

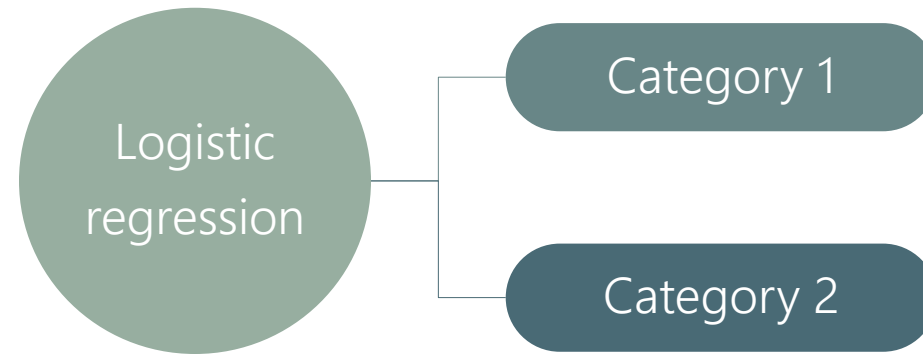
# **COURSE NOTES: LOGISTIC REGRESSION**

# Logistic regression vs Linear regression

Logistic regression implies that the possible outcomes are **not** numerical but rather categorical.

Examples for categories are:

- Yes / No
- Will buy / Won't Buy
- 1 / 0

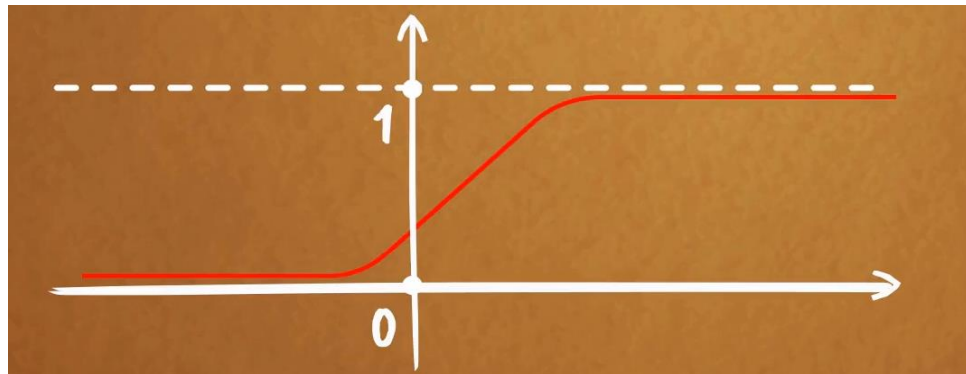


Linear regression model:  $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$

Logistic regression model:  $p(X) = \frac{e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$

# Logistic model

The logistic regression predicts the probability of an event occurring.



Visual representation of a logistic function

# Logistic regression model

## Logistic regression model

$$\frac{p(X)}{1-p(X)} = e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}$$

The logistic regression model is not very useful in itself. The right-hand side of the model is an exponent which is very computationally inefficient and generally hard to grasp.

## Logit regression model

When we talk about a 'logistic regression' what we usually mean is 'logit' regression – a variation of the model where we have taken the log of both sides.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \log(e^{(\beta_0 + \beta_1 x + \dots + \beta_k x_k)})$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 x + \dots + \beta_k x_k$$

