Introduction to Data Analytics

ITE 5201

Lecture8-Logistic Regression

Instructor: Parisa Pouladzadeh
Email: parisa.pouladzadeh@humber.ca

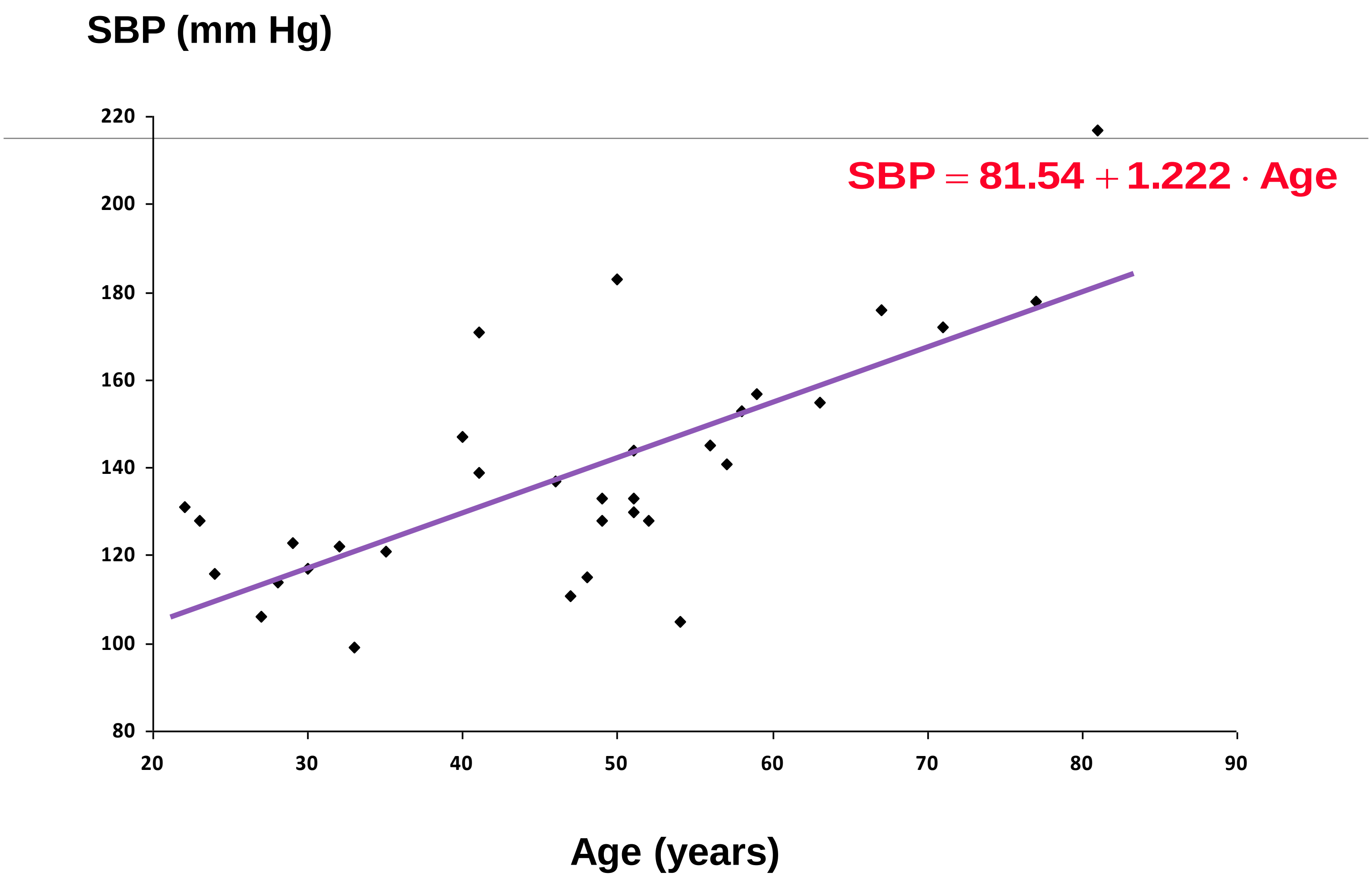
www.udemy. /course/python-for-data-science-and-machine-learning.com

