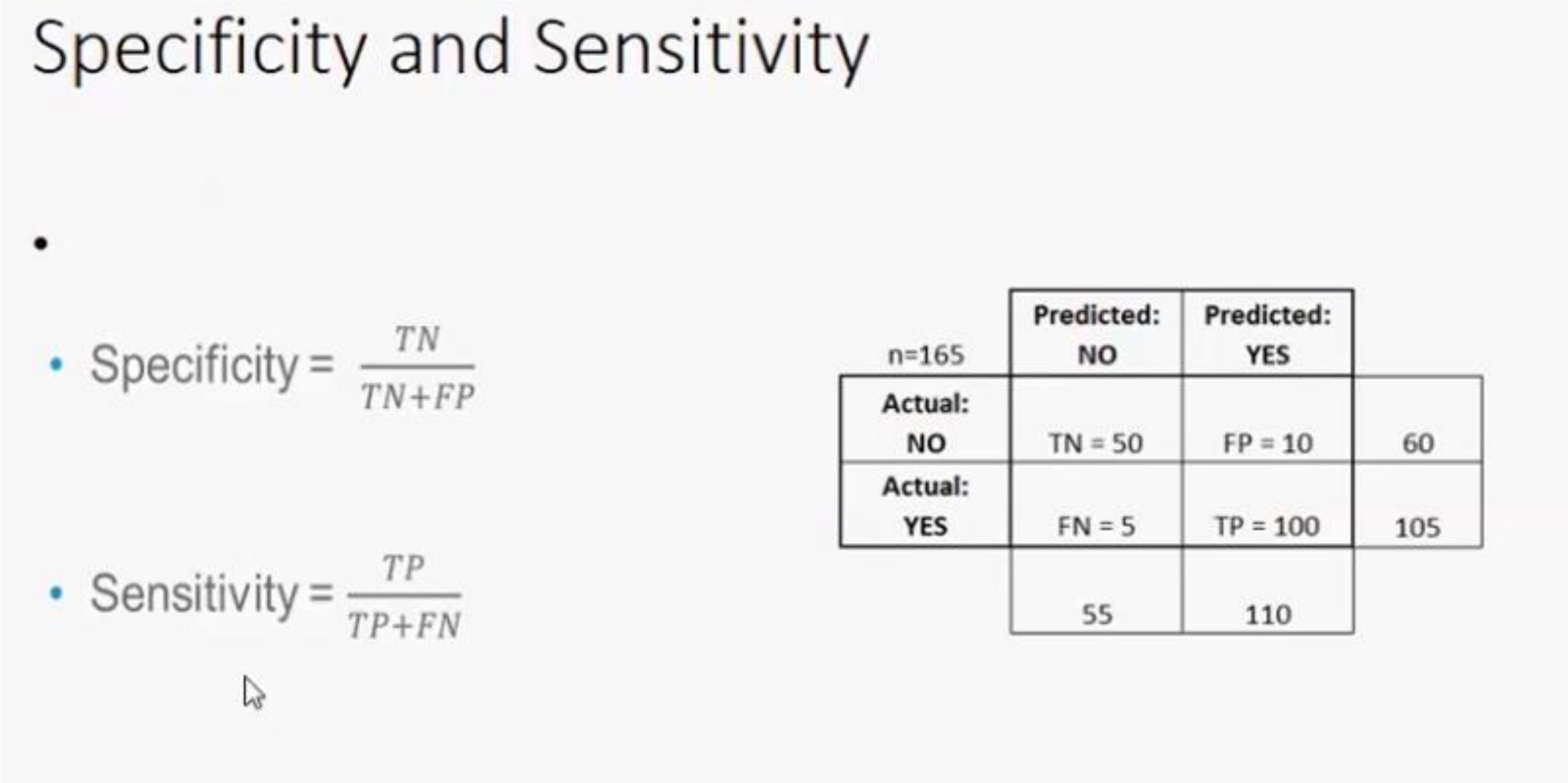
Logistic Regression



Sensitivity

- **Sensitivity** - the proportion of true positives or the proportion of cases correctly identified by the test **as meeting** a certain condition
- **Specificity** - the proportion of true negatives or the proportion of cases correctly identified by the test as **not meeting** a certain condition

Logistic Regression



Specificity and Sensitivity

-

- Specificity = $\frac{TN}{TN+FP}$

- Sensitivity = $\frac{TP}{TP+FN}$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Logistic Regression
- Bivariate Analysis