$$\log(\text{odds}) = \beta_0 + \beta_1 x + \dots + \beta_k x_k$$

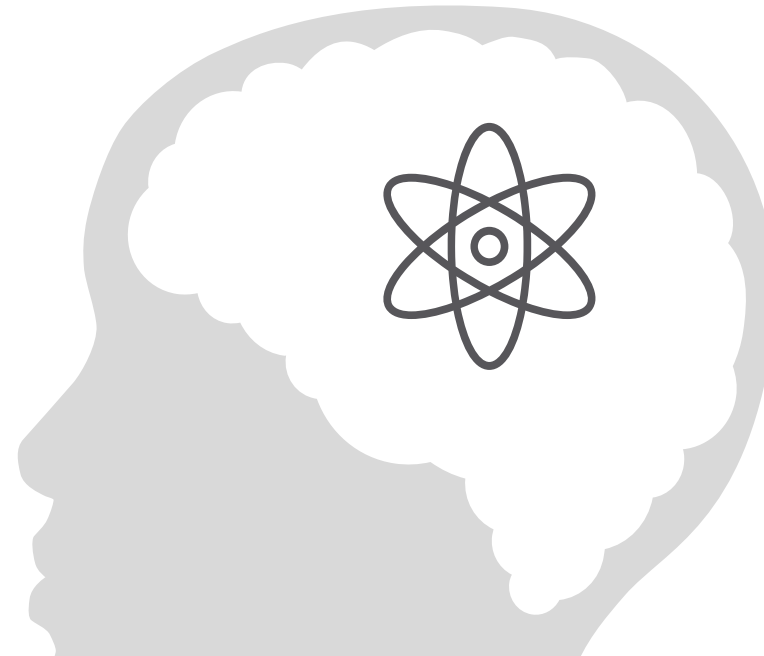
$$\text{ODDS} = \frac{p(X)}{1-p(X)}$$

### Coin flip odds:

The odds of getting heads are 1:1 (or simply 1)

### Fair die odds:

The odds of getting 4 are 1:5 (1 to 5)



# Logistic regression model

The dependent variable,  $y$ ;  
This is the variable we are  
trying to predict.

Indicates whether our  
model found a solution or  
not.

Coefficient of the  
intercept,  $b_0$ ; sometimes  
we refer to this variable as  
constant or bias.

Coefficient of the independent variable  $i$ :  $b_i$ ; this is usually the most important metric – it shows us the relative/absolute contribution of each independent variable of our model. For a logistic regression, the coefficient contributes to the log odds and cannot be interpreted directly.

<b>Dep. Variable:</b>	y	<b>No. Observations:</b>	518
<b>Model:</b>	Logit	<b>Df Residuals:</b>	516
<b>Method:</b>	MLE	<b>Df Model:</b>	1
<b>Date:</b>	Thu, 28 Nov 2019	<b>Pseudo R-squ.:</b>	0.2121
<b>Time:</b>	15:01:00	<b>Log-Likelihood:</b>	-282.89
<b>converged:</b>	True	<b>LL-Null:</b>	-359.05
		<b>LLR p-value:</b>	5.387e-35
	<b>coef</b>	<b>std err</b>	<b>z</b> <b>P&gt; z </b> <b>[0.025</b> <b>0.975]</b>
<b>const</b>	-1.7001	0.192	-8.863 0.000 -2.076 -1.324
<b>duration</b>	0.0051	0.001	9.159 0.000 0.004 0.006

McFadden's pseudo-R-squared, used for comparing variations of the same model. Favorable range [0.2,0.4].

