Logistic Regression

- Sonal Ghanshani

When to use?

How to model Qualitative Response Variable?

- Example
  - Labour force participation (Yes=1, no=0) depends on unemployment rate, average wage rate, education, family income etc.
  - US presidential elections: Vote Democratic candidate (=1), vote Republican candidate (=0) depends on rate of GDP growth, unemployment, whether a candidate runs for re-election (a dummy)
  - Onset of heart disease depends on age, exercise (yes/no), smoking (yes/no)
- All the response variables are qualitative in nature.

Two critical questions

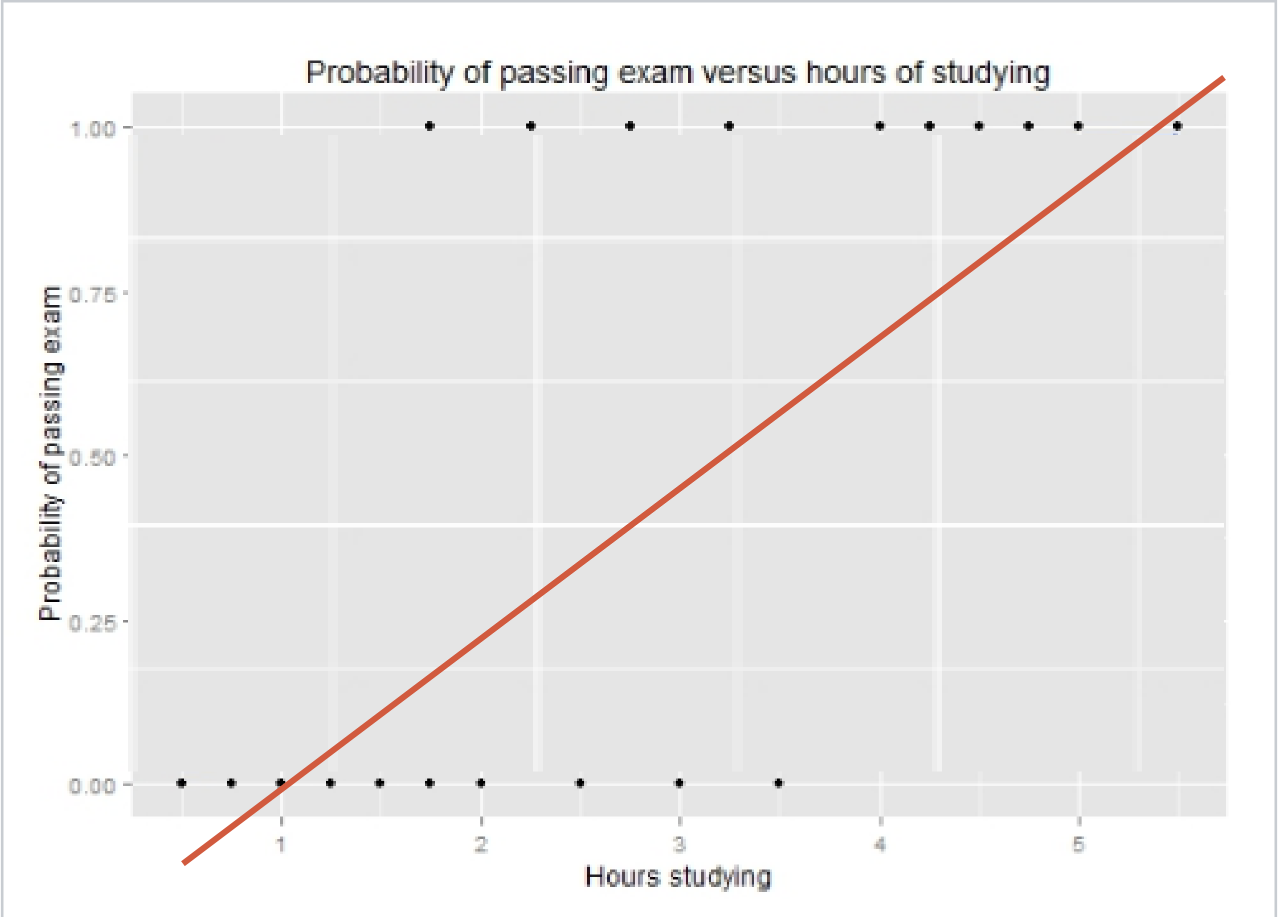
- Can we run a usual linear regression and interpret the outcome?
- Since the response variable is qualitative in nature, what do you predict in this case?

