Logistic Regression: Getting Numbers from Levels

Logistic Regression is a subset of Classification (more on that later)

Logistic Regression

At this point we have covered:

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor

- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven’t seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Logistic Regression: A Morbid Example

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming.

There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake.

The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)

Let’s look at this data in R!

Logistic Regression: A Morbid Example

It seems clear that both age and gender have an effect on someone’s survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn’t something we can transform our way out of (there is no 0.5 “dead”) - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Logistic Regression: A Morbid Example

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable

2. A linear model
   - $\eta=\beta_0+\beta_1 X_1+\cdots+\beta_n X_n$

3. A link function that relates the linear model to the parameter of the outcome distribution
   - $g(p)=\eta$ or $p=g^{-1}(\eta)$

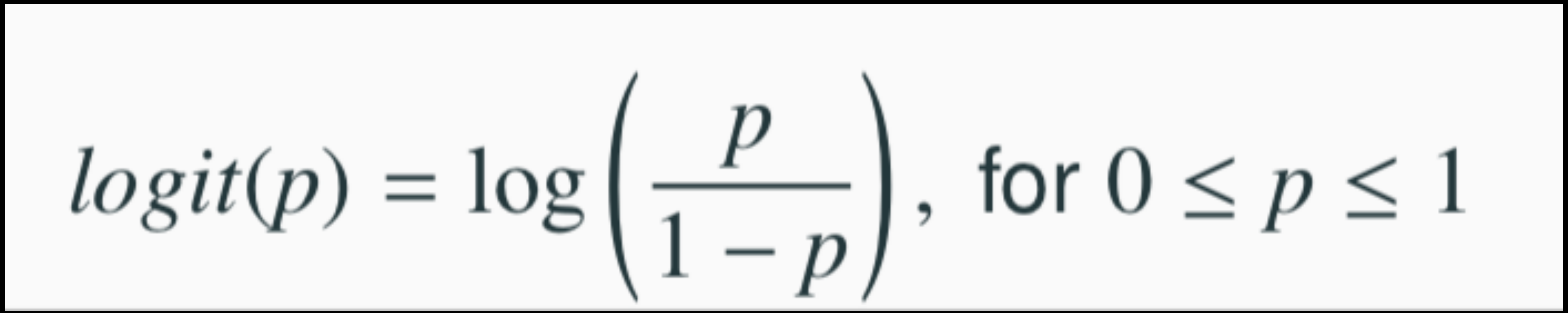
g turns probability into a number, g^{-1} turns linear fit to probability

Logistic Regression: A Morbid Example

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model *p* the probability of success for a given set of predictors.

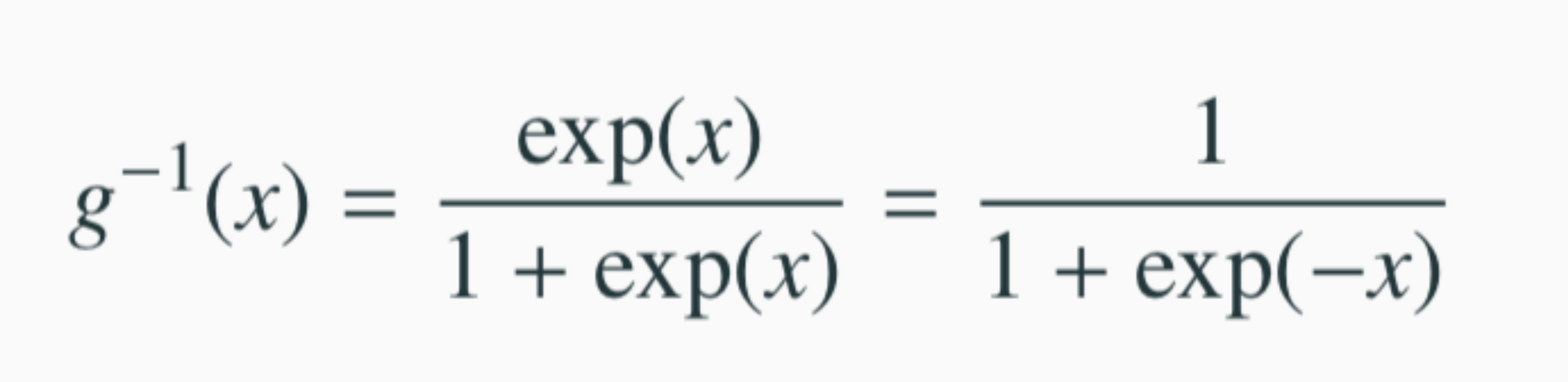
To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to *p*. There are a variety of options but the most commonly used is the logit function.