HUMBER

Simple linear regression

Table 1 Age and systolic blood pressure (SBP) among 33 adult women

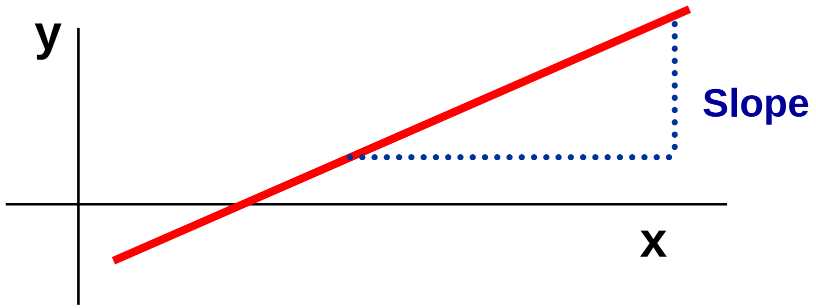
Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217



adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

Simple linear regression

Relation between 2 continuous variables (SBP and age)



$$y = \alpha + \beta_1 x_1$$

Regression coefficient β_1

- Measures association between y and x
- Amount by which y changes on average when x changes by one unit
- Least squares method

HUMBER

Logistic regression

Table 2 Age and signs of coronary heart disease (CD)

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

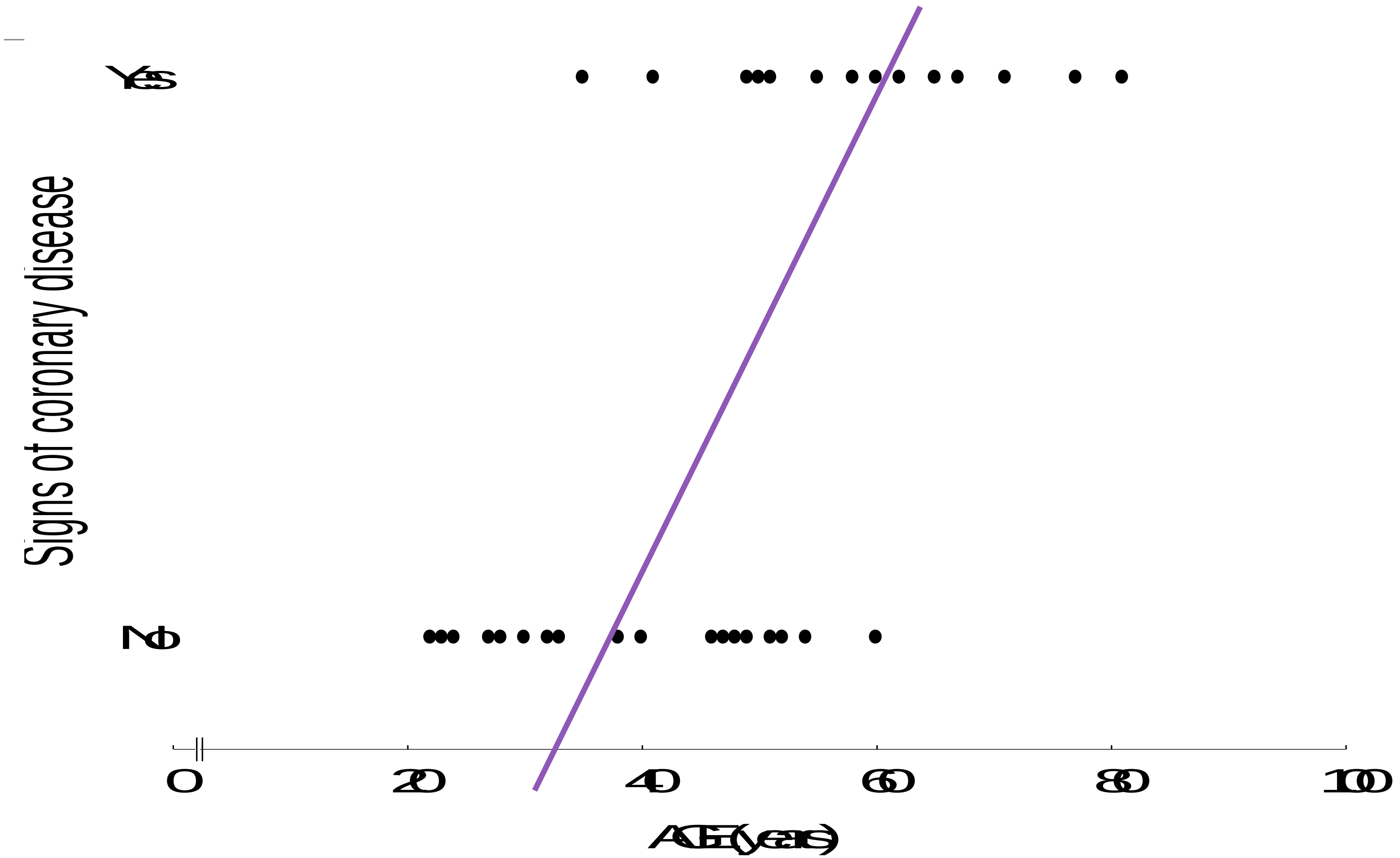
How can we analyse these data?

