Logistic regression - a *very brief* introduction

We simply don't have the time to go into the details here, so a brief outline will have to do.

Logistic regression is usually used if the response variable (our Y) has only two possible outcomes. For example, *present* vs. *absent* or 0 vs. 1, etc.

It can also be used if there are more than two outcomes (e.g., three or four), but the math gets more complicated.

We can't use a linear regression approach since the usual regression equation, $\hat{Y} = b_0 + b_1 X$ assumes that Y is continuous and (presumably) can take on any value between $-\infty$ to $+\infty$.

Instead, we have to use a *non-linear* approach. We model the following equation:

$$Pr(Y = 1) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

In other words, as usual, we find an estimate for β_0 and β_1. This equation then give us the *probability* of getting the specified value of Y.

To illustrate this, let's use an example based on the BIOL 214 text (*Statistics for the Life Sciences*, Samuels, Witmer & Shaffner, 4^th ed.):

We're interested in predicting whether or not a tumor has spread, so we let $Y = 1$ indicate that the *tumor has spread* and let $Y = 0$ indicate that the *tumor has not spread*.

Then we measure the size of the tumor, and try to use this to predict whether or not the tumor has spread.

We can then use the equation above to tell us the probability of the tumor having spread.

Some comments on the calculations:

We can not solve this equation analytically and get values for b_0 or b_1.

So how do we get values for b_0 and b_1?

We use an approach based on something called *maximum likelihood*. We can't really explain this without a lot more lecture (and math!), but essentially we look for the values of b_0 and b_1 that are *most likely* to have given us the data we observe.

(A really simple example: suppose we somehow didn't know the value for p (the probability of heads for a coin). So we toss a coin 10 times and get five heads, the most likely value for p is 0.5).

Unfortunately, even using a maximum likelihood approach, there is no analytical solution (we can't get values of b_0 and b_1 just by solving various equations).

To figure out the values of b_0 and b_1 we need to use some sophisticated guesswork (well, it's not really guesswork).

We make an educated guess (or let R do this) at the values of b_0 and b_1.

Then we use a series of iterative steps to find a solution (some of these are based on ideas developed by Newton!).

Eventually, these iterations will yield values of b_0 and b_1 that we can show are the solutions to the equation.

We will *not* learn how to do this. Instead we'll walk through an example using R to see what logistic regression is all about.

Let's walk through the example in the 214 textbook. We are given the following data:

Tumor size (cm), X	Spread, Y	Tumor size (cm), X	Spread, Y
6.5	1	6.2	1
6.3	0	2.0	0
3.8	1	9.0	1
7.5	1	4.0	0
4.5	1	3.0	1
3.5	1	6.0	1
4.0	0	4.0	0
3.7	0	4.0	0
6.3	1	4.0	0
4.2	1	5.0	1
8.0	0	9.0	1
5.2	1	4.5	1
5.0	1	3.0	0
2.5	0	3.0	1
7.0	1	1.7	0
5.3	0		

From this we want to predict whether or not a tumor is present.

Let's walk through the steps in R:

First, we enter the data as usual:

```
size <- scan(nlines = 2) 6.5 6.3 3.8 7.5 4.5 3.5 4.0 3.7 6.3 4.2 8.0 5.2
5.0 2.5 7.0 5.3
6.2 2.0 9.0 4.0 3.0 6.0 4.0 4.0 4.0 5.0 9.0 4.5 3.0 3.0 1.7

spread <- scan(nlines = 1) 1 0 1 1 1 1 0 0 1 1 0 1 1 0 1 0 1 0 1 0 1 1 0
0 0 1 1 1 0 1 0
```

Then we tell R to do a logistic regression (note the use of `glm`, which stands for **G**eneral **L**inear **M**odel, which is an extension of `lm`):

```
tumormodel <- glm(spread ~ size, family = binomial(link = "logit")
summary(tumormodel)
```

And R will give us the following result:

```
Call:
glm(formula = spread ~ size, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0657  -1.1288   0.5657   0.9844   1.4185

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0858     1.2256  -1.702   0.0888 .
size          0.5117     0.2561   1.998   0.0457 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 42.165  on 30  degrees of freedom
Residual deviance: 37.002  on 29  degrees of freedom
AIC: 41.002

Number of Fisher Scoring iterations: 4
```