Introduction to Logistic Regression

Logistic Regression

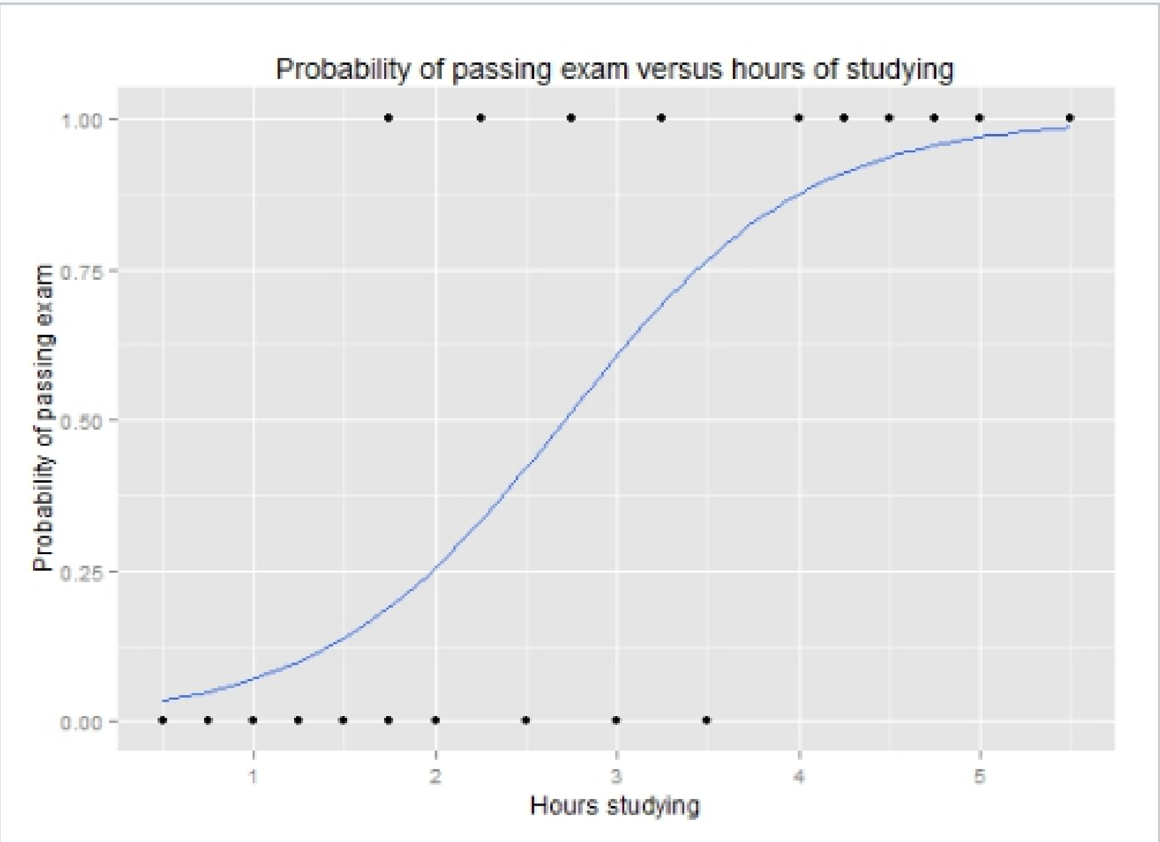
- This method is widely used for binary classification problems. It can also be extended to multi-class classification problems.
- Here, the dependent variable is categorical: $y \in \{0, 1\}$.
- A binary dependent variable can have only two values, like 0 or 1, win or lose, pass or fail, healthy or sick, etc.

