LOGISTIC REGRESSION

Logistic regression just like Linear Regression is a generalized linear model (GLM) procedure using the same basic formula of Linear Regression, but instead of the continuous Y, it is regressing for the probability of a categorical outcome. In simplest form, this means that we're considering just one outcome variable and two states of that variable- either 0 or 1. In Linear Regression the output is a continuous variable however in Logistic Regression it is a probability.

Note: Once you understand Logistic Regression with Gradient Descent Optimization, understanding Feed forward Back propagation Neural Networks is a cakewalk.

Reference - Wikipedia: https://en.wikipedia.org/wiki/Logistic_regression

Example - Probability of passing an exam versus hours of study

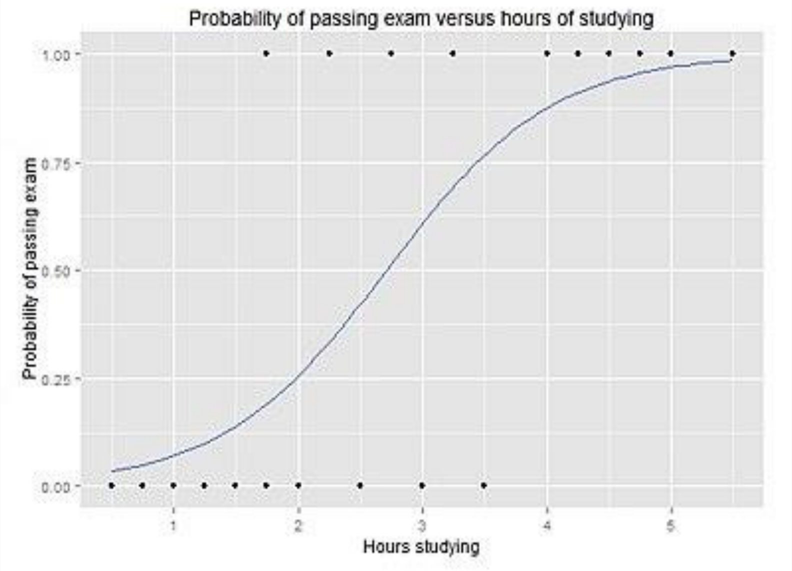
To answer the following question: A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used.

The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

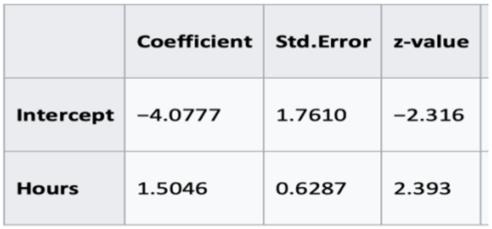
Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.

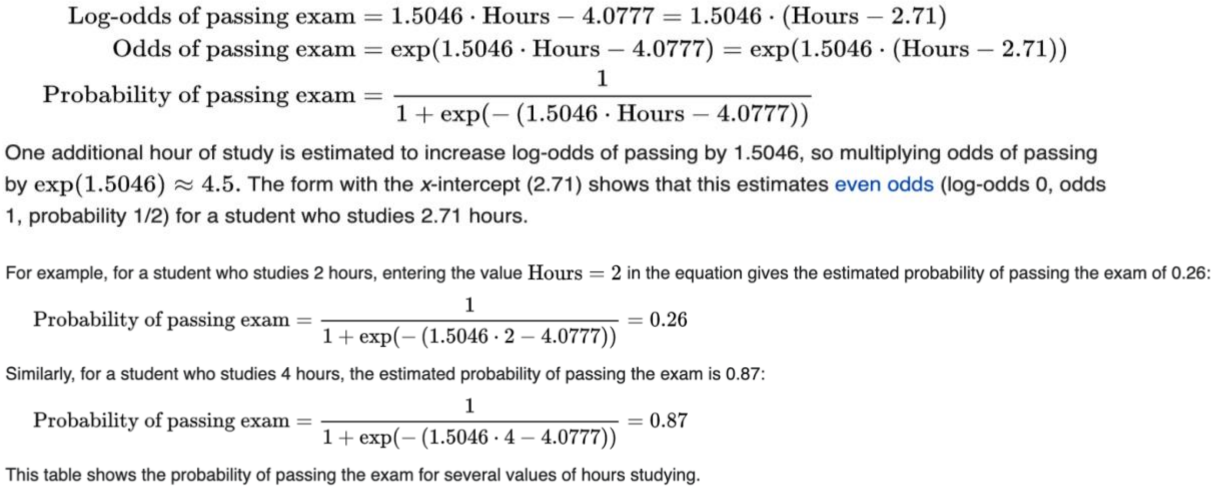


Graph of a logistic regression curve showing probability of passing an exam versus hours studying

The logistic regression analysis gives the following output.



