Biostatistics (MATH11230)

Estimation in the Logistic Regression Model

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Estimation of logistic regression model parameters

The likelihood function

- Again here, we follow Chapter 13 of Jewell (2003) in most of the exposition.
- Before jumping into the estimation of the parameters of the logistic regression model, we start with a simpler situation, namely the estimation of a proportion.
- Suppose that a random sample of 100 births were drawn from infants born in the USA in 1991 and that in this sample 35 births were associated with unmarried mothers.
- Let p be the unknown population proportion of newborns in 1991 whose mothers were unmarried at the time of birth.

Estimation of logistic regression model parameters

The likelihood function

- We have the following binomial likelihood for p:

$$L(p) = \binom{100}{35} p^{35}(1-p)^{65}.$$

- A sensible way to estimate the parameter p given the data is to maximise the likelihood function, choosing the parameter value (or vector when applicable) that makes the data actually observed as likely as possible.

- Formally, we define the maximum likelihood estimate as the value of the parameter that maximises the likelihood function.

Estimation of logistic regression model parameters

The likelihood function

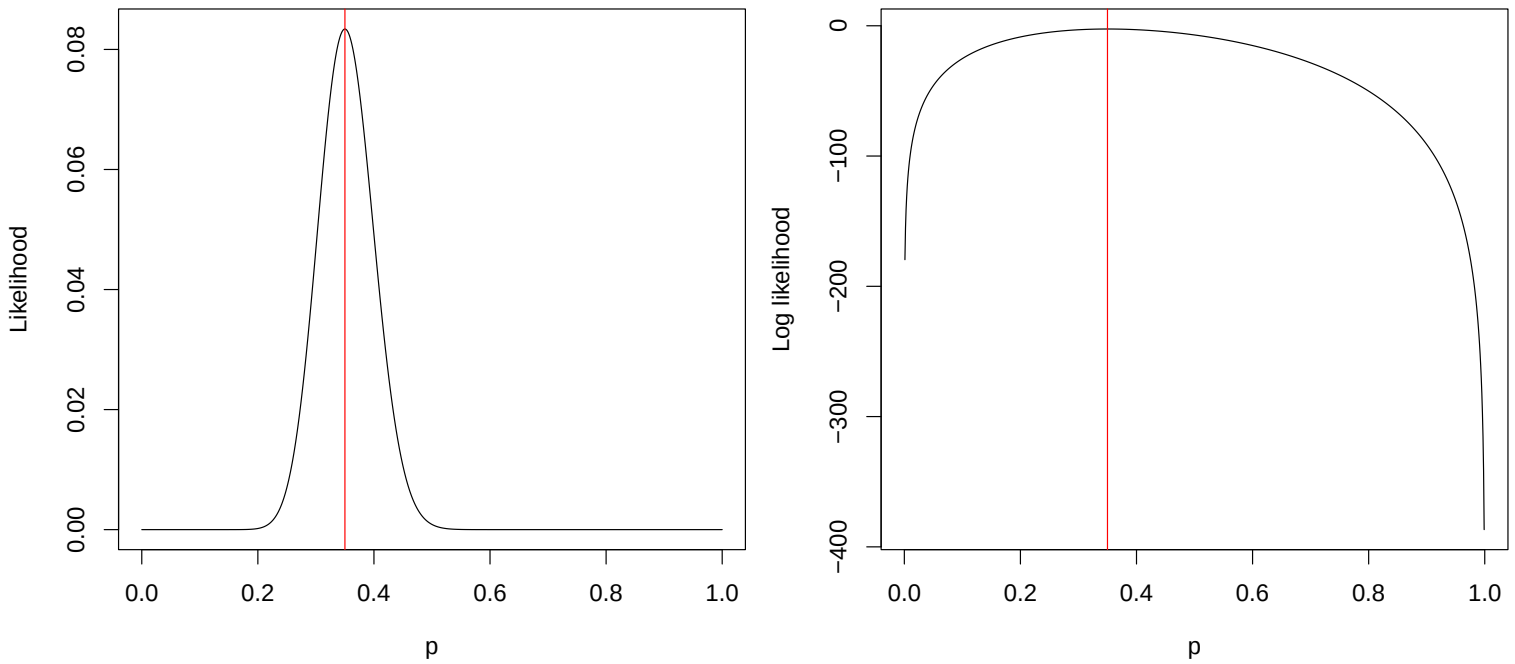
- It is often numerically convenient to use the log likelihood function for computation of the maximum likelihood estimator/estimate.
- Because the logarithm is a strictly increasing function, the value of p that maximises the log likelihood function is also the value that maximises the likelihood function.
- The log likelihood function for our toy example is given by

$$\log L(p) = \log \left\{ \binom{100}{35} \right\} + 35 \log(p) + 65 \log(1-p).$$

- Setting the derivative of the log likelihood to zero and solving for p, we get that the maximum likelihood estimate is $\widehat{p} = 0.35$, which is a reasonable and familiar estimate for the population proportion based on our random sample.

Estimation of logistic regression model parameters

The likelihood function



Estimation of logistic regression model parameters

The likelihood function

- The log likelihood function is extremely useful.
- The log likelihood function is not only useful for finding the maximum likelihood estimator, but it also gives information about the precision of the estimator.
- Qualitatively, this can be seen by considering the shape of the log likelihood function close to where the maximum occurs.
- If the log likelihood climbs (and falls) gently both toward and away from its peak, it means that there are other values of p relatively far from $\widehat{p}$ that gives almost the same value of the likelihood function.
- This suggests that the sampling variability of $\widehat{p}$ is high, reflecting that we must consider a sizeable range of values as plausible estimates of p, that is, that have a similarly high level of likelihood as $\widehat{p}$.