$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

Logistic Regression: A Morbid Example

The logit function takes a value between 0 and 1 and maps it to a value between −∞ and ∞.



$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

The inverse logit function takes a value between −∞ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

Logistic Regression: A Morbid Example

Ok, so what does the totality of our model look like?

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

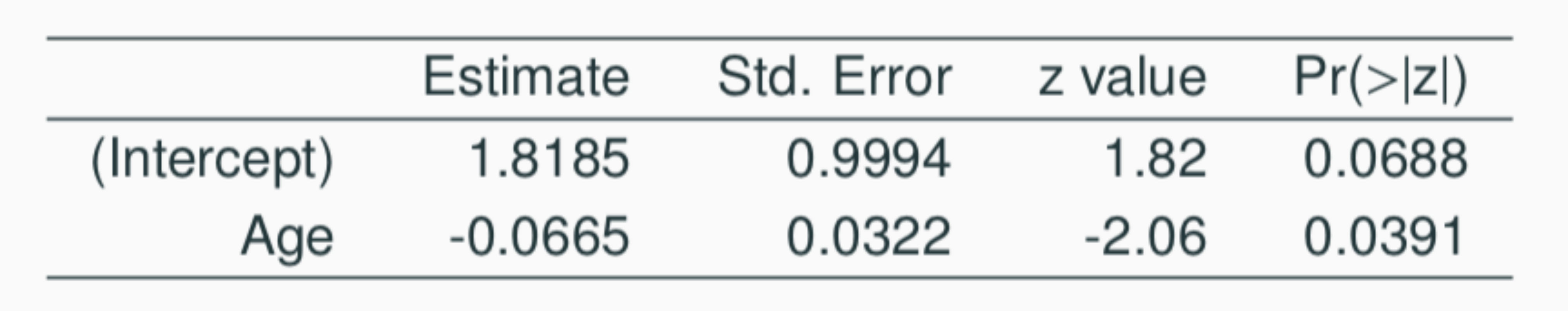
$$\text{logit}(p) = \eta$$

From which we back out the probability of survival based on parameters 1-n, for the *i*th observation:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

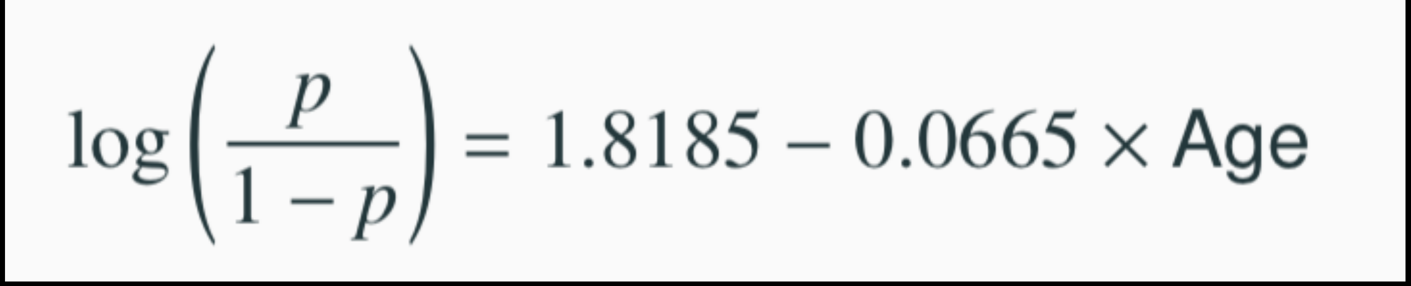
Logistic Regression: A Morbid Example

Give me an example!



