Topic 1
Introduction to Statistical Data Modelling

Topic 1: Contents

1.1 Statistical models

- By now we are aware of what statistical data modelling is in general terms:
  - It's concerned with describing a "data generating process" using mathematics, in way which **explicitly captures uncertainty through probability**.
- Statistical (data) modelling is a fundamental part of data science because in many situations where we are trying to understand data:
  - the phenomenon we are interested in, is itself intrinsically subject to uncertainty (e.g. election outcome);
  - there is a lack of scientific understanding (e.g. a theory) of the observed process (e.g. link between depression and access to green space);
  - the data values are subject to observational error.

1.1 Purpose of statistical models

- We 'build' statistical models in order to:
  - Understand relationships between various real-world processes (e.g. relationship between green space and depression);
  - Make predictions (e.g. predict how intense storms in the UK will get in the future under a changing climate);
  - Reduce the amount of data to small number of interpretable parameters (e.g. compare multidimensional data arising from complex physical models).
- Uncertainty is quantified using probability (e.g. confidence/credible intervals, hypothesis tests, p-values).

1.1 Framework for simple statistical models

- The simplest statistical models assume that the observed data $(y_1, \ldots, y_n)$ are **independent** realisations of a univariate **response variable** (i.e. the variable to be modelled).
- Mathematically, $(y_1, \ldots, y_n)$ are each a realisation of independent and identically distributed (i.i.d.) random variables $Y_1, \ldots, Y_n$, each following the same probability distribution which for this module we define (for simplicity) as

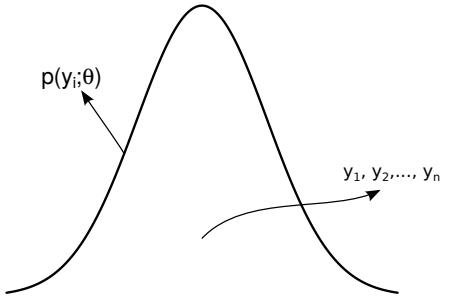
$$p(y_i; \boldsymbol{\theta}) = \begin{cases} Pr(Y_i = y_i), & \text{if } Y_i \text{ is discrete} \\ f(y_i) \text{ (prob. dens. function)} & \text{if } Y_i \text{ is continuous} \end{cases}$$

  where $\boldsymbol{\theta}$ is a vector of parameters associated with the probability distribution.
- E.g. if Y_i is Normal (Gaussian) then $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$p(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$$

1.1 Framework for simple statistical modelling

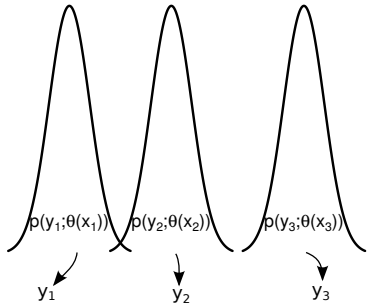
- Statistical modelling is about finding a **probability distribution** that (in some sense) best captures the **data generating process**, e.g.:



- Always remember that the actual data did not actually come from a probability distribution. The aim of a good statistical model is to be able to **explain** both the "random" nature of any data set (e.g. variance), as well as any structural patterns (e.g. mean).

1.1 Framework for simple statistical modelling

- More complex (and realistic) models involve relationships between the response and **covariates** $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, through parametrisation of $\boldsymbol{\theta}$ so the model becomes $p(y_i; \boldsymbol{\theta}(\boldsymbol{x}_{1,i}, \ldots, \boldsymbol{x}_{p,i}))$.



- This includes categorical covariates or **factors** which form groupings in the data (e.g. gender). Adding factors allows for non-independence in the response variable (e.g. temperature measurements in winter are more similar to each other than ones in summer).

1.1 Broad recipe for building statistical models

In practical terms, statistical modelling involves:

- Choosing an appropriate probability distribution respecting the nature of the data as much as possible: discrete/continuous, bounded at zero etc.
- Fitting the model to the data – this is part of statistical inference.
- Check that the model fits the data – very important!
- Use the model to learn about the the process that was modelled, e.g. quantify strength of relationships, predict unknown data, estimate probability extreme values etc.

1.2 The linear model

- Our main tool at this point is the linear (regression) model.
- The linear model:

$$\begin{aligned} Y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \qquad (Y_i \text{ independent given } \boldsymbol{x}_i) \\ \mu_i &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} = \sum_{j=0}^{p} \beta_j x_{i,j} \qquad (x_{i,0} = 1) \end{aligned}$$

  is applicable to a wide range of situations where the mean of y_i can be assumed Normally distributed, and its mean is linear in x_i.
- This framework also includes the case where x_i is a factor and so can be used to model differences between grouping structures
  - e.g. for modelling temperature y_i where the mean μ_i is different for each of the 4 seasons, we would include factor $x_i \in \{\text{winter, autumn, spring, summer}\}$.

1.2 The linear model

- It also includes the case where some covariates are **functions** of the raw covariates (e.g. x_i^2) and so can be used to model **non-linear** relationships, as long as these relationships can be expressed as linear functions of the unknown parameters involved.
- We can also sometimes appropriately transform the response variable so that the linear model framework becomes applicable where it is not suitable for the original response (e.g. 'Box-Cox' family of transformations).
- So the linear model is a much richer class of models for modelling relationships than it first appears.