Compare mean age of diseased and non-diseased

- Non-diseased: 38.6 years
- Diseased: 58.7 years ($p<0.0001$)

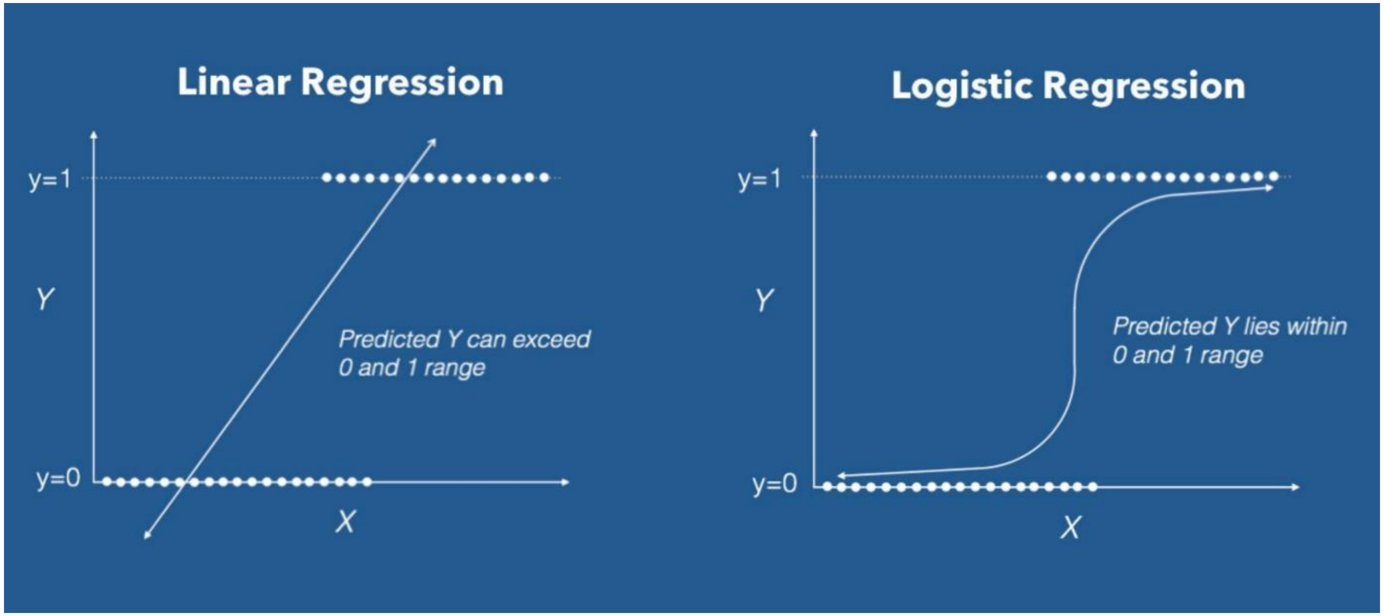
Linear regression?

Dot-plot: Data from Table 2



Logistic Regression

- Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.



HUMBER

Logistic Regression

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function.