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

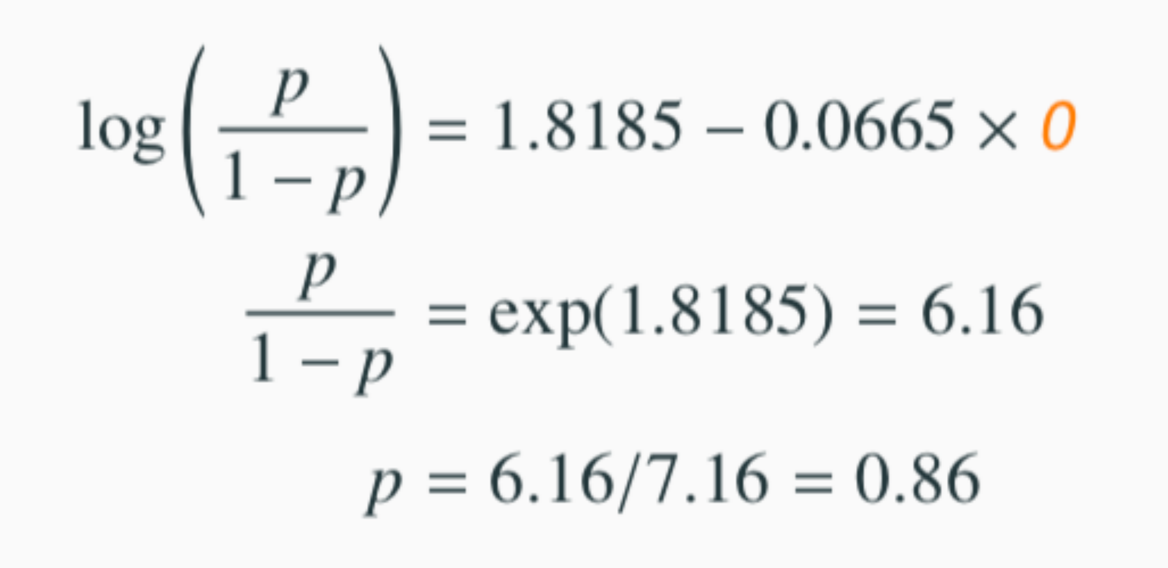
Model:



$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

So, for example, the odds of survival of a newborn (age = 0):

Can I get an R example??