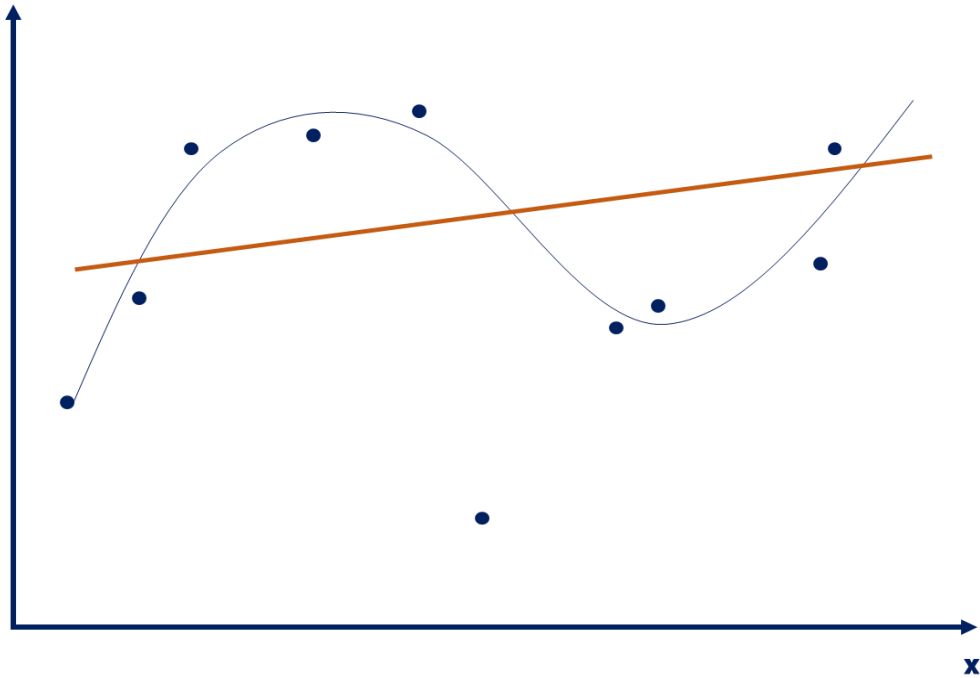
Log-Likelihood\* (the log of the likelihood function). Always negative. We aim for this to be as high as possible.

Log-Likelihood-Null is the log-likelihood of a model which has no independent variables. It is used as the benchmark 'worst' model.

Log-Likelihood Ratio p-value measures of our model is statistically different from the benchmark 'worst' model.

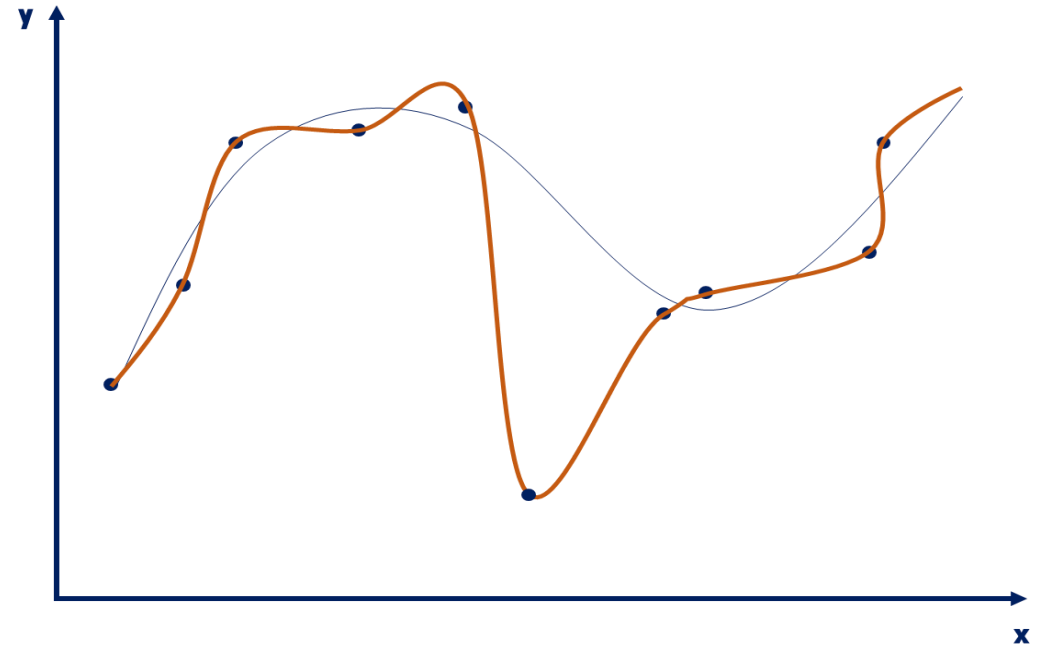
\*Likelihood function: a function which measures the goodness of fit of a statistical model.  
MLE (Maximum Likelihood Estimation) tries to maximize the likelihood function.

# Underfitting



The model has not captured the underlying logic of the data.

# Overfitting



Our training has focused on the particular training set so much it has "missed the point".