1.2 Beyond the linear model

- The validity of the linear model depends upon three key assumptions:
  - The relationship between the **mean** value of the response and the covariates can be expressed as a **linear function** of unknown parameters;
  - The distribution of the response (about its mean) can be assumed to be Normal;
  - The variance of the response distribution is constant (it does not depend upon the mean value or indeed anything else).
- **However these will not always apply** – think about very non-linear relationships in the mean, or count data, or binary data, or data whose variance increases as the mean increases.
- The purpose of this module is to expand your statistical skill set to go well beyond the linear model.

1.3 Examples of modelling needs beyond the linear model

- By way of motivation let's look at some examples where modelling is required that does not fit into the linear model framework.
- We start with an example concerning effects of different types of fishing net on the strict legal controls imposed on commercial trawling to protect fish stocks;
  - Fish below a certain size cannot be landed and any fish which are 'too small' are often dumped overboard.
- One particular experiment involves surrounding the *'real'* net of a trawler with an extremely fine mesh net. The idea is that the *'real'* net should catch only larger fish, but the fine mesh net catches all fish.
- By studying the catches in both nets, we can get an idea as to how many fish of each length are caught by the *'real'* net.

1.3 Fish data

Length	Escaping	Total	Length	Escaping	Total
21	1	1	44	0	13
23	1	1	45	0	3
25	6	6	46	0	5
26	6	6	47	0	3
27	9	9	48	0	3
28	7	8	49	0	4
29	3	3	50	0	1
30	8	11	52	0	3
31	6	7	53	0	1
32	2	6	54	0	3
33	6	21	55	0	1
34	5	23	56	0	1
35	2	15	57	0	1
36	0	14	63	0	1
37	0	12	65	0	2
38	1	14	66	0	1
39	0	8	67	0	1
40	1	10	68	0	3
41	0	14	85	0	1
42	0	6	86	0	1
43	0	10	87	0	1

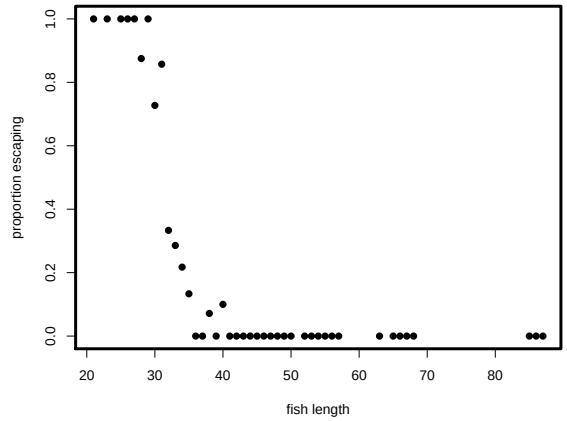
- Length: of fish in cm.
- Total: number of fish of particular length caught by fine mesh
- Escaping: number of fish of particular length that have "escaped" the real net.

1.3 Fish data modelling considerations

- Interest lies in understanding the propensity of fish escaping, as a function of their length.
- Fish of different sizes will have different chances of escaping (from the real net), but it seems reasonable to assume that fish of the *same* size will each have the same probability of escaping.
- First, we need to define the response variable. We have several fish at most lengths and let's call y_i the number of fish escaping, out of N_i entering the netting.
- Clearly we cannot use y_i as the response (why not?), but instead we could focus on the proportion $z_i = \frac{y_i}{N_i}$ of fish of length x_i escaping from the real net.

1.3 Fish data modelling considerations

- Let's plot observed proportion escaping $z_i = \frac{y_i}{N_i}$ against fish length x_i.



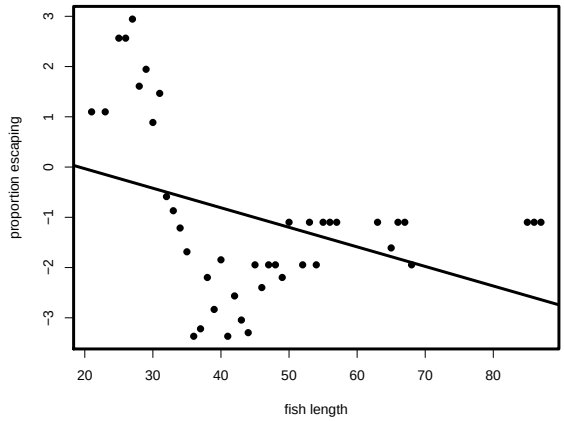
- We see that we have a classic 'S-shaped' (in this case a backward 'S') relationship between proportion of fish escaping and their length.
- How should we model such data? Does a linear model make sense?

1.3 Linear model for fish

- The response variable z_i is bounded in $[0, 1]$ so we should think about transforming this if we are to fit a linear model. (Transforming may also help in "straightening" the non-linear relationship.)
- Something that may do the trick is the so-called 'logit' function: $\text{logit}(z) \equiv \log\left(\frac{z}{1-z}\right)$ for any $z \in (0, 1)$ so that $\text{logit}(z) \in (-\infty, \infty)$.
- We cannot actually use this directly because of problems with some observed z_i being 0 or 1.
- We can get round this by a simple 'trick' of calculating an adjusted proportion as $z_i' = \frac{y_i+0.5}{N_i+1}$ and then calculating the $\text{logit}(z_i')$.