Logistic Regression

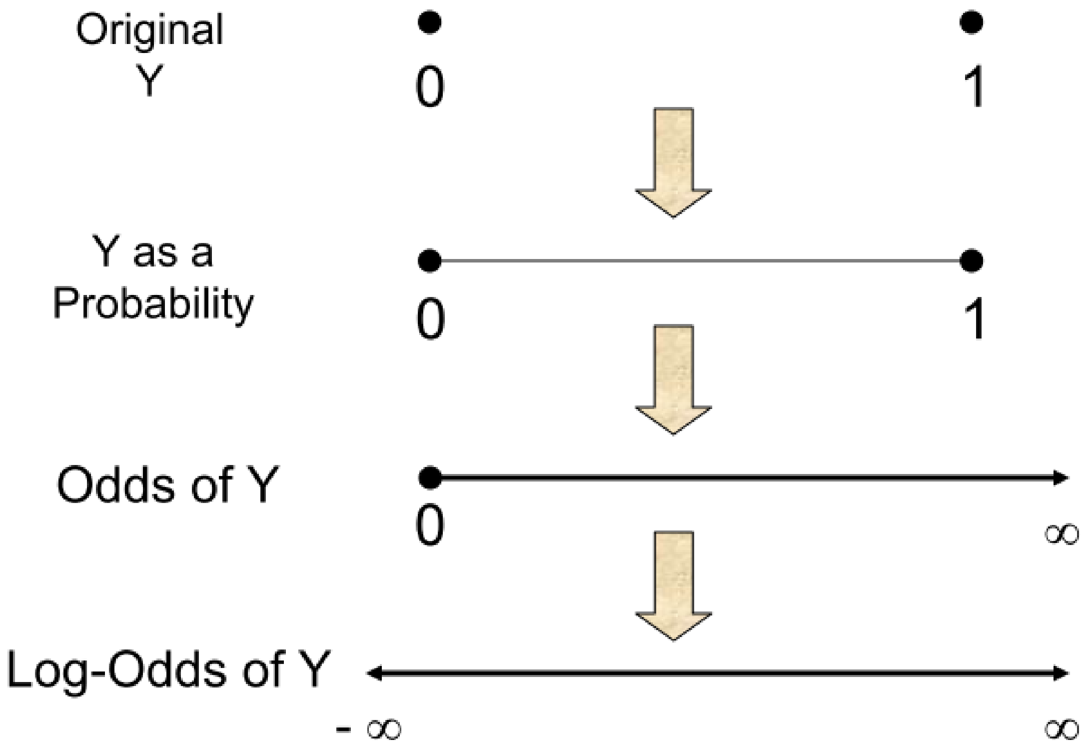


Graph of a logistic regression curve showing probability of passing an exam versus hours studying

