Classification

- Supervised learning - there is a chosen response variable
- Individuals or items belong to one and only one of several possible classes, groups or categories
- **Aim to predict** which class a new individual belongs to using a classification rule or model
- The model uses one or more predictor variables, called feature variables

Classification

- Fitting the classification rule is called <u>training</u>
- Training requires <u>labelled data</u>
  - data available on the feature variables for a set of items already classified
- The model is fitted on this training data
- Assessment of classifier performance usually uses <u>error rates</u> or <u>correct classification rates</u>
  - how well does it classify new cases?

Logistic Regression Model

Links one or more explanatory variables, *X*, to a binary response variable *Y*

$Y \sim Binomial(N, \pi)$

Y is number of positive responses out of *N* independent trials where for each trial:

P(success) = $P(1) = \pi$

P(fail) = $P(0) = 1 - \pi$

Linear model linking the log odds (LHS) to the linear predictor (RHS):

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

the logit of π

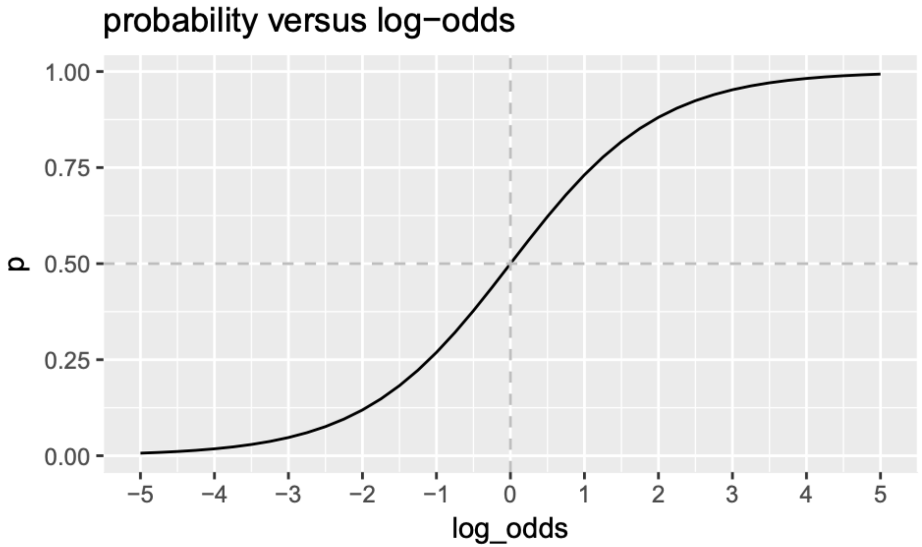
Intercept,
Log odds if X=0

Slope coefficient,
Change in log odds for a unit change in X

Why use log (odds) in model?

Probabilities P(E) lie between 0 and 1

- Odds P(E)/(1-P(E)) defined between 0 and ∞
- Using natural logs, <u>ln(odds)</u> defined between -∞ and ∞
- Able to predict log odds of event E using $\beta_0 + \beta_1 X$



Binary X variable

$$Logit[\pi(x)] = Ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1(x)$$

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

Odds for X=0

$$\frac{\pi(0)}{1-\pi(0)} = e^{\beta_0+\beta_1(0)} = e^{\beta_0}$$

Odds for X=1

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

Binary X variable

Odds for X=0

$$\frac{\pi(0)}{1-\pi(0)} = e^{\beta_0+\beta_1(0)} = e^{\beta_0}$$

Odds for X=1

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

$$\left(\frac{\pi(1)}{1-\pi(1)}\right) \Big/ \left(\frac{\pi(0)}{1-\pi(0)}\right) = e^{\beta_0+\beta_1} / e^{\beta_0} = e^{\beta_1}$$

- So taking exponential of coefficient β_1 gives the <u>odds ratio</u>

Another formulation

Equivalently we are modelling the probability P(E) or π :

$$\pi(X) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}}$$

- This is the **predicted probability of a positive response** at a value of the explanatory variable X

Binary Classification

- Using a model to predict which one of two groups an individual belongs to

- The model here gives the **predicted probability of event** of interest as

$$\pi(X) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}}$$

- Model can be used for <u>classification</u> by using a cut-off value for the probability or for the linear predictor $\beta_0 + \beta_1 X$
  - the event is unlikely if the predicted probability is low

Logistic regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. Logistic Regression is used to find the probability of a certain class or event existing such as Pass/Fail, True/False, or Alive/Dead or the probability of a certain event occurring.


A GOOD EXAMPLE OF
OVERFITTING