1.3 Linear model for fish

- The transformed z_i' plotted against length x_i:



- We could now contemplate using a linear model on this transformed data (but not really very satisfactory as indicated by the fit—see `topic1.R`).

1.3 Linear model for fish

- So with this data there are problems with trying to transform the response into the linear model framework:
  - We had to make an arbitrary modification to the data to avoid numerical problems;
  - The transformation did not really 'linearise' the problem very well;
  - The variance does not look very constant for different fish lengths;
  - What justification do we have for believing that the transformed response, $\log\left(\frac{z_i}{1-z_i}\right)$, would be Normally distributed?
- It would be better (and more elegant) to leave the data untransformed and propose a statistical model that respect the 'nature' of the data, namely
  - counts (non-negative integers) of a binary event (escape or not) out of total number of 'trials'.

1.4 A Binomial fish model

- If each fish has theoretical probability π_i of escaping (independently of other fish), then we can regard the number escaping, y_i, as having a Binomial distribution, from the corresponding total number N_i. That is:

$$Y_i \sim Bin(N_i, \pi_i) \qquad \text{(with } Y_i \text{ being independent)}.$$

- We then need a suitable 'backward S' relationship between the **response mean** $\mu_i = N_i\pi_i$ and the fish length x_i. Or alternatively (since N_i is a known constant) between π_i and x_i.
- Using the same reasoning as previously we could use a **logistic** relationship:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad \Longrightarrow \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

1.4 A Binomial model for fish

- So in summary the a-priori more sensible and justifiable model we are proposing is:

$$\begin{aligned} Y_i &\sim Bin(N_i, \pi_i) \qquad Y_i = 0, 1, \ldots, N_i \quad \text{(indep.)} \\ \pi_i &= \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \end{aligned}$$

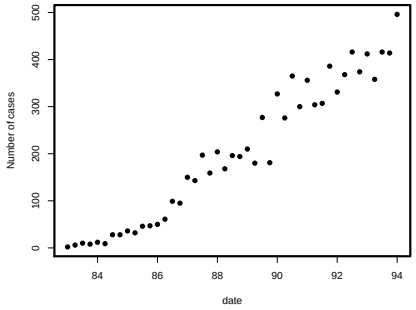
- But how do we estimate the unknown parameters β_0 and β_1 and quantify uncertainty? How do we assess if this is in fact a 'good' model or not?
- We come back to that later. For the moment just note that this is new territory (e.g. Y_i is not Gaussian, $\mathrm{E}[Y_i] = N_i\pi_i$ and $\mathrm{var}[Y_i] = N_i\pi_i(1-\pi_i)$, so if the mean varies with x_i then so does the variance—it's not a constant)
- Binomial data are in fact quite common and a special case is binary (Bernoulli) data where $N_i = 1$ and $y_i = 0$ or 1 so that

$$Y_i \sim Bin(1, \pi_i) \qquad \text{or} \qquad Y_i \sim Ber(\pi_i)$$

1.4 Examples of modelling needs beyond the linear model

- And we don't need to stop here in looking for common kinds of data which can't sensibly be fitted into the linear model framework.

- For example consider data on numbers of quarterly reported AIDS cases in the U.K. from Jan 1983 to Mar 1994.



- Clearly cases are increasing. But how do we model the underlying rate of increase?

1.4 Count data

- Sensible model might be that number of cases, Y_i, are Poisson distributed (count is effectively unbounded) with a parameter λ_i (the mean) which has an exponential relationship to time x_i i.e.

$$\begin{aligned} Y_i &\sim Poisson(\lambda_i) \qquad Y_i = 0, 1, 2, \ldots \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_i \implies \lambda_i = e^{\beta_0 + \beta_1 x_i} \end{aligned}$$

- Are there quarterly effects? We might want to add a "quarter" factor into the λ_i (mean) relationship.

- The point is, Y_i is not Normal and $\mathrm{E}[Y_i] = \lambda_i$, $\mathrm{var}[Y_i] = \lambda_i$, so if the mean varies with x_i then so does the variance—it's not constant.

- Again, not difficult to think of many situations where similar Poisson 'count' data might arise e.g.:
  - Disease counts in space and time (e.g. daily Covid-19 cases in each England county);
  - Numbers of storms per season;
  - etc.

1.4 Examples of modelling needs beyond the linear model

Other examples include:

- Multiway **contingency tables**: surveys or studies which record the counts of observational units (e.g. people) falling into different categories. Here the model may be multinomial (more than two categories of response) where the probability of being in each of these categories may depend on the other classifying factors.
- **Strictly positive or non-negative data** such as rainfall or income or duration data where the time to an event is of interest, e.g. difference in survival times of leukaemia patients under different treatments.
- There are a lot of data out there for which the linear model framework (despite its versatility) will be inadequate. We need to push that framework further and find ways to provide the same kind of unifying approach and computational convenience, but in a broader framework.