	Coefficient	Std.Error	z-value
Intercept	−4.0777	1.7610	−2.316
Hours	1.5046	0.6287	2.393



$$\begin{aligned}\text{Log-odds of passing exam} &= 1.5046 \cdot \text{Hours} - 4.0777 = 1.5046 \cdot (\text{Hours} - 2.71)\\ \text{Odds of passing exam} &= \exp(1.5046 \cdot \text{Hours} - 4.0777) = \exp(1.5046 \cdot (\text{Hours} - 2.71))\\ \text{Probability of passing exam} &= \frac{1}{1 + \exp(-(1.5046 \cdot \text{Hours} - 4.0777))}\end{aligned}$$

One additional hour of study is estimated to increase log-odds of passing by 1.5046, so multiplying odds of passing by $\exp(1.5046) \approx 4.5$. The form with the *x*-intercept (2.71) shows that this estimates even odds (log-odds 0, odds 1, probability 1/2) for a student who studies 2.71 hours.

For example, for a student who studies 2 hours, entering the value $\text{Hours} = 2$ in the equation gives the estimated probability of passing the exam of 0.26:

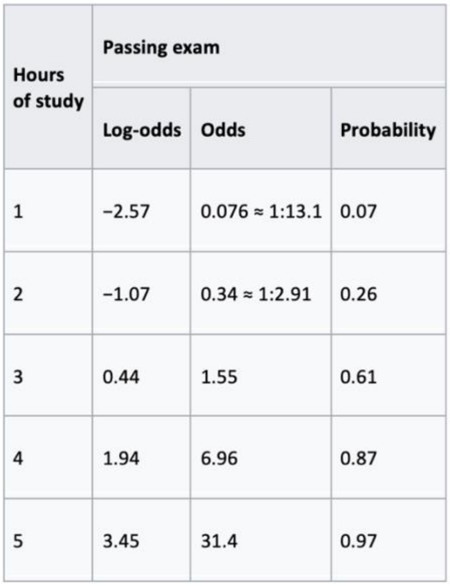
$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 2 - 4.0777))} = 0.26$$

Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

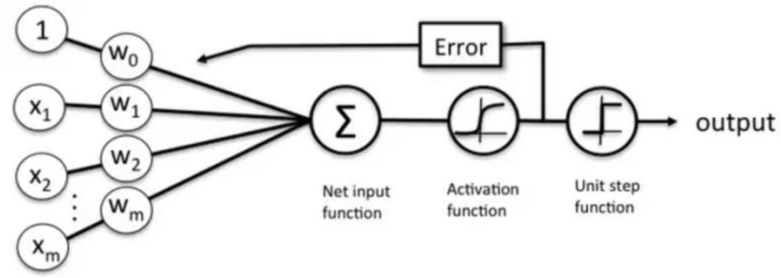
$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 4 - 4.0777))} = 0.87$$

This table shows the probability of passing the exam for several values of hours studying.

This table shows the probability of passing the exam for several values of hours studying.



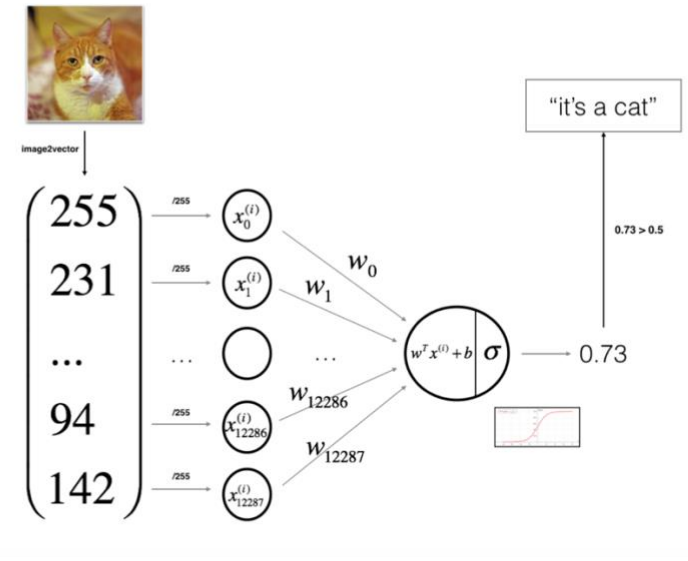
Hours of study	Passing exam		
	Log-odds	Odds	Probability
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97



