SMIII lectures

Week 6

Logistic Regression

Logistic Regression

In a series of experiments, beetles were exposed to gaseous carbon disulphide at various concentrations for a period of 5 hours and the numbers of deaths recorded.

Concentration (x)	Number exposed (n)	Deaths (y)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	62	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

The purpose of the analysis is to relate the probability of death to the concentration of carbon disulphide.

The binomial model

In this experiment, can consider the numbers of deaths as independent binomial observations.

That is, $Y_i \sim B(n_i, \pi_i)$ where

- n_i is the number beetles exposed at concentration x_i;
- π_i is the probability of death at concentration x_i;
- Y_i is the number of deaths observed at concentration x_i.

Need a model to relate π_i to x_i. That is, a model of the form

$$\pi_i = f(\boldsymbol{x}_i).$$

Linear model for the probability

In analogy to linear regression, a simple suggestion would be to consider the linear model

$$\pi_i = f(x_i) = \beta_0 + \beta_1 x_i$$

However, such a model can be seen to be unsuitable because probabilities are constrained to satisfy $0 \leq \pi_i \leq 1$ whereas linear functions are not.

The logistic regression model

To overcome this difficulty the logistic regression model is defined by

$$\eta_i = \beta_0 + \beta_1 x_i.$$

where

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{logit}(\pi_i).$$

is the **logit transformation**.

The model can be written as

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i.$$

Properties of the logit transformation

- For $0 < \pi_i < 1$ the **logit**,

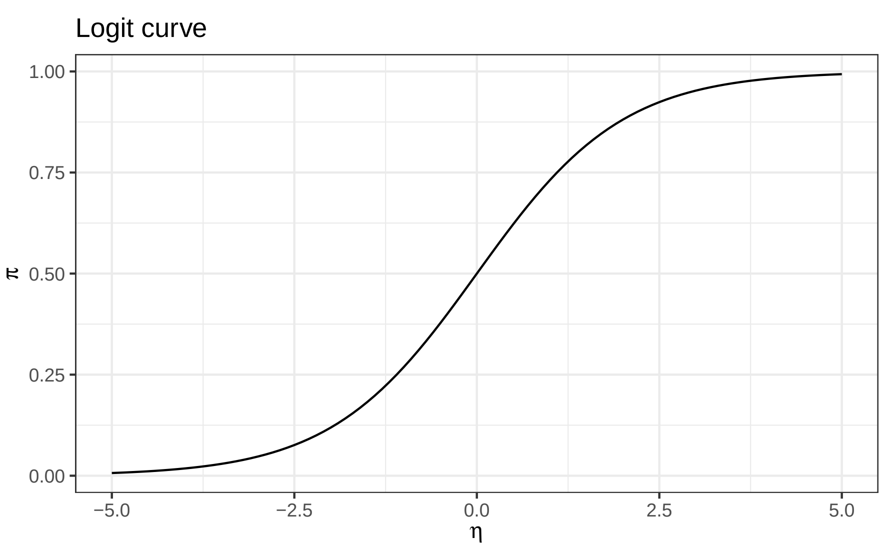
$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

  satisfies $-\infty < \eta_i < \infty$
- The logit transformation is invertible. In particular,

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$$

- The logit transform and its inverse are strictly increasing.
- Note the logit is also the log-odds.

The logistic curve



Beetle example

For the beetles dataset, we get the estimated logistic regression model is

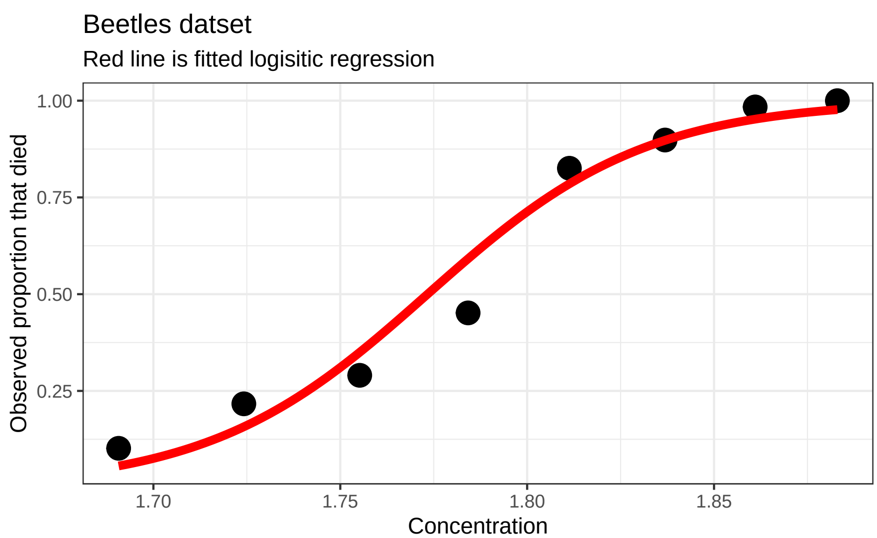
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -60.56 + 34.15 \times \text{Concentration}_i.$$

or, equivalently,

$$\pi_i = \frac{\exp(-60.56 + 34.15 \times \text{Concentration}_i)}{1 + \exp(-60.56 + 34.15 \times \text{Concentration}_i)}.$$

Example

```
## Rows: 8 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): Concentration, Exposed, Dead
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```



Remarks

- The interpretation of logistic regression is analogous to linear regression.
- For example, the slope coefficient in the model

  $$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

  implies that a one-unit increase in x corresponds to a change of β_1 to the log-odds.
- In the beetle example, a one-unit increase in poison concentration leads to an increase of 34.15 in the log-odds of death.

No additive error

- The specification of linear regression is

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

  where $e_i \sim N(0, \sigma^2)$ independently.
- The specification of the logistic regression model is

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i.$$

- Note there is no additive error in that part of the model.
- Random variation in the data is modelled in the binomial assumption,

$$Y_i \sim B(n_i, \pi_i).$$