1.5 Approaches to fitting models and associated inference

- So what general purpose **inferential frameworks** are out there allowing us to fit a wider range of models than the linear model?

- We seek a 'framework' that allows us to:
  - Obtain 'good' estimates of the unknowns (parameters) in the model;
  - quantify the associated uncertainty (e.g. less uncertainty for same model fitted to larger data set)
  - Test hypotheses about these unknowns;
  - Assess the overall 'goodness of fit' of the resulting model;
  - Compare and formally test the difference between competing models.

1.5 Approaches to fitting models and associated inference

- There are three mainstream approaches:
  1. **Likelihood theory** (and variants of it such as 'quasi likelihood', 'restricted likelihood', 'penalised likelihood' etc.);
  2. **Bayesian inference** which together with Markov Chain Monte Carlo offers a unifying modelling framework;
  3. **Resampling** methods ('monte carlo methods', 'jacknifing', 'bootstrapping' etc.) are very useful when the models are mathematically intractable, or when one wishes to avoid specific parametric assumptions about the form of the model.
- In this course the focus is on the first of these. However, we should not think of these as competing approaches, but as complementary—*use no more than is needed*.
- We start by reviewing and introducing some key aspects of **likelihood theory**.

1.5 What is likelihood?

- Consider data $\boldsymbol{y} = (y_1, \ldots, y_n)$ that we assume are each respectively an independent realisation of random variables $Y_1, \ldots, Y_n$, where Y_i all have the same probability distribution (they are i.i.d.).
- Denote this distribution by $p(y_i; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are the unknown parameters.
- Consider y_1. The probability of this particular value according to our model is $p(y_1; \boldsymbol{\theta})$. The probability of the second data point is $p(y_2; \boldsymbol{\theta})$ and so on.
- Since we assumed independence, the joint probability of all the data given this model is simply the product:

$$p(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i; \boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{y})$$

- Viewed as a function of $\boldsymbol{y}$ for fixed $\boldsymbol{\theta}$, $p(\boldsymbol{y}; \boldsymbol{\theta})$ is a probability distribution. Viewed as a function of $\boldsymbol{\theta}$ for fixed $\boldsymbol{y}$ (i.e. the observed $(y_1, \ldots, y_n)$) it is called the **likelihood** and is denoted as $L(\boldsymbol{\theta}; \boldsymbol{y})$.

1.5 What is likelihood – simple example

- Suppose data $\boldsymbol{y}$ are discrete and we assume the Geometric distribution as our model, so that $Y_i \sim Geom(\pi)$ with

$$p(y_i; \pi) = (1-\pi)^{y_i}\pi, \qquad \pi \in [0,1], y = 0, 1, 2, \ldots$$

- The likelihood is

$$L(\pi; \boldsymbol{y}) = \prod_{i=1}^{n}(1-\pi)^{y_i}\pi = \pi^n \prod_{i=1}^{n}(1-\pi)^{y_i}$$

- Note that the likelihood is **not a probability distribution** for π but it does represent *"how likely it is that we would get the observed $\boldsymbol{y}$, for different values of π"*.
- Likelihood is the function that links the real world (i.e. data) to the "mathematical" world (i.e. the model). We use it to determine the unknown parameters (π in this example) from what we have observed (the data $\boldsymbol{y}$).

1.5 What is likelihood – coin example

- Suppose data $\boldsymbol{y}$ are binary and represent the outcome of n coin tosses, where $y_i = 1$ is "Heads" while $y_i = 0$ is "Tails".
- Assume the tosses are independent and that the model is a Bernoulli distribution so that $y_i \sim \text{Bern}(\pi)$ and

$$p(y_i, \pi) = \pi^{y_i}(1-\pi)^{1-y_i}, \qquad \pi \in [0,1], y_i \in \{0,1\}.$$

- The likelihood is

$$L(\pi; \boldsymbol{y}) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}.$$

- Let's plot $L(\pi; \boldsymbol{y})$ against π for $n = 1, 2, \ldots, 100$ and see what the function is telling us about π – noting that we know that $\pi = 0.5$.

1.5 Example: flip coin 100 times

1.6 The likelihood 'engine'

- Likelihood theory then says that you should 'fit' the model (i.e. find 'good' estimates for the unknowns $\boldsymbol{\theta}$) by choosing values, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ that maximise the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ (**maximum likelihood estimation**).
- It is much more convenient to do this this by maximising the **log-likelihood** $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \log\left(L(\boldsymbol{\theta}; \boldsymbol{y})\right)$.
- In the Geometric example,