High level Algorithm (Logistic Regression using Gradient Descent Optimization, very similar to NN’s):

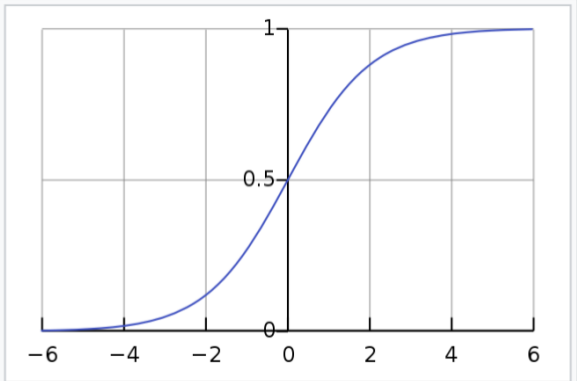
1. Define the model structure (such as number of input features)
2. Initialize the model’s parameters
3. Loop until specified convergence criteria:
    - Pass the sum of x’s (features) and w’s (weights) (refer diagram above) to an activation function
    - Calculate the loss
    - Update the weights

Example:



Note: We want to maximize the likelihood that a random data point gets classified correctly, which is called Maximum Likelihood Estimation. Maximum Likelihood Estimation is a general approach to estimating parameters in statistical models. You can maximize the likelihood using different methods like Newton's method or optimization method like Gradient Descent.

The Logistic Curve and its equation



The logistic curve

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

Reference: A great resource to understand the derivation of the above Logistic equation: https://stats.idre.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/

Let's begin with probability. Probabilities range between 0 and 1. Let's say that the probability of success is .8, thus

p = .8

Then the probability of failure is

q = 1 – p = .2

Odds are determined from probabilities and range between 0 and infinity. Odds are defined as the ratio of the probability of success and the probability of failure. The odds of success are

odds(success) = p/(1-p) or p/q = .8/.2 = 4,

that is, the odds of success are 4 to 1. The odds of failure would be

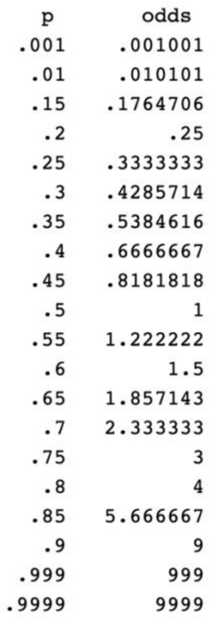
odds(failure) = q/p = .2/.8 = .25.

This looks a little strange but it is really saying that the odds of failure are 1 to 4. The odds of success and the odds of failure are just reciprocals of one another, i.e., 1/4 = .25 and 1/.25 = 4. Next, we will add another variable to the equation so that we can compute an odds ratio.

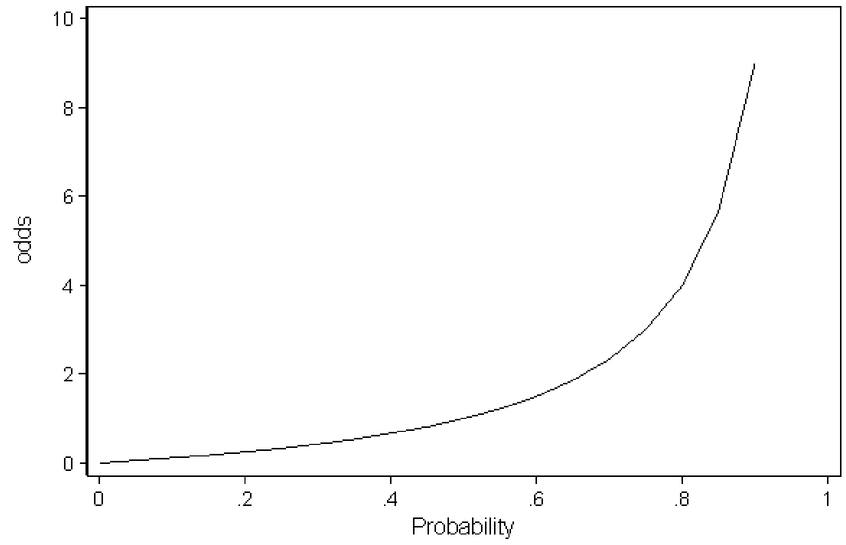
From probability to odds to log of odds

Everything starts with the concept of probability. Let's say that the probability of success of some event is .8. Then the probability of failure is 1 – .8 = .2. The odds of success are defined as the ratio of the probability of success over the probability of failure. In our example, the odds of success are .8/.2 = 4. That is to say that the odds of success are 4 to 1. If the probability of success is .5, i.e., 50-50 percent chance, then the odds of success is 1 to 1.