The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

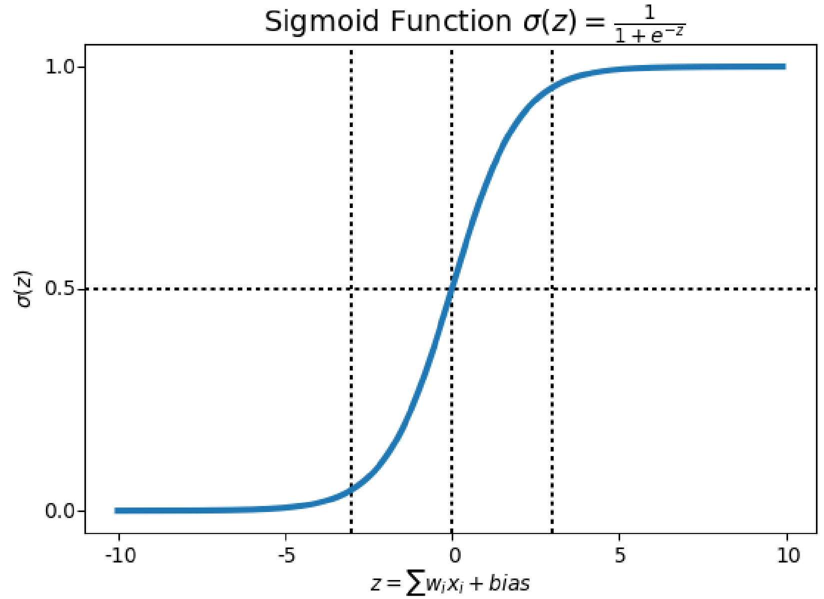
Logistic regression hypothesis expectation

What is the Sigmoid Function?

What is the Sigmoid Function?

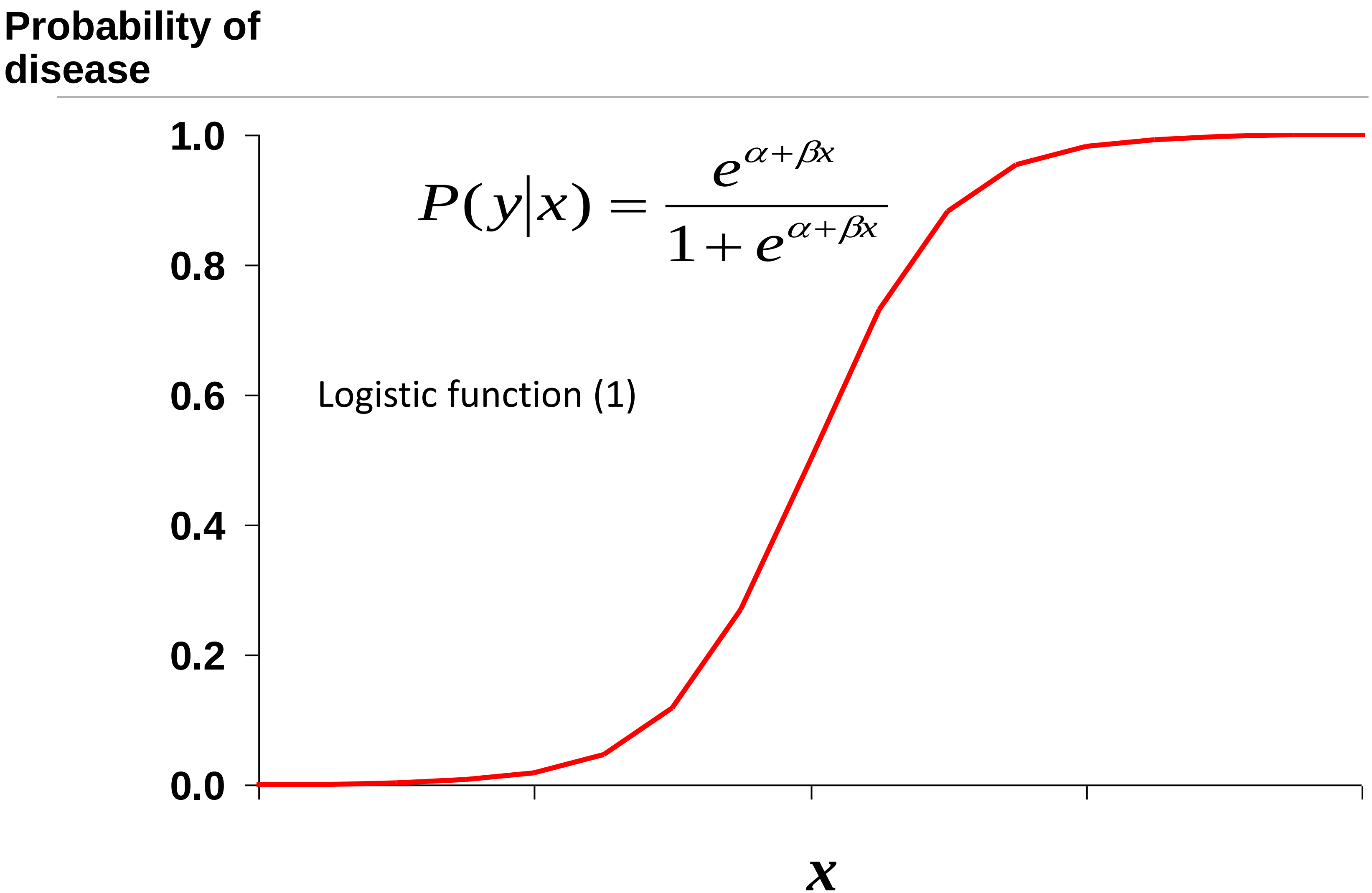
In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning,
probabilities.