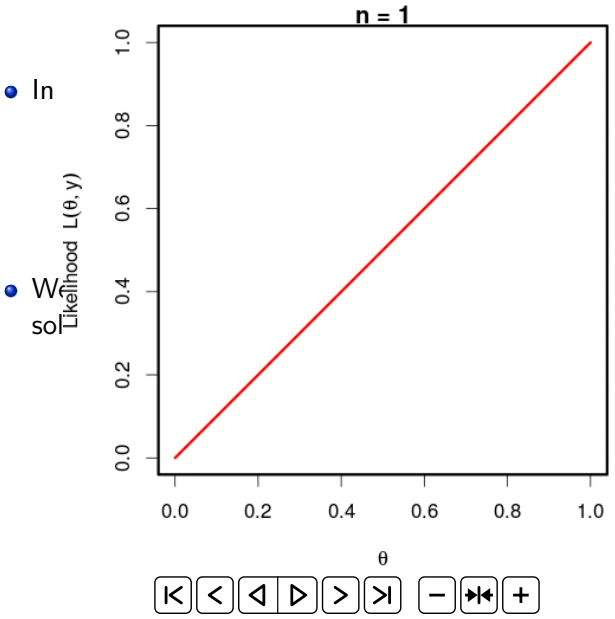
$$\ell(\pi; \boldsymbol{y}) = \log\left(\pi^n \prod_{i=1}^{n} (1-\pi)^{y_i}\right) = n\log(\pi) + \log(1-\pi)\sum_{i=1}^{n} y_i.$$

- We find the maximum by differentiating and setting equal to zero, solving for π:

$$\frac{d\ell(\pi; \boldsymbol{y})}{d\pi} = n/\pi - \frac{\sum_{i=1}^{n} y_i}{1-\pi} = 0 \implies \hat{\pi} = \frac{1}{\sum_{i=1}^{n} y_i/n + 1}.$$

1.6 The likelihood 'engine'

- In
- W
so



$-\sum_{i=1}^{n} y_i)$.

ng equal to zero,

$\hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$.

1.6 The likelihood 'engine'

- For more complex models where $\boldsymbol{\theta}$ is a vector, maximizing $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ implies solution of the system of simultaneous equations:

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \qquad \text{i.e. } \frac{\partial \ell(\theta_1; \boldsymbol{y})}{\partial \theta_1} = 0 \text{ and } \frac{\partial \ell(\theta_2; \boldsymbol{y})}{\partial \theta_2} = 0 \cdots$$

- The first derivative of the log likelihood $\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}}$ is called the **score function**, $\boldsymbol{u}(\boldsymbol{\theta})$. It is a vector of first partial derivatives, one for each element of $\boldsymbol{\theta}$. So the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}$ is found by setting the score to zero, i.e. $\boldsymbol{u}(\hat{\boldsymbol{\theta}}) = 0$ and solving the resulting set of equations.

1.6 The likelihood 'engine'

- The method of maximum likelihood is broadly applicable and provides estimates that have many desirable statistical properties (at least approximately).
- Maximum likelihood estimators are
  - **consistent** (vary less from true value as sample size increases);
  - *asymptotically* **unbiased** (on average equal to true value);
  - *asymptotically* **efficient** (vary less from true value than other estimators for given sample size).
- They also have another desirable property known as **invariance**. This says that if $\hat{\theta}$ is the MLE of θ, but you are interested in some function of θ, e.g. $g(\theta)$, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ e.g. if $\hat{\theta}$ is the MLE of θ, then $\frac{1}{\hat{\theta}}$ is be the MLE of $\frac{1}{\theta}$.

1.6 The likelihood 'engine'

- Once the MLE is derived, the *shape of the likelihood* (in particular 'how peaked' it is at the maximum) can provide a measure of the confidence that can be placed in such estimates (i.e. standard errors and confidence intervals).
- The second derivative of the log-likelihood indicates the extent to which the function is peaked rather than flat. This notion is encapsulated in the **expected information matrix** $\mathcal{I}(\boldsymbol{\theta})$—which is the expected value under the model (over possible data values, $\boldsymbol{y}$) of minus the matrix of second derivatives of the log-likelihood i.e.:

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2 \log\left(L(\boldsymbol{\theta}; \boldsymbol{y})\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = -\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{\theta})\right]$$

where $\boldsymbol{H}(\boldsymbol{\theta})$ is called the **Hessian**—the matrix of second derivatives of the log-likelihood.

1.6 The likelihood 'engine'

- Under reasonable conditions, the MLE $\hat{\boldsymbol{\theta}}$ can be shown to have an asymptotic (multivariate) Normal distribution with variance covariance matrix given by the *inverse of the information matrix*:

$$\hat{\boldsymbol{\theta}} \sim MVN(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta})).$$

- Since $\mathcal{I}^{-1}(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ (which is unknown) we can estimate its value by replacing $\boldsymbol{\theta}$ with its MLE estimate $\hat{\boldsymbol{\theta}}$ so obtaining: $\mathrm{cov}(\hat{\boldsymbol{\theta}}) \approx \mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$.
- Further, since the *expectation* involved in $\mathcal{I}(\boldsymbol{\theta}$ can sometimes be difficult to derive, we may approximate $\mathcal{I}^{-1}(\boldsymbol{\theta})$ by the **observed information matrix**

$$\boldsymbol{I}^{-1}(\boldsymbol{\theta}) = -\boldsymbol{H}^{-1}(\boldsymbol{\theta})|_{\boldsymbol{y}},$$

where the **Hessian** matrix of second derivatives of the log-likelihood is evaluated at the data values $\boldsymbol{y}$ that were actually observed.