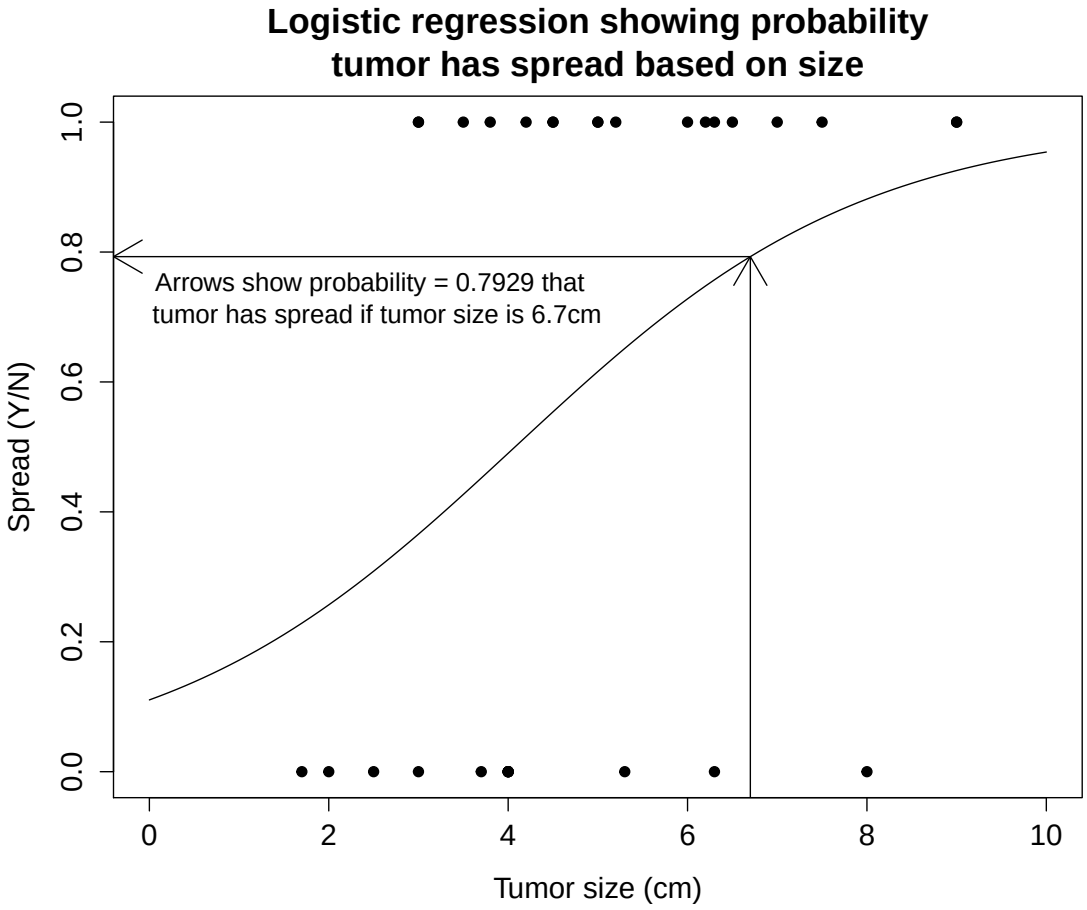
If we look at this, we can then plug in the the estimates for b_0 and b_1 into our equation:

$$Pr(Y = 1) = \frac{e^{-2.0858+0.5117x}}{1 + e^{-2.0858+0.5117x}}$$

(Which, incidentally, matches what's in the 214 textbook).

So what does it all mean?

Let's graph this equation and see what it looks like, together with the original data:



From the graph we see the original data plotted at $Y = 1$ or $Y = 0$.

Using the equation for the probability that $Y = 1$ as given above, we generate the curve drawn on the graph.

We can also use our equation to calculate the probability of the tumor has spread if we have a tumor of a certain size.

For example, if we have tumor that's 6.7 cm in size, we can do the following:

$$Pr(\text{tumor has spread}) = \frac{e^{-2.0858+0.5117(6.7)}}{1+e^{-2.0858+0.5117(6.7)}} = \frac{e^{1.34259}}{1+e^{1.34259}} = \frac{3.8289}{4.8289} = 0.7929$$

This is indicated by the arrows in the graph above.

So let's make some final comments on logistic regression:

Logistic regression is useful when we are trying to predict the probability of an event with (usually) two outcomes.

It is not the only way of doing this - other techniques include Poisson regression or Probit models (there are others).

Logistic regression can easily be expanded to multiple logistic regression (including techniques similar to stepwise, etc.).

Logistic regression does have some assumptions, but they're not as strict as those for least squares regression (what we've been doing up until now):

Probably the most important you need to worry about are the usual assumption about random data, and the assumption that all the Y's are independent (which should be obvious since we're looking at the probability of Y).

There are several other assumptions, but we will not discuss them as they require quite a bit more explanation. Let's just say if you find you need to use logistic regression a lot, you should probably look into this or talk to someone.