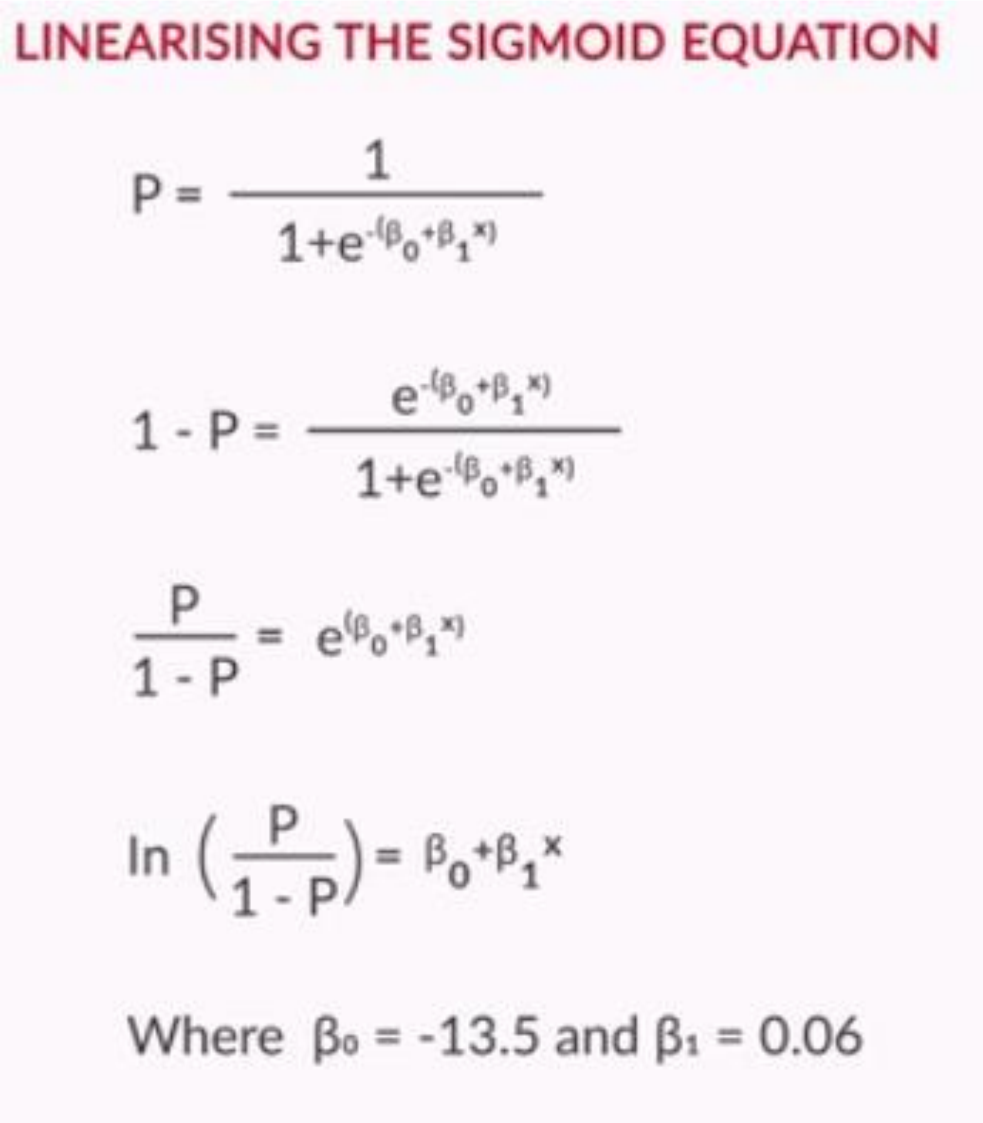
Odds and Log Odds

In the previous segment, you saw that by trying different values of β0β0 and β1β1, you can manipulate the shape of the sigmoid curve. At some combination of β0β0 and β1β1, the 'likelihood' (length of yellow bars) will be maximised.

The question is - how do you find the optimal values of β0β0 and β1β1 such that the likelihood function is maximized?

In python, logistic regression can be implemented using libraries such as sklearn and statsmodels, though looking at the coefficients and the model summary is easier using statsmodels.