1.6 The likelihood 'engine'

- As with $\mathcal{I}^{-1}(\boldsymbol{\theta})$ we estimate the value of $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ by $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})$ (replacing $\boldsymbol{\theta}$ with its MLE).
- So overall we have the useful result that maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ are approximately (multivariate) Normally distributed as:

$$\hat{\boldsymbol{\theta}} \sim MNV(\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})).$$

- Note that this is called the sampling distribution of the MLE and it expresses **estimation uncertainty** – different data from the same process will result in different estimates.
- This provides approximate standard errors, confidence intervals and hypothesis tests for such estimates.

1.7 Hypothesis tests and confidence intervals

- In this module, we will mainly be performing hypothesis tests in two situations. The first is for regression coefficients β_j.
- The sampling distr. of $\hat{\beta}$ (the MLE estimate) is $N\left(\beta, \text{var}\left[\hat{\beta}\right]\right)$ where β is the true value.
- The hypothesis we would like to test is whether β takes a particular value, say $\beta = a$ (usually $a = 0$). The null hypothesis is $H_0 : \beta = a$ to be tested against the alternative hypothesis $H_a : \beta \neq a$.
- If H_0 is true then the test statistic:

$$Z = \frac{\hat{\beta} - a}{\sqrt{\text{var}(\hat{\beta})}} = \frac{\hat{\beta} - a}{\text{se}\left(\hat{\beta}\right)} \sim N(0,1)$$

  is approximately Normal with mean zero and variance 1. If Z is extreme w.r.t. the $N(0,1)$ then we reject H_0.

1.7 Hypothesis tests and confidence intervals

- To establish whether Z is extreme we first need to quantify what we mean by extreme and this is what the significance level does, conventionally taken to be 5%.
- If Z lies within the bulk of the $N(0,1)$, i.e. in the range of values for which the area under the $N(0,1)$ curve is 95% then the value is not extreme and H_0 is accepted.
- The test is two sided since H_a spans both the negative and positive range of β. So we need to consider extremes at both ends of the distribution and for 5% we consider the 2.5% and 97.5% quantiles of the $N(0,1)$, these being -1.96 and 1.96 respectively.
- So, negative values of Z less than -1.96 or positive values of Z greater than 1.96 mean that we reject H_0 at the 5% level. The p-value is the area of the $N(0,1)$ on either the lower or the upper tails depending on the sign of Z. p-values less than 5% indicate Z is extreme w.r.t. $N(0,1)$ so test is rejected.

1.7 Hypothesis tests and confidence intervals

- A 95% confidence interval (CI) for an estimated value (e.g. $\hat{\beta}$ or $\hat{\mu}_i$) that has a Normal sampling distribution can be calculated by

$$\hat{\beta} \pm 1.96\text{se}\left(\hat{\beta}\right)$$

  where the value 1.96 depends on the confidence level.
- Testing whether β takes any particular value a at the 5% level is equivalent to looking at whether a lies inside the 95% CI.

1.8 Model comparison

- Given a fitted model with assumptions that have been assessed as reasonable, an obvious next consideration is how good is it? To what extent does the model 'explain' the data? Are all the covariates necessary or can some be discarded without worsening the fit? Is the model formulation preferable to alternatives?
- One way to consider these questions is via likelihood ratios. Given a specified statistical model, a **saturated model** can always be proposed which fits the observed data perfectly (residuals are zero).
- This is a model which has as many parameters as data values so that predicted values from the model $\hat{y}_i = \hat{\mu}_i = y_i$ are exactly equal to the observed values y_i. Denote the associated likelihood for this model as L_{M_s}.
- Now suppose that some other model M has a likelihood L_M maximised over $p < n$ unknowns, then the **likelihood ratio** $\Lambda = \frac{L_M}{L_{M_s}}$ intuitively measures how well the model M fits relative to the saturated model M_s.

1.8 Model comparison

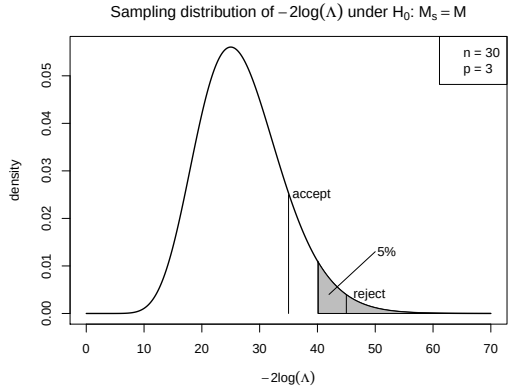
- Likelihood theory tells us that $-2\log\Lambda = 2(\log L_{M_s} - \log L_M)$ has approximately a χ^2_{n-p} distribution if M is as good as M_s. The expected value of a χ^2_q distribution is equal to its d.f. q, and so an 'adequate' model with p unknowns should have $-2\log\Lambda$ roughly equal to $n-p$.
- This is the **likelihood ratio test** (LRT) for comparing models.
- The LRT can also be used to compare **nested** models. Suppose $M_1 \subset M_2$ where M_1 has p_1 unknowns and M_2 has p_2 where $p_2 > p_1$.
- Then the log-likelihood ratio statistic $-2(\log L_{M_1} - \log L_{M_2})$ has approximately a $\chi^2_{p_2-p_1}$ distribution under the hypothesis of no difference in fit between the models and this provides a test of whether the fit of the two models differs significantly.
- Note that models with more parameters will always fit the data better than "smaller" models and LRT quantifies they fit is significantly better.