$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

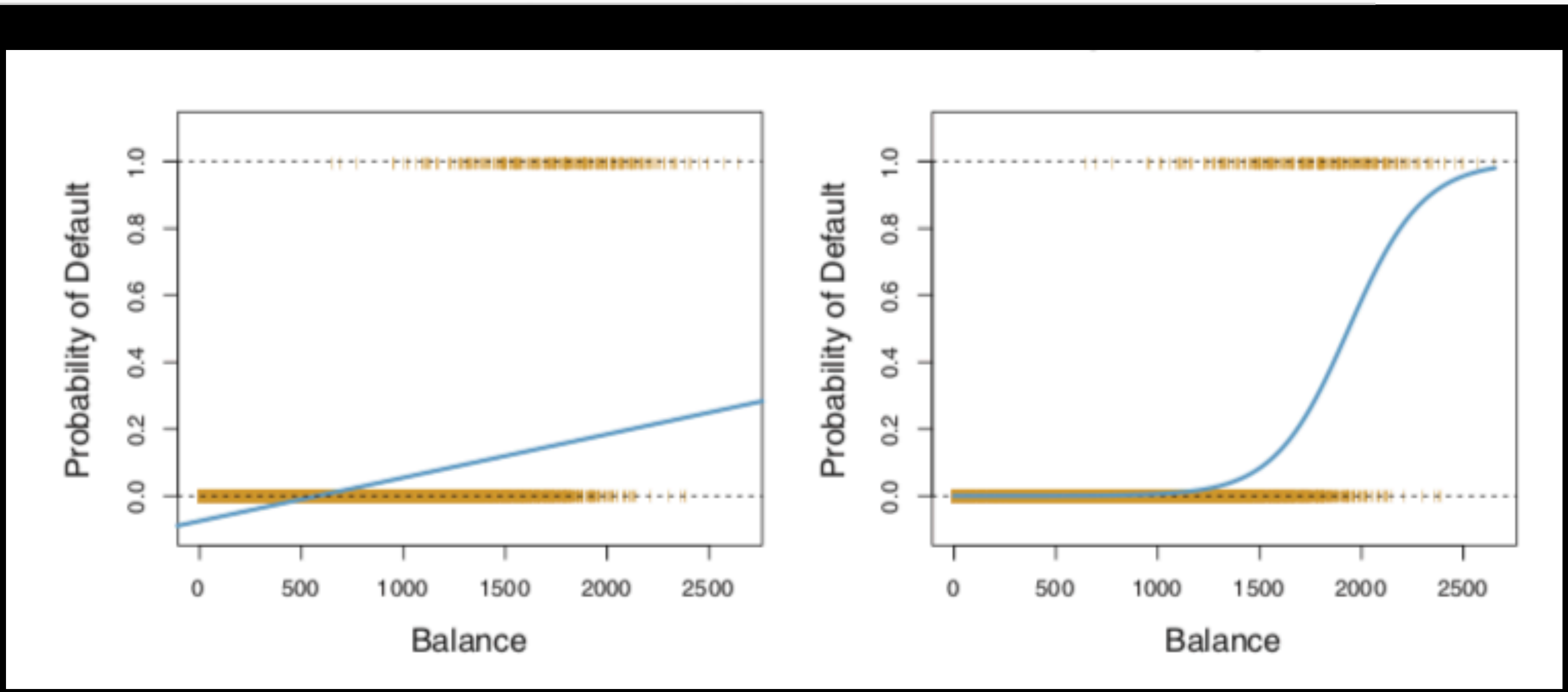
$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

A note about the logit function

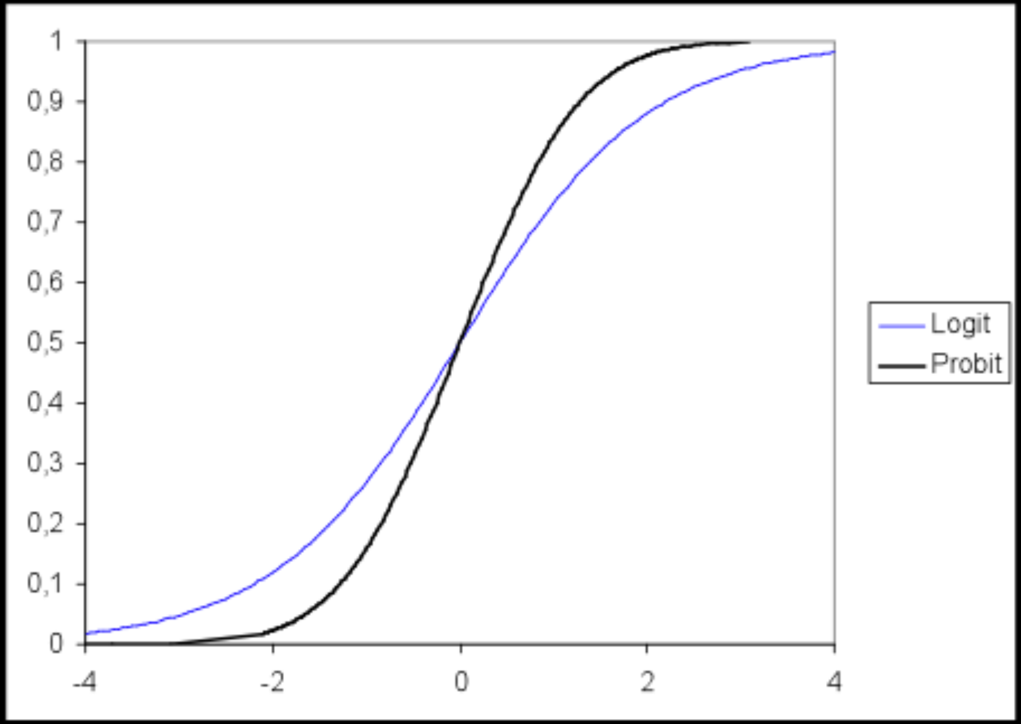
$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Weird mapping between a linear fit & probability of success