$$f(x) = \frac{1}{1+e^{-(x)}}$$



HUMBER

Logistic function



Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure.

We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.

Here the value of Y ranges from 0 to 1 and it can represented by following equation.

p is the probability of presence of the characteristic of interest.

```
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk
```

Transformation

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\frac{P(y|x)}{1-P(y|x)}$$

$$\underbrace{\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]}_{\text{logit of } P(y|x)} = \alpha + \beta x$$

- ✓ **α = log odds of disease in unexposed**
- ✓ **β = log odds ratio associated with being exposed**
- ✓ **e^{β} = odds ratio**

Fitting equation to the data

- Linear regression: Least squares
- Logistic regression: Maximum likelihood
- Likelihood function
  - Estimates parameters a and b
  - Practically easier to work with log-likelihood

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i)\ln\left[1 - \pi(x_i)\right] \right\}$$

Maximum likelihood

- **Iterative computing**
  - **Choice of an arbitrary value for the coefficients (usually 0)**
  - **Computing of log-likelihood**
  - **Variation of coefficients' values**
  - **Reiteration until maximisation (plateau)**

- **Results**
  - **Maximum Likelihood Estimates (MLE) for α and β**
  - **Estimates of P(y) for a given value of x**

Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure.

We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.

Here the value of Y ranges from 0 to 1 and it can represented by following equation.

p is the probability of presence of the characteristic of interest.

```
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk
```

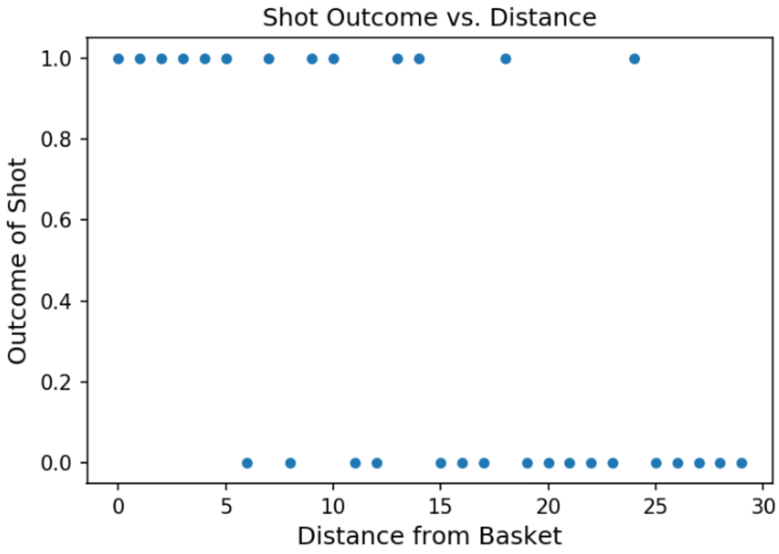
Ex. Shooting Basket

Let's say I wanted to examine the relationship between my basketball shooting accuracy and the distance that I shoot from.

More specifically, I want a model that takes in "distance from the basket" in feet and spits out the probability that I will make the shot.