1.8 Model comparison

- So the LRT is the second type of test we will be using. The log-likelihood ratio statistic will be χ^2_q if our model fits well (if testing for model fit) or if a model is just as good as a bigger one (if comparing nested models).
- So H_0 : "M fits well compared M_S" or H_0 : "M_1 fits as well as M_2". The χ^2_q is bounded below at zero, so we are only looking at the upper tails to judge whether the LRT value is extreme or not.
- Therefore, this is a one sided test where we check whether the value of the LRT is beyond the 95% quantile of the χ^2_q. If it is, then the associated p-value is small (< 0.05) and H_0 is rejected.

1.8 Model comparison

- For instance, suppose we are checking how well our model fits w.r.t. the saturated model M_s, and suppose there are $n = 30$ data points and our model has $p = 3$ parameters.



1.8 Model comparison

- What about comparing non-nested models?
- A familiar problem with comparing 'goodness of fit' of both nested and non-nested models, is that adding a parameter to a model will increase the fit – even if the new parameter represents pure noise.
- So it's also useful to have a general measure of 'goodness of fit' of a model which incorporates a 'trade off' between how well it fits relative to how many parameters have been used (analogous to the adjusted R^2_a used for linear models).
- A most commonly used measure is the **Akaike Information Criterion** (AIC) defined for a model M with p unknown parameters as:

$$\text{AIC} = -2\log(L_M) + 2p$$

- The AIC is an estimator of the **out-of-sample** predictive error, so it can be used for both nested and non-nested models – the smaller the AIC, the better the fit. So a model is preferred if it has a lower AIC.

1.9 Prediction

- A statistical model is basically a probability distribution, so predictions are by definition probabilistic.
- When we seek a single but **meaningful** value the conventional choice is to use the **mean**, with the associated uncertainty represented e.g. by a 95% confidence interval.
- E.g. if $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ then for a particular value say x^* we can predict the response by $\mu = \beta_0 + \beta_1 x^*$. So if x^* is age and Y_i is health outcome, then $\mu = \beta_0 + \beta_1 x^*$ is our prediction of the **average** health outcome for people of age x^*.
- But what about a prediction about a **specific individual**? Then clearly we need to predict Y_i which contains the variability of the individuals.
- Our prediction for an individual is then the probability distribution $\mathrm{N}(\beta_0 + \beta_1 x^*, \sigma^2)$ which we can still summarise using the mean, although the uncertainty is now expressed by the variance σ^2.

1.9 Prediction

- The uncertainty about an individual prediction will always be larger than the one relating to the mean – we are more certain about what happens on average than what happens at one particular occasion.
- Our predictions of Y_i can be obtained by plugging in the MLEs of the model parameters, so predictions are based on $\mathrm{N}(\hat{\beta}_0 + \hat{\beta}_1 x^*, \hat{\sigma}^2)$.
- Note that this **ignores** the uncertainty in the having to estimate the parameters.
- We can only obtain predictions that incorporate both estimation uncertainty and (individual) variability of the response in very few cases
  - one of which being the linear model, where we can have both confidence intervals (estimation uncertainty) and **prediction intervals** (estimation uncertainty plus indiv. variability).

1.9 Simulation

- Computer simulation of random samples from our model is a very powerful way of
  - more robustly doing inference (hyp. tests and conf. intervals);
  - predicting functions of the response variable, i.e. $g(Y_i)$.
- This is also known as **parametric bootstrapping**, and it involves simulating **hypothetical** data sets from our model, e.g. use `rnorm()` in `R` to simulate from a linear model.
- For instance we can compare model M_2 with model $M_1 \subset M_2$ by simulating many data sets of equal length to the original data from our fitted M_2, and fit M_1 to those simulations to see if the likelihood of M_1 is 'close' to M_2.
- Or suppose we want to predict $\exp(Y_i)$. We can then simulate from our fitted model, say $\mathrm{N}(\hat{\beta}_0 + \hat{\beta}_1 x^*, \hat{\sigma}^2)$ and then exponentiate the samples.

1.10 Example of using of the likelihood 'engine'

- Having reviewed the general likelihood 'engine', let's return to the fish net data and try to use this approach. Recall model was of the form:

$$\begin{aligned} Y_i &\sim \mathrm{Bin}(N_i, \pi_i) \qquad (Y_i \text{ indep.}) \\ \mathrm{logit}\,(\pi_i) &= \beta_0 + \beta_1 x_i \quad \left(\text{or} \quad \pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\right) \end{aligned}$$

  where Y_i is number escaping from net out of a total N_i, and x_i is fish length.
- Here $p(y_i; \pi_i, N_i) = \binom{N_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{N_i - y_i}$. So with $\boldsymbol{\theta} = (\beta_0, \beta_1)$:

$$\begin{aligned} L(\beta_0, \beta_1; y_1, \ldots, y_n) &= L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} \binom{N_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{N_i - y_i} \\ \text{or: } \ell(\boldsymbol{\theta}; \boldsymbol{y}) &= \text{const} + \sum y_i \log \pi_i + \sum (N_i - y_i) \log(1-\pi_i) \\ &= \text{const} + \sum \left[y_i \left(\beta_0 + \beta_1 x_i\right) - N_i \log\left(1 + e^{\beta_0+\beta_1 x_i}\right) \right] \end{aligned}$$