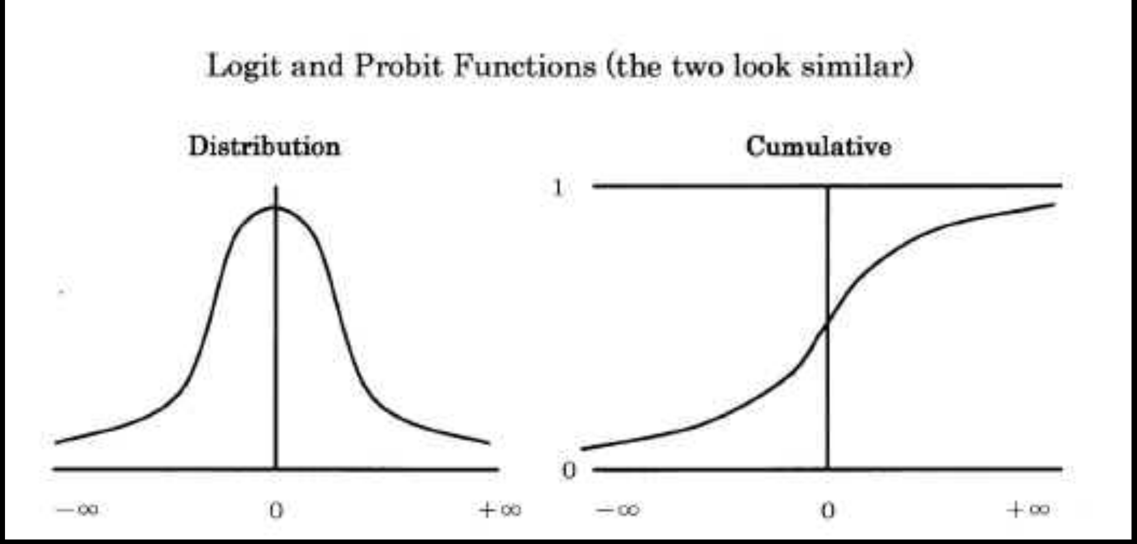
This is essentially to map observed successes and failures to probabilities.

Want to avoid probabilities < 0 or > 1

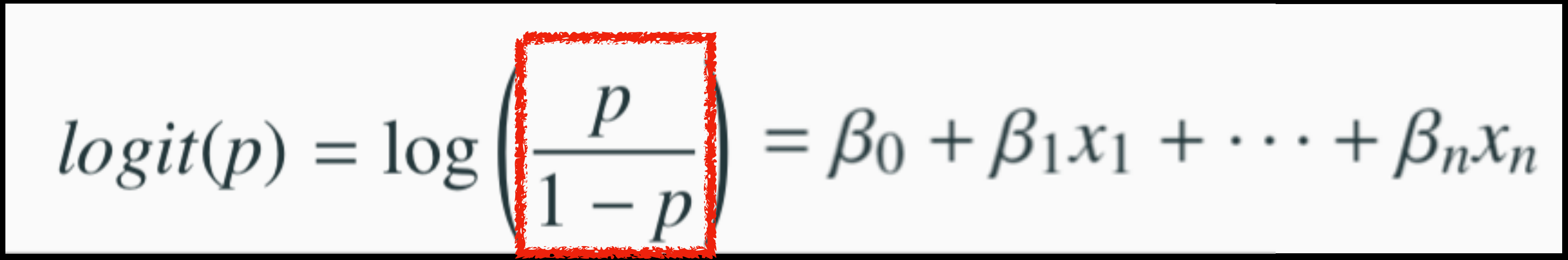


Other possible mappings however.

Logit the most (currently) popular.



Another note about the logit function



$$logit(p) = \log\left(\boxed{\frac{p}{1-p}}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

“the odds” or “odds”

Think horse racing:
1 in 20 odds means

1/20 = p/(1-p)

$$p = \frac{1/20}{1 + 1/20} = 0.048$$