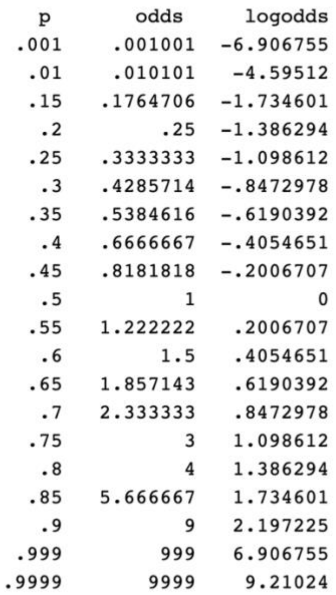
The transformation from probability to odds is a monotonic transformation, meaning the odds increase as the probability increases or vice versa. Probability ranges from 0 and 1. Odds range from 0 and positive infinity. Below is a table of the transformation from probability to odds and we have also plotted for the range of p less than or equal to .9.



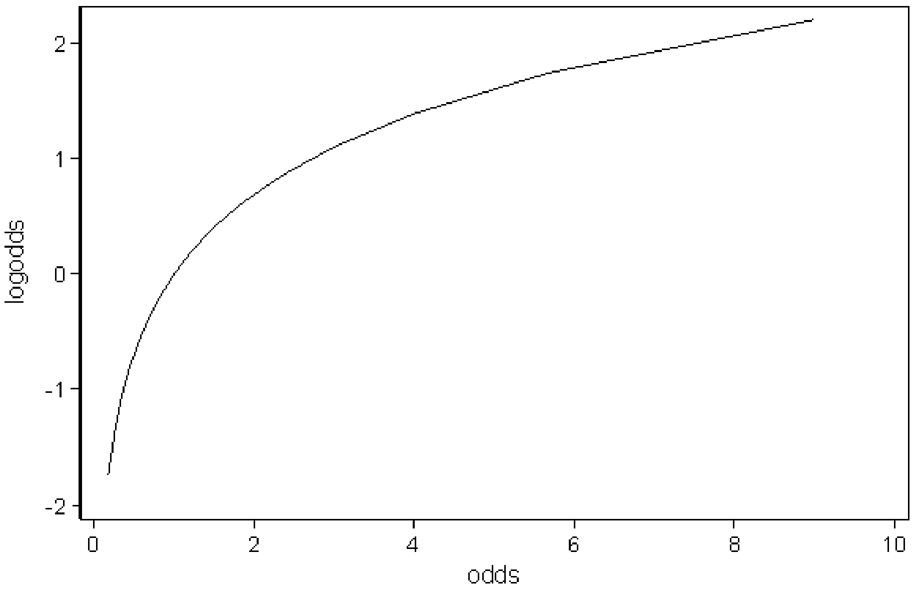
```
    p        odds
 .001     .001001
  .01     .010101
  .15    .1764706
   .2         .25
  .25    .3333333
   .3    .4285714
  .35    .5384616
   .4    .6666667
  .45    .8181818
   .5           1
  .55    1.222222
   .6         1.5
  .65    1.857143
   .7    2.333333
  .75           3
   .8           4
  .85    5.666667
   .9           9
 .999         999
.9999        9999
```