Logistic Regression
- Bivariate Analysis



LINEARISING THE SIGMOID EQUATION

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

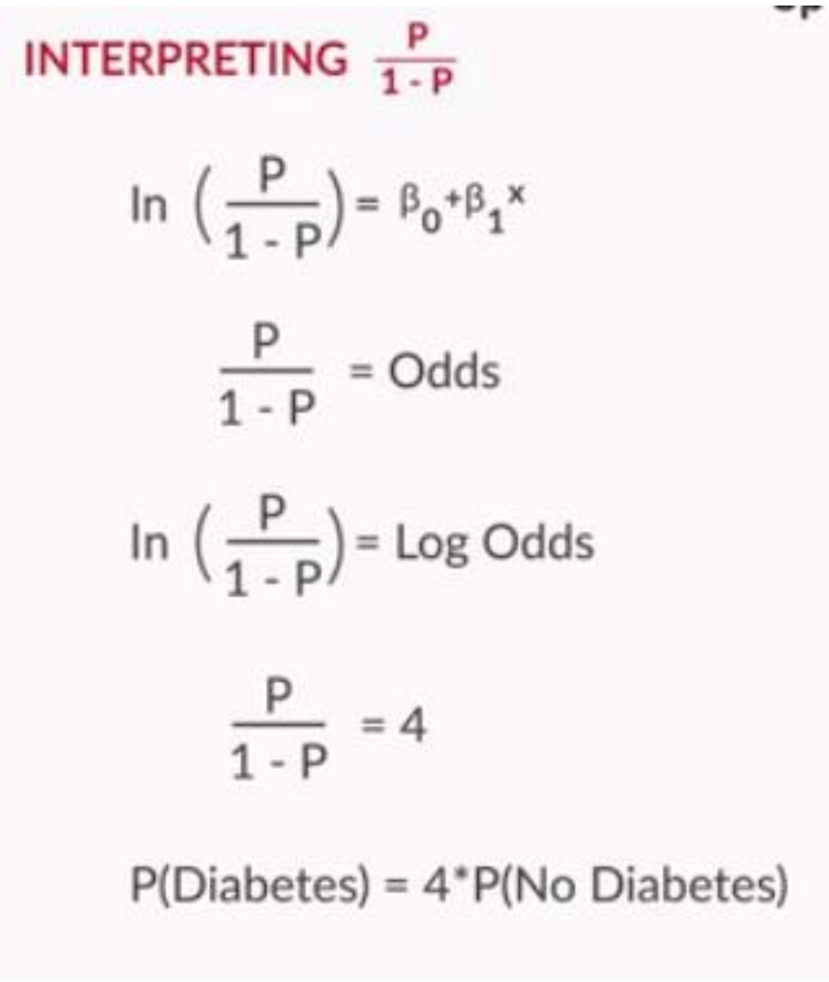
$$1 - P = \frac{e^{-(\beta_0+\beta_1 x)}}{1+e^{-(\beta_0+\beta_1 x)}}$$

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

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

Where $\beta_0 = -13.5$ and $\beta_1 = 0.06$

Logistic Regression
- Bivariate Analysis



INTERPRETING $\frac{P}{1 - P}$

$$\text{In}\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 x$$

$$\frac{P}{1 - P} = \text{Odds}$$

$$\text{In}\left(\frac{P}{1 - P}\right) = \text{Log Odds}$$

$$\frac{P}{1 - P} = 4$$

P(Diabetes) = 4*P(No Diabetes)

Logistic Regression
- Multivariate Analysis

Data Preparation & Cleaning:

A Note on Feature Scaling Methods
Note that feature standardization is a type of **feature rescaling method,** the other common methods being normalization, rescaling etc.

The most commonly used are standardisation and normalisation, and here's the difference between them (people often get confused between the two):

Standardisation: $x=x-mean(x)sd(x)$

(Mean) Normalisation: $x=x-mean(x)max(x)-min(x)$

Logistic Regression
- Multivariate Analysis

Recall that, for **continuous variables**, We scaled the variables to **standardise** the three continuous variables — tenure, monthly charges and total charges. What the scale command basically does is — it converts values to the z-scores.

For example, let's say that, for a particular customer, tenure = 72. After standardizing, the value of scaled tenure becomes 72−32.424.6=1.61, because for the variable tenure, mean(μ)= 32.4 and standard deviation(σ)= 24.6.

The variables had these ranges before standardisation:
Tenure = 1 to 72
Monthly charges = 18.25 to 118.80
Total charges = 18.8 to 8685

After standardisation, the ranges of the variables changed to:
Tenure = -1.28 to +1.61
Monthly charges = -1.55 to +1.79
Total charges = -0.99 to 2.83

Clearly, none of the variables will have a disproportionate effect on the model's results now.

