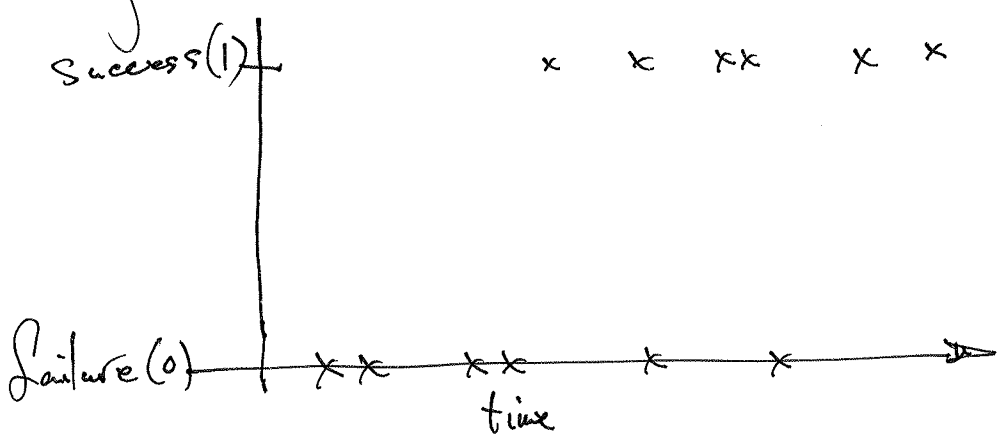
Logistic Regression

Suppose we want to predict the chance $p(x)$ of accomplishing a goal (like passing a course, or finishing a race in a certain amount of time) as a function of some quantitative variable x (like amount of study time, or time spent training). The response is binary, and data might look like this:



We seek a function $p(x)$ that sort-of "best fits" these data so that

$$p(x) = P(Y = 1 \mid X = x)$$

Y is the response and has a Bernoulli(p) distribution (or could be binomial...)

Simple linear (or polynomial) regression is not a good model in this case since we need the function $p(x)$ to satisfy $\lim_{x \to \infty} p(x) = 1$, $\lim_{x \to -\infty} p(x) = 0$, and $p(x)$ should be monotone (note these are properties of cumulative distribution functions, so $p(x)$ is a cdf.)

Consider the odds of an event E:

$$\text{odds of } E := \frac{P(E)}{1-P(E)} \in [0, \infty].$$

While probabilities are always in $[0,1]$, odds can take values from the extended non-negative real numbers... the odds of an event CAN BE ∞... Now, the natural log of the odds obeys

$$-\infty \leq \log \frac{P(E)}{1-P(E)} \leq \infty$$

So maybe it would make sense to build a linear model for the natural log of the odds ratio:

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

After some algebra...

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

$$\Longrightarrow \quad p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

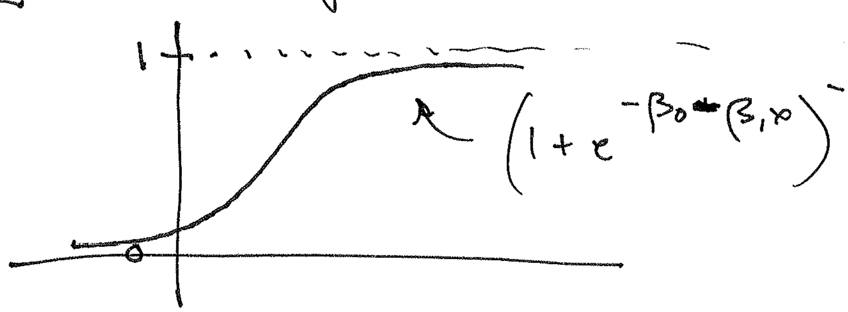
or $$\boxed{p(x) = \left(1+e^{-\beta_0-\beta_1 x}\right)^{-1}.}$$

You should graph several instances of this $p(x)$ function with different values of β_0 and β_1, like

β_0	0	0	0	0	2	2	2	2	4	4	4	4
β_1	1	2	3	-4	1	2	3	-4	1	2	3	-4

When you do, you'll see that β_0 is a location parameter, and β_1 is a shape parameter. What happens as β_0 increases? as β_1 increases? when $\beta_1 < 0$?

Regardless of the β_0 and β_1 values, these curves have the same basic shape, which is a sigmoid or s-shape:



Like the c.d.f. of a normal random variable, this curve has rotational symmetry. Specifically, the point of rotational symmetry here is $\left(-\beta_0/\beta_1 \,,\, 1/2\right)$.