Ex. Shooting Basket

First I need some data. So I went out and shot a basketball from various distances while recording each result (1 for a make, 0 for a miss). The result looks like this when plotted on a scatter plot:



Ex. Shooting Basket

Generally, the further I get from the basket, the less accurately I shoot.

So we can already see the rough outlines of our model:

- when given a small distance, it should predict a high probability and when given a large distance it should predict a low probability.

- So let's start with the familiar linear regression equation:
  - Y = B0 + B1*X
- In linear regression, the output Y is in the same units as the target variable (the thing you are trying to predict).

Ex. Shooting Basket

However, in logistic regression the output Y is in log odds.

Odds is just another way of expressing the probability of an event, P(Event).

Odds = P(Event) / [1-P(Event)]

Continuing our basketball theme, let's say I shot 100 free throws and made 70.

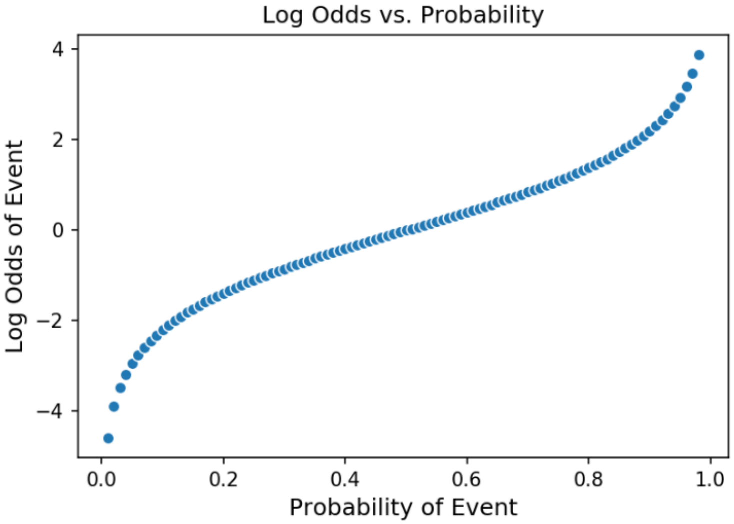
Based on this sample, my probability of making a free throw is 70%. My odds of making a free throw can be calculated as:

Odds = 0.70 / (1–0.70) = 2.333

Ex. Shooting Basket

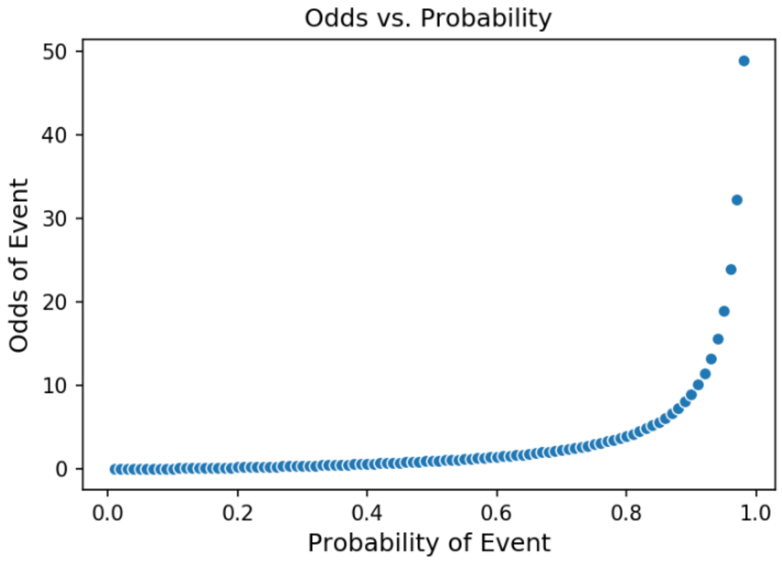
And if we take the natural log of the odds, then we get log odds which are unbounded (ranges from negative to positive infinity) and roughly linear across most probabilities!

Since we can estimate the log odds via logistic regression, we can estimate probability as well because log odds are just probability stated another way.



Ex. Shooting Basket

Probabilities are bounded between 0 and 1, which becomes a problem in regression analysis. Odds as you can see below range from 0 to infinity.



HUMBER

Ex. Shooting Basket

We can write our logistic regression equation:

- Z = B0 + B1*distance_from_basket
- where Z = log(odds_of_making_shot)