Logistic Regression
- Multivariate Analysis

Model Building:

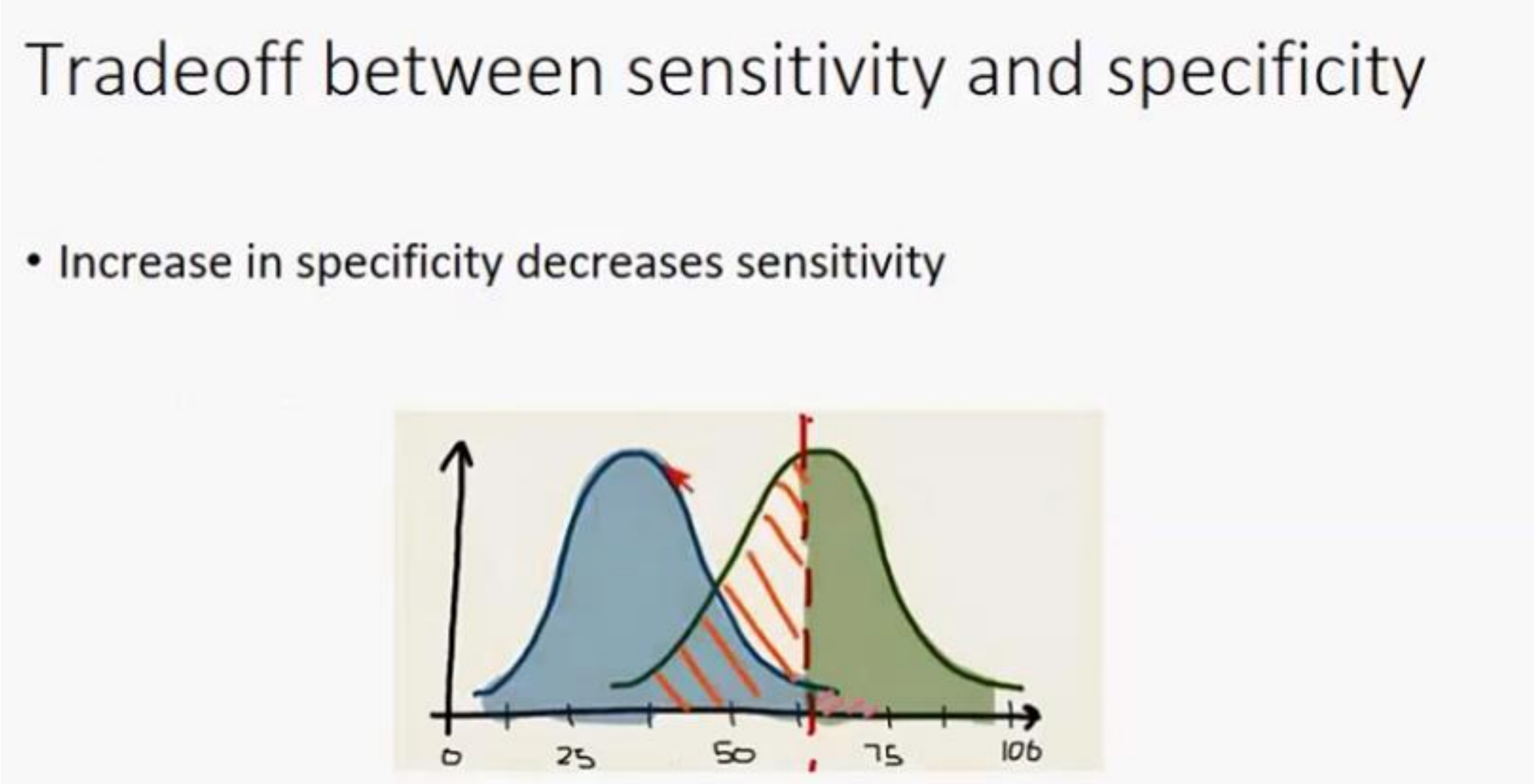
To build the logistic regression model in python, we will use two libraries - **statsmodels** and **sklearn**. In statsmodels, displaying the statistical summary of the model is easier, such as the significance of coefficients (p-value), the coefficients themselves, etc., which is not so straightforward in sklearn.

On the other hand, sklearn comes with some really useful methods such as **recursive feature elimination (RFE).** Thus, we'll use a combination of both the libraries in the upcoming lectures.

Feature Selection Process :

When you have a large number of predictor variables, such as 20-30, it is useful to use an automated feature selection technique such as RFE to reduce the number of variables to a smaller number (say 10-12) and then manually eliminate a few more.