There are other ways to transform $p(x)$ for which a linear model might be appropriate. We are mainly concerned with these three ways (they seem to be the most popular):

1. The <u>logit function</u>:

$$\eta(x) = \log\left(\frac{x}{1-x}\right)$$

(which is what we've just been using)

2. The <u>probit function</u>:

$$\eta(x) = \Phi^{-1}(x),$$

where $\Phi(x) = P(\text{a \underline{standard} normal r.v.} \leq x)$.

3. The <u>complementary log-log function</u>

$$\eta(x) = \log(-\log(1-x)).$$

The transformation η is typically called a <u>link function</u> (as in, the <u>logit-link</u>, or the <u>probit link</u>, etc.)

Back to the model in ☺ that uses the logit link...

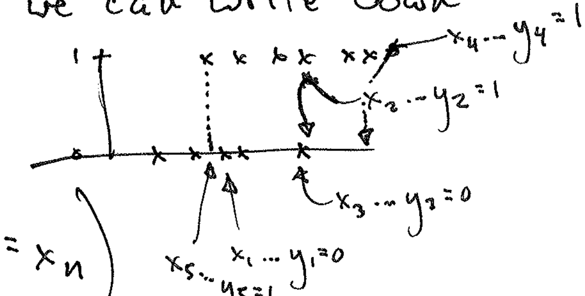
How do we estimate the β_0 and β_1 coefficients? We typically need a computer to carry out a numerical procedure, but to shed some light ... we could write

$$P(Y=y \mid X=x) = p(x)^{y}\,(1-p(x))^{1-y} \quad \text{for } y = 0 \text{ or } 1.$$

Y is the Bernoulli(p(x))

response : $y \in \{0,1\}$

Now suppose we make n independent observations of Y when $X = x_1, x_2, \dots, x_n$. Then we can write down the likelihood function as



$$P(Y_1 = y_1, \dots, Y_n = y_n \mid X_1 = x_1, \dots, X_n = x_n)$$

by independence

$$= \prod_{k=1}^{n} P(Y_k = y_k \mid X_k = x_k)$$

$$= \prod_{k=1}^{n} p(x_k)^{y_k}\,(1-p(x_k))^{1-y_k} \qquad (*)$$

We want to maximize the chance of seeing what we actually saw over all model parameter values (β_0 and β_1). This is called maximum likelihood. So fix the x_k and y_k

values in the conditional probability / likelihood function in (*), and consider that function to be a function of the β_0 and β_1. ~~We basically want to differentiate the~~ It turns out (as the case often is) it's easier to maximize the log of the likelihood function (which is okay since the log function is monotone). We call this the log-likelihood function:

$$L = \log P(Y_1 = y_1, \dots, Y_n = y_n \mid X_1 = x_1, \dots, X_n = x_n, \beta_0, \beta_1)$$

We need to think of the log-likelihood function as a function of the model parameters... the data values are fixed.

$$= \log \prod_{k=1}^{n} \left[p(x_k)^{y_k} (1-p(x_k))^{1-y_k} \right] = \sum_{k=1}^{n} \left[\log p(x_k)^{y_k} + \log (1-p(x_k))^{1-y_k} \right]$$

$$= \sum_{k=1}^{n} y_k \log p(x_k) + \sum_{k=1}^{n} (1-y_k) \log(1-p(x_k))$$

$$= \sum_{k=1}^{n} \log(1-p(x_k)) + \sum_{k=1}^{n} y_k \left(\log(p(x_k)) - \log(1-p(x_k)) \right)$$

$$= \sum_{k=1}^{n} \underbrace{\log(1-p(x_k))}_{= -\log\left[1+e^{\beta_0+\beta_1 x_k}\right]} + \sum_{k=1}^{n} y_k \underbrace{\log \frac{p(x_k)}{1-p(x_k)}}_{= \beta_0 + \beta_1 x_k}$$

And so our problem becomes finding the values of $\hat{\beta}_{0_{MLE}}$ and $\hat{\beta}_{1_{MLE}}$ that maximize

"maximum likelihood estimator"

$$\sum_{k=1}^{n} y_k\left(\hat{\beta}_{0_{MLE}} + \hat{\beta}_{1_{MLE}} x_k\right) - \sum_{k=1}^{n} \log\left[1 + e^{\hat{\beta}_{0_{MLE}} + \hat{\beta}_{1_{MLE}} x_k}\right].$$

This is difficult to do by hand! You can try ... try differentiating this expression with respect to $\hat{\beta}_{0_{MLE}}$ and $\hat{\beta}_{1_{MLE}}$ and setting to 0 ...

Statistical software packages (like SAS and R) will do this.