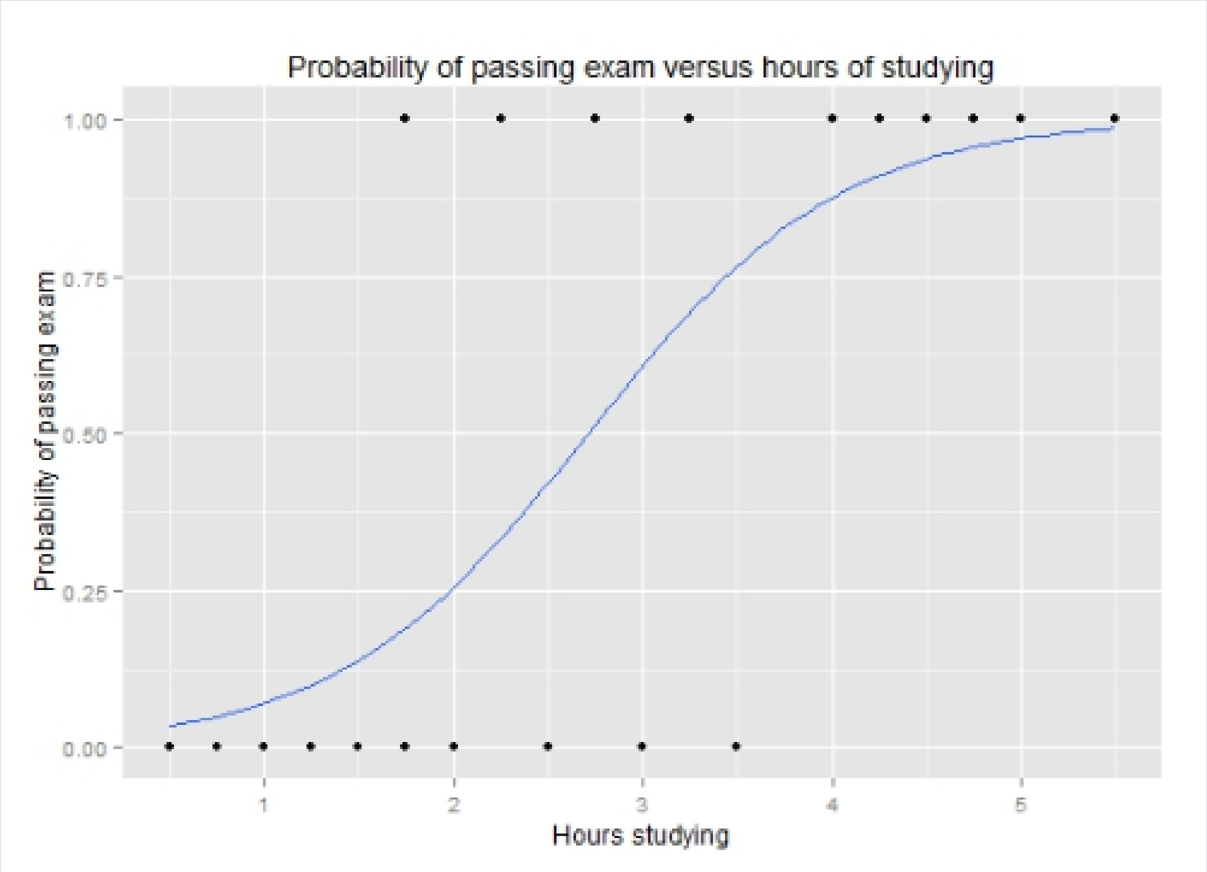
Logistic Regression



Graph of a logistic regression curve showing probability of passing an exam versus hours studying



Logistic Regression



Graph of a logistic regression curve showing probability of passing an exam versus hours studying

- The probability in the logistic regression is often represented by the Sigmoid function (also called the logistic function or the S-curve):

$$S(t) = \frac{1}{1 + e^{-t}}$$

- In this equation, t represents data values * number of hours studied and S(t) represents the probability of passing the exam.
- The points lying on the sigmoid function fits are either classified as positive or negative cases. A threshold is decided for classifying the cases.

Logistic Regression

$$P("Success"|X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

- Here, Success is P(Y = 1 | X)
- This method is widely used for binary classification problems. It can also be extended to multi-class classification problems.
- Here, the dependent variable is categorical: y ϵ {0, 1}.
- A binary dependent variable can have only two values, like 0 or 1, win or lose, pass or fail, healthy or sick, etc.