Logistic Regression
- Multivariate Analysis



Tradeoff between sensitivity and specificity

- Increase in specificity decreases sensitivity

Logistic Regression
- Multivariate Analysis

Sensitivity and specificity measure what is called the **discriminatory power** of a model, since they measure how well a model is able to discriminate between two (or more) classes. For example, if the sensitivity is very low, say 10%, it implies that many positives are being predicted as negatives. In other words, the model is not able to discriminate well between these classes.

In general, for a classification model, model evaluation metrics measure one of these two powers of a model:

Accuracy
Discriminatory power

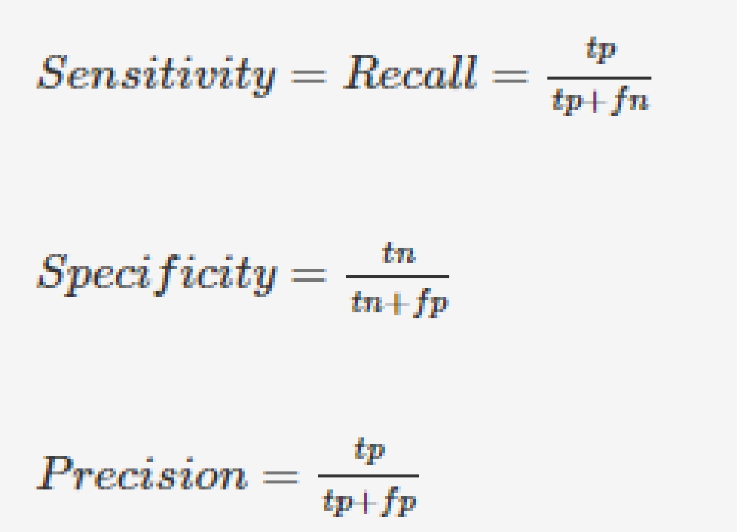
Logistic Regression
- Multivariate Analysis

Precision, Recall and F1-Score
There are other model evaluation metrics as well. Three extremely common ones you will come across are **precision, recall and the F1-score (or F-measure).**

Precision is the fraction of correctly predicted positives out of all *predicted* positives, i.e. it measures 'out of all those the model has predicted to be positive, how many are correct'. It differs from sensitivity only in the denominator.

Recall is the same as sensitivity, and is commonly used along with precision.

To summarise, the expressions for various measures of discriminatory power are as follows:



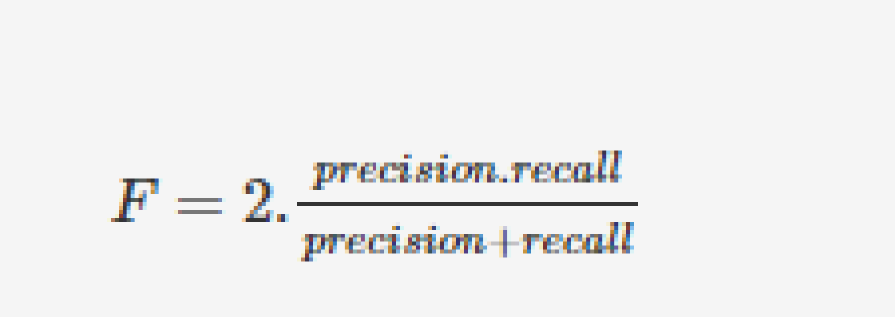
$$Sensitivity = Recall = \frac{tp}{tp + fn}$$