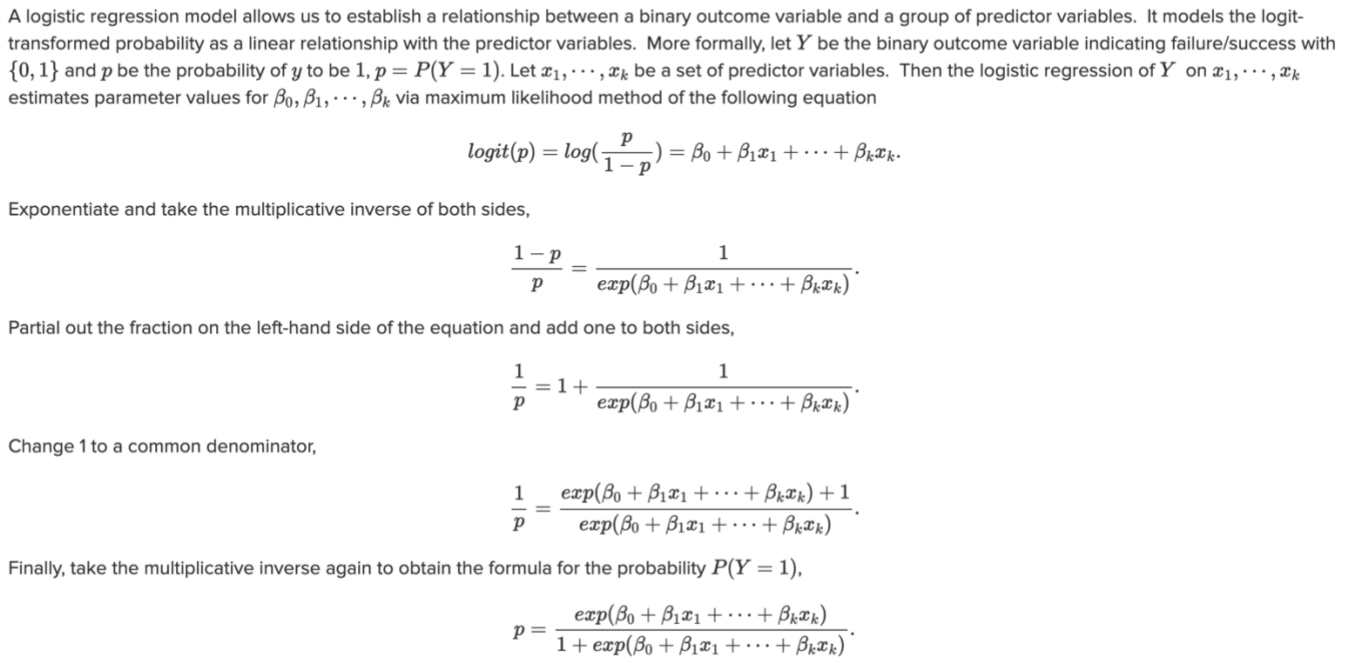
The transformation from odds to log of odds is the log transformation. Again this is a monotonic transformation. That is to say, the greater the odds, the greater the log of odds and vice versa. The table below shows the relationship among the probability, odds and log of odds. We have also shown the plot of log odds against odds.



```
     p       odds     logodds
  .001    .001001   -6.906755
   .01    .010101    -4.59512
   .15   .1764706   -1.734601
    .2        .25   -1.386294
   .25   .3333333   -1.098612
    .3   .4285714   -.8472978
   .35   .5384616   -.6190392
    .4   .6666667   -.4054651
   .45   .8181818   -.2006707
    .5          1           0
   .55   1.222222    .2006707
    .6        1.5    .4054651
   .65   1.857143    .6190392
    .7   2.333333    .8472978
   .75          3    1.098612
    .8          4    1.386294
   .85   5.666667    1.734601
    .9          9    2.197225
  .999        999    6.906755
 .9999       9999     9.21024
```



Why do we take all the trouble doing the transformation from probability to log odds? One reason is that it is usually difficult to model a variable which has restricted range, such as probability. This transformation is an attempt to get around the restricted range problem. It maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity. Another reason is that among all of the infinitely many choices of transformation, the log of odds is one of the easiest to understand and interpret. This transformation is called logit transformation.



A logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables. More formally, let Y be the binary outcome variable indicating failure/success with $\{0,1\}$ and p be the probability of y to be 1, $p = P(Y = 1)$. Let $x_1, \cdots, x_k$ be a set of predictor variables. Then the logistic regression of Y on $x_1, \cdots, x_k$ estimates parameter values for $\beta_0, \beta_1, \cdots, \beta_k$ via maximum likelihood method of the following equation

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Exponentiate and take the multiplicative inverse of both sides,

$$\frac{1-p}{p} = \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Partial out the fraction on the left-hand side of the equation and add one to both sides,

$$\frac{1}{p} = 1 + \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Change 1 to a common denominator,

$$\frac{1}{p} = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + 1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Finally, take the multiplicative inverse again to obtain the formula for the probability $P(Y = 1)$,

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

So that's how we arrive at the Logit transformation:

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

Side note: Why is logistic regression considered a linear model?

Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters. The output cannot depend on the product or quotient etc. of its parameters.

References:

A great resource to understand the derivation of the Logistic equation:
https://stats.idre.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/

Wikipedia: https://en.wikipedia.org/wiki/Logistic_regression