1.10 Example of using of the likelihood 'engine'

- Maximising to obtain the MLE $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1)$ we get:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_0} &= \sum \left[y_i - \frac{N_i e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right] = \sum \left[y_i - \frac{N_i}{1+e^{-\beta_0-\beta_1 x_i}} \right] \\ \frac{\partial \ell}{\partial \beta_1} &= \sum \left(y_i x_i - \frac{N_i x_i e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) = \sum \left[y_i x_i - \frac{N_i x_i}{1+e^{-\beta_0-\beta_1 x_i}} \right]\end{aligned}$$

- So to get $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1)$ we need to solve the score equations:

$$\boldsymbol{u}(\hat{\beta}_0, \hat{\beta}_1) = \left(\sum \left[y_i - \frac{N_i}{1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i}} \right], \sum \left[y_i x_i - \frac{N_i x_i}{1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i}} \right] \right) = \mathbf{0}$$

- **So that's the first problem: we cannot do this analytically and we would need to use a numerical approach (e.g. Newton-Raphson)**.

1.10 Example of using of the likelihood 'engine'

- Then to do any inference we need to use $\text{cov}(\hat{\boldsymbol{\theta}}) \approx -\boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}})|_{\boldsymbol{y}}$, where the Hessian is:

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{\theta}) &= \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ell}{\partial^2 \beta_1} \end{pmatrix} \\ -\boldsymbol{H}(\beta_0, \beta_1) &= \begin{pmatrix} \sum \frac{N_i e^{-\beta_0-\beta_1 x_i}}{\left[1+e^{-\beta_0-\beta_1 x_i}\right]^2} & \sum \frac{N_i x_i e^{-\beta_0-\beta_1 x_i}}{\left[1+e^{-\beta_0-\beta_1 x_i}\right]^2} \\ \sum \frac{N_i x_i e^{-\beta_0-\beta_1 x_i}}{\left[1+e^{-\beta_0-\beta_1 x_i}\right]^2} & \sum \frac{N_i x_i^2 e^{-\beta_0-\beta_1 x_i}}{\left[1+e^{-\beta_0-\beta_1 x_i}\right]^2} \end{pmatrix}\end{aligned}$$

- So we need to evaluate the above matrix at the MLE solution $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1)$ then invert it and finally evaluate it at the actual data values $(y_1, \ldots, y_n)$ and $(N_1, \ldots, N_n)$ that we observed in order to get approximate confidence intervals for $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1)$.

- **This is all looking a bit hard and hand crafted to this particular model!**

1.10 Example of using of the likelihood 'engine'

- Finally, what about the 'goodness of fit'? Well we are supposed to be looking at the log-likelihood ratio statistic $-2\log\Lambda = 2\left(\ell_{M_s} - \ell_M\right)$ which intuitively measures how well our model M fits relative to the saturated model M_s. In this case, this quantity should have an approximate χ^2_{n-2} distribution if the model fit is 'acceptable'.

- Here:

$$\ell_M = \text{const} + \sum \left[y_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) - N_i \log \left(1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_i} \right) \right]$$

and

$$\ell_{M_s} = \text{const} + \sum \left[y_i \log \left(\frac{y_i}{N_i} \right) + (N_i - y_i) \log(1 - \frac{y_i}{N_i}) \right]$$

- So $-2\log\Lambda$ is:

$$\sum \left[y_i \log \left(\frac{y_i}{N_i} \right) + (N_i - y_i) \log(1 - \frac{y_i}{N_i}) - y_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) - N_i \log \left(1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_i} \right) \right]$$

- **Again this all looks rather specific to this particular model!**

1.10 Value of general modelling frameworks based on likelihood

- So what are the lessons from this example? Well the general likelihood approach is 'doable with difficulty' in this case and the details are effectively 'hand crafted' to the model.

- For example, if we add an additional explanatory variable then we have to redo the mathematics of the estimation and the inference from scratch (e.g. we have a system of three rather than two simultaneous score equations to solve and the Hessian is a 3×3 matrix rather than 2×2).

- If we changed the distribution of the response to Poisson rather than Binomial and the mean relationship from logistic to logarithmic then the form of the likelihood would change completely and the resulting derivations would need to be worked out again from scratch.

- Different models are not just a question of specific cases of a general framework of equations – each different model is a whole new ball game!

1.10 Value of general modelling frameworks based on likelihood

- The power of the **Normal theory linear model** was that the general likelihood approach led to a fairly simple mathematical framework for estimation and inference in all such models which could then be encapsulated in easy-to-use software functions.

- Clearly this is not going to be the case more generally. We need to find some intermediate position
  - a broader class of models than the Normal theory framework;
  - but sufficiently limited to imply that likelihood results allow the development of flexible software for fitting and inference.

- We can do this through the concept of **Generalised Linear Models** (GLMs) which we consider in the next topic.