$$Specificity = \frac{tn}{tn + fp}$$

$$Precision = \frac{tp}{tp + fp}$$

Logistic Regression
- Multivariate Analysis

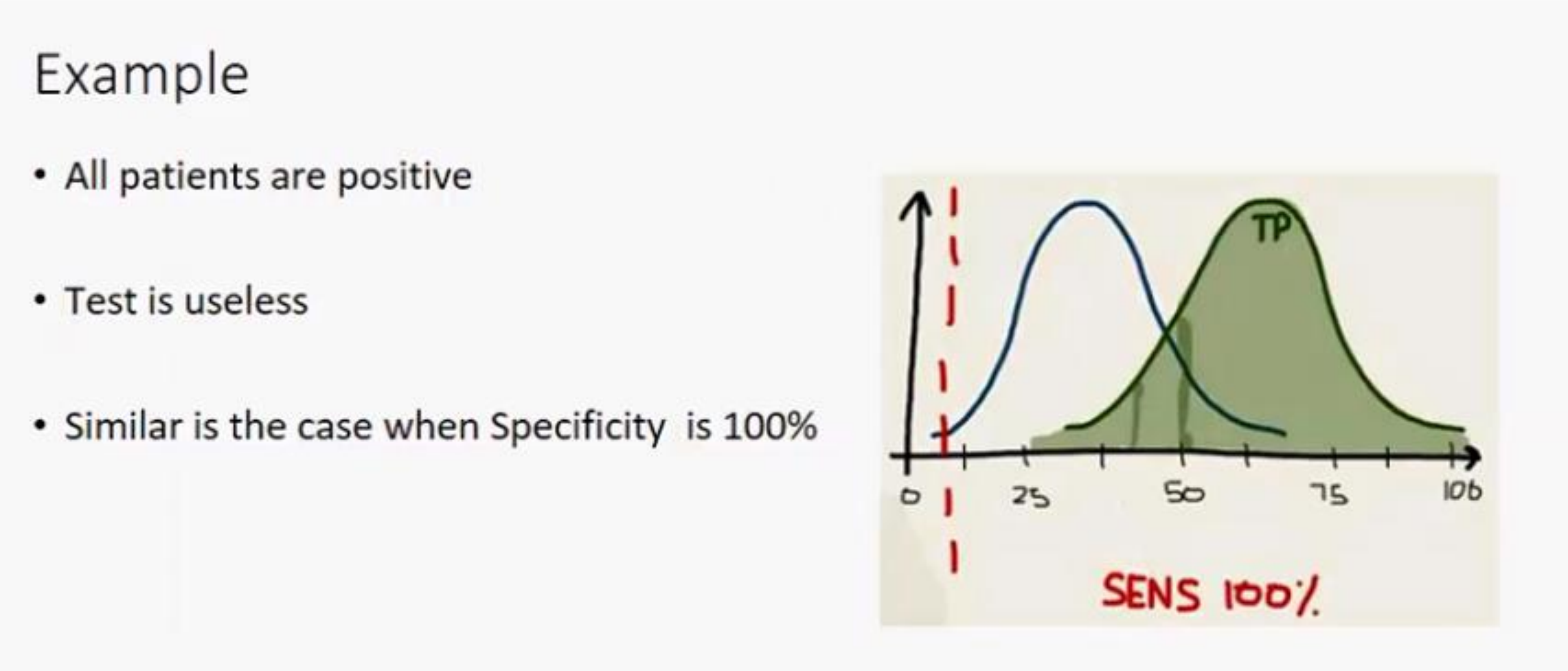
F1-score is a measure that combines both precision and recall. It is the harmonic mean of precision and recall:



$$F = 2.\frac{precision.recall}{precision+recall}$$

Now, you often need to choose a metric which you want to maximise, as you saw on the previous page. This is called **tuning of the model.**