Model Validation

Confusion Matrix

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

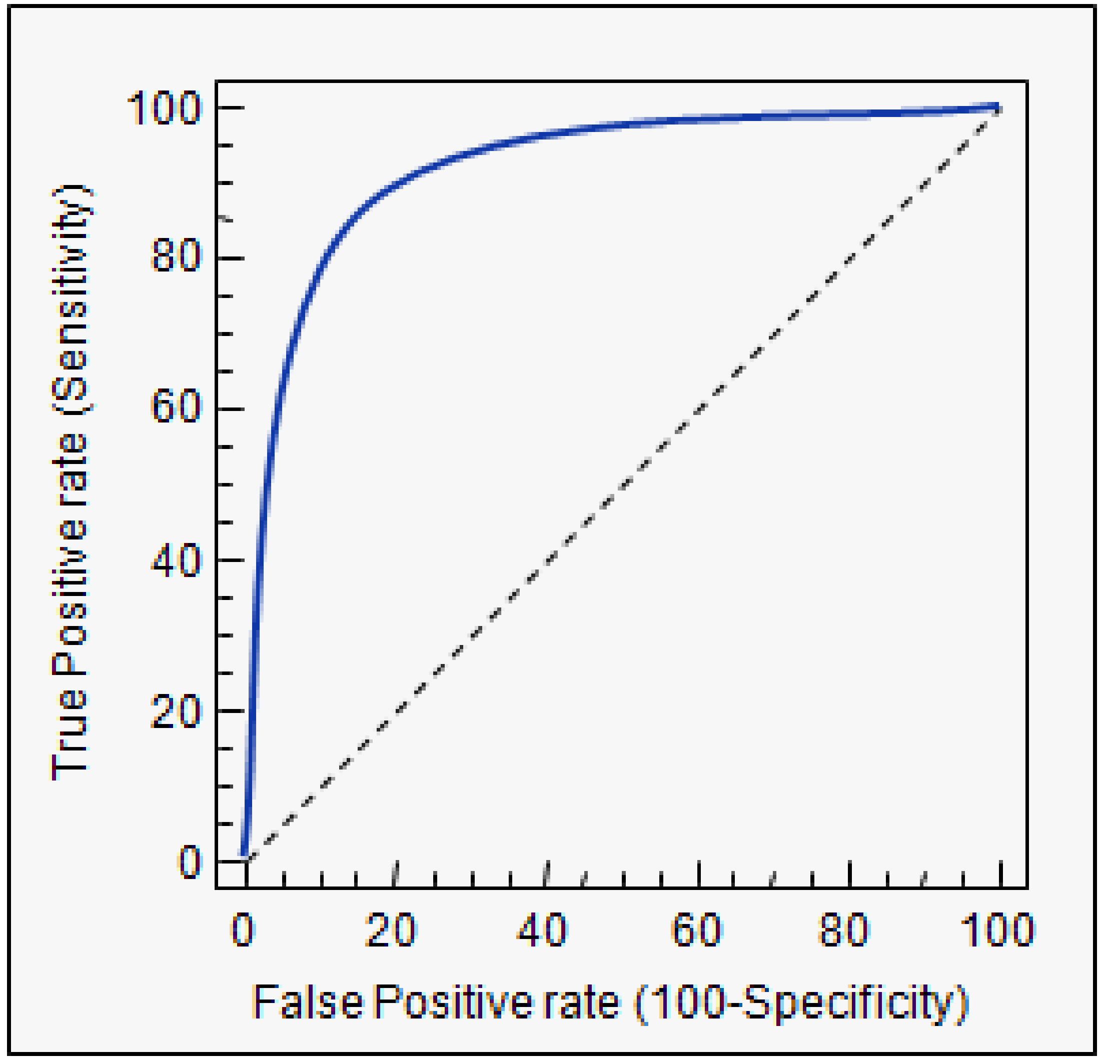
Sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Specificity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

ROC - AUC


100
80
60
40
20
0
True Positive rate (Sensitivity)
0
20
40
60
80
100
False Positive rate (100-Specificity)