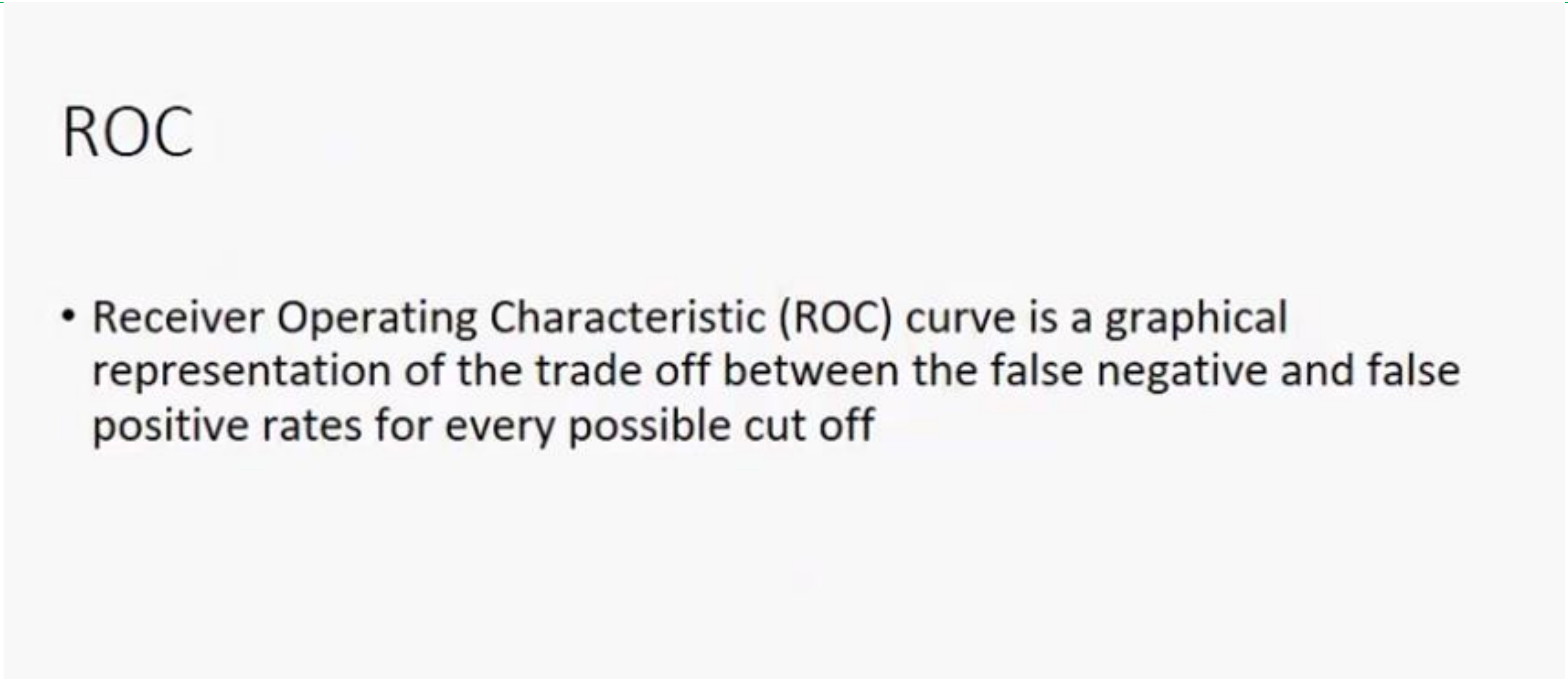
Logistic Regression



Example

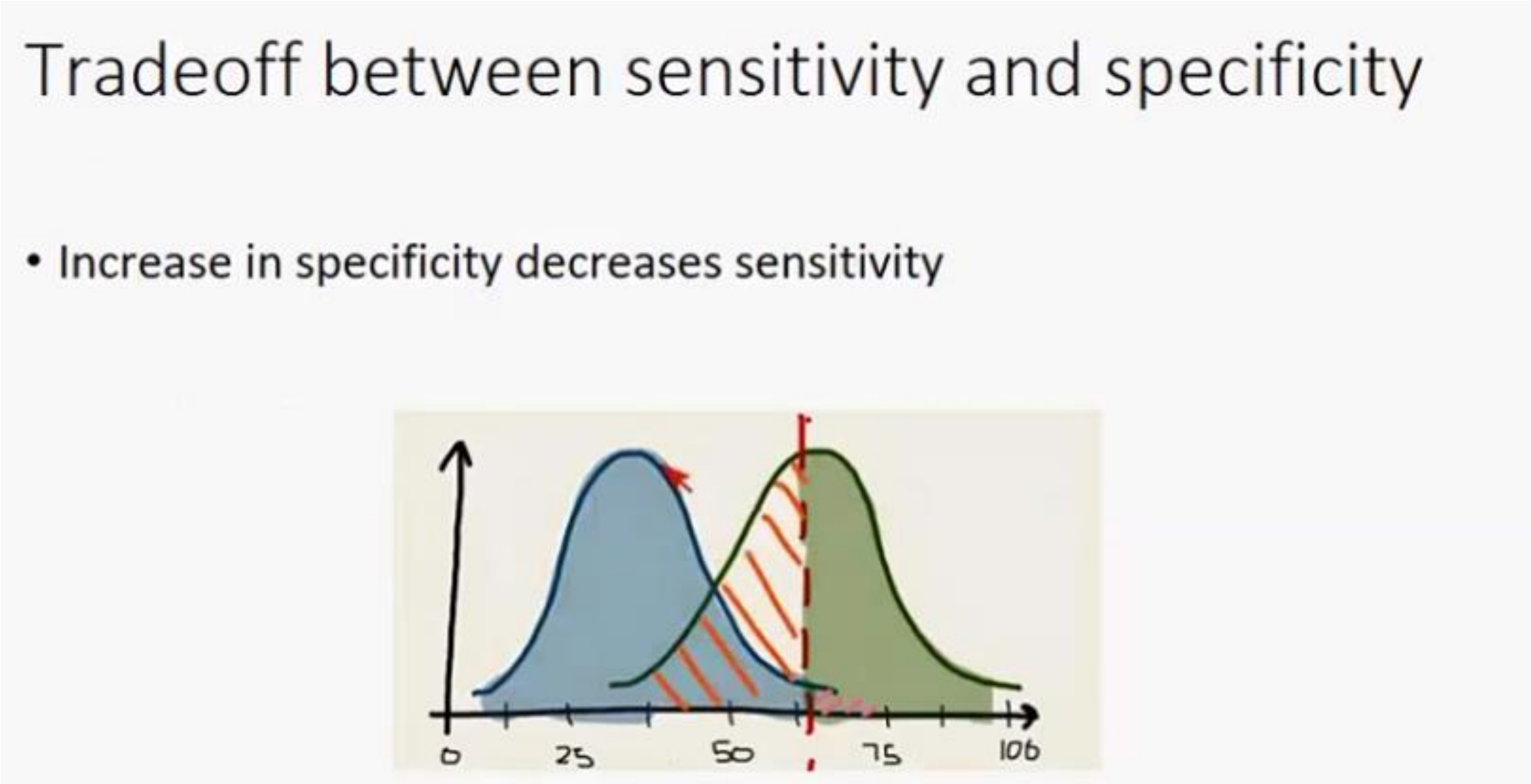
- All patients are positive
- Test is useless
- Similar is the case when Specificity is 100%

Logistic Regression



ROC

- Receiver Operating Characteristic (ROC) curve is a graphical representation of the trade off between the false negative and false positive rates for every possible cut off



Tradeoff between sensitivity and specificity

- Increase in specificity decreases sensitivity

Logistic Regression
- Types of LR

In general, logistic regression by definition tries to predict what state a particular individual or system will be in the future. You learnt about the two **types of logistic regression**:
Binary logit
Multinomial logit

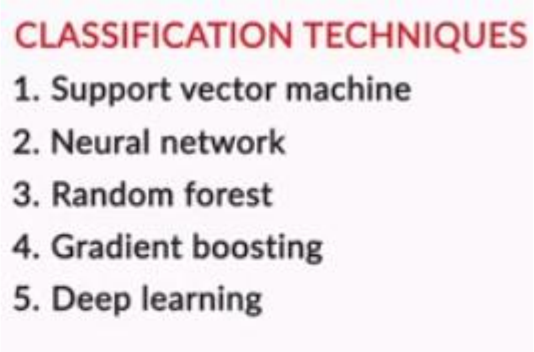
Binary logit involves two levels of the dependent variable. For example, the telecom churn example you learnt in earlier sessions is a binary logistic regression problem, as it classifies customers into two levels, churns and non-churns. **Multinomial logit**, however, involves more than 2 levels of dependent variables, such as whether a customer will purchase product A, product B or not purchase anything.

So, the rule of thumb for deciding whether the problem is a binary classification problem or multinomial classification problem is that you should first understand the dependent variable.

Multinomial logit is typically solved by plotting multiple Binary logit models

The two main important differences between logistic and linear regression are: 1. Dependent/response variable in linear regression is continuous whereas, in logistic regression, it is the discrete type. 2. Cost function in linear regression minimise the error term Sum(Actual(Y)-Predicted(Y))^2 but logistic regression uses maximum likelihood method for maximising probabilities.

Logistic Regression
- Types of LR



CLASSIFICATION TECHNIQUES

1. Support vector machine
2. Neural network
3. Random forest
4. Gradient boosting
5. Deep learning

There may be multiple classification techniques as stated above. But logistic regression is a widely used technique in various types of industries. This is because of two **main reasons**:

1) It is very easy to **understand** and offers an **intuitive explanation** of the variables
2) The output (i.e. the probabilities) has a linear relationship with the log of odds, which can be very useful for explaining results to managers

Logistic Regression
- Sample selection

Selecting the right sample is essential for solving any business problem. As discussed in the lecture, there are major errors you should be on the lookout for while selecting a sample. These include:

Cyclical or **seasonal fluctuations** in the business that need to be taken care of while building the samples. E.g. Diwali sales, economic ups and downs, etc. so that all the behaviors in the model are covered

The sample should be **representative of the population** on which the model will be applied in the future.
For **rare events samples**, the sample should be balanced before it is used for modelling
Incase of credit card data , bank collecting salaried for till three years back. From last 3 years they are selecting students data strategically. Based on your model you select which data whether student or salaries or both should be considered. Model also should consider representative population and rare incident population

Logistic Regression
- Data Preparation

1. Missing values treatment(Remove or impute)
2. Column significance
3. Outlier treatment
4. Category to numerical
5. Category to dummy variables (no of dummy features/columns = no of categories-1)
6. Mathematical transformations (square, cube, log)
7. Standardization/normalization
8. Column significance using RFE(recursive feature elimination), P-Value, VIF(Variance Inflation factor). Another excellent way is PCA which ensures low multicollinearity