Estimation of logistic regression model parameters

The likelihood function

- In turn, if the log likelihood function has a very sharp peak, meaning that the likelihood drops quickly as p moves away from $\widehat{p}$, then we obtain a much lower value for the sampling variability of $\widehat{p}$.

- This indicates that only values of p in a narrow range around the maximum likelihood estimate give a value of the likelihood close to the maximum, that is, that seem reasonable given the data.

- To sum up, how quickly the slope of the log likelihood function changes near the peak provides direct information on the sampling variability of $\widehat{p}$.

Estimation of logistic regression model parameters

The likelihood function

- Under some regularity conditions, the maximum likelihood estimator has approximately, in large samples, a normal distribution with mean equal to the true parameter p and variance given by the inverse of the observed Fisher information at $\widehat{p}$, i.e.,

$$\widehat{p} \dot{\sim} \mathrm{N}(p, J(\widehat{p})^{-1}).$$

- The observed Fisher information at $\widehat{p}$ is given by

$$J(\widehat{p}) = -\frac{\mathrm{d}^2}{\mathrm{d}p^2} \log L(p) \Big|_{p=\widehat{p}}.$$

- Returning to our running example, we have that $J(\widehat{p})^{-1}$ is given by 0.002275 and therefore we can construct a 95% confidence interval for p as follows:

$$(0.35 - 1.96 \times \sqrt{0.002275}, 0.35 + 1.96 \times \sqrt{0.002275}) = (0.257, 0.443).$$

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

- We now discuss the construction of the likelihood function for data that have been sampled from a population where it is assumed that a logistic regression model holds.

- For simplicity, we assume that we have only a single exposure variable measured by X and that the risk for the outcome D follows the logistic regression model

$$\log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x, \quad p_x = \Pr(D \mid X = x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}.$$

- The ideas can easily be extended to the case where there are several risk factors/exposure variables.

- We now focus on data generated from a population-based or cohort design. We then extend these ideas to data arising from a case-control study.

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

- Let us introduce some notation to facilitate the writing of the likelihood function.
- Let $d \in \{0, 1\}$ be the observed value of D, where $d = 1$ stands for disease (or health outcome of interest) present and $d = 0$ stands for disease (or health outcome of interest) absent.
- The contribution to the likelihood of an individual for which $D = 1$ and $X = x$ is given by

$$\begin{aligned}\Pr(D = 1, X = x) &= \Pr(D = 1 \mid X = x) \Pr(X = x) \\ &= p_x \Pr(X = x) \\ &= \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \Pr(X = x).\end{aligned}$$

- Similarly, the contribution to the likelihood of an individual for which $D = 0$ and $X = x$ is given by

$$\begin{aligned}\Pr(D = 0, X = x) &= \Pr(D = 0 \mid X = x) \Pr(X = x) \\ &= (1 - p_x) \Pr(X = x) \\ &= \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \Pr(X = x).\end{aligned}$$

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

↪ Assuming we observe $\{(x_i, d_i)\}_{i=1}^n$ and that the individual observations are independent of each other, we have that the likelihood for β_0 and β_1 can be written as

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \Pr(D = d_i \mid X = x_i) \Pr(X = x_i) \\ &= \prod_{i=1}^{n} \left\{ [p_{x_i} \Pr(X = x_i)]^{d_i} [(1 - p_{x_i}) \Pr(X = x_i)]^{1-d_i} \right\}. \end{aligned}$$

↪ The terms involving $P(X = x_i)$ do not depend on β_0 and β_1 and so can be ignored when maximizing the likelihood:

$$\begin{aligned} L(\beta_0, \beta_1) &\propto \prod_{i=1}^{n} \left\{ p_{x_i}^{d_i} (1 - p_{x_i})^{1-d_i} \right\} \\ &= \prod_{i=1}^{n} \left\{ \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{d_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-d_i} \right\}. \end{aligned}$$

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

↪ The corresponding log likelihood function is given by

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left\{ d_i \log \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) + (1-d_i) \log \left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}} \right) \right\}$$

$$= \sum_{i=1}^{n} \{ d_i(\beta_0 + \beta_1 x_i) - \log(1+e^{\beta_0+\beta_1 x_i}) \}$$

↪ Setting the partial derivatives with respect to β_0 and β_1 to zero gives rise to a system of nonlinear equations in these two parameters that can only be solved numerically.

↪ However, standard iterative numerical techniques can quickly locate this maximum, yielding the maximum likelihood estimates of β_0 and β_1, say $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

- Confidence intervals for β_0 and β_1 (or $\beta_0, \beta_1, \ldots, \beta_k$ in the case of a multiple logistic regression model) can be derived from the fact that maximum likelihood estimates have an approximately Normal sampling distribution when the overall sample size is large (or both sample sizes are large for cohort or case-control designs).

- Letting $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)'$ it then holds that

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}_2(\boldsymbol{\beta}, J(\widehat{\boldsymbol{\beta}})^{-1}),$$

$$J(\widehat{\boldsymbol{\beta}}) = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log L(\boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}.$$

- Note that here $J(\widehat{\boldsymbol{\beta}})$ is the observed Fisher information matrix and N_2 stands for the bivariate normal distribution.

- Remember that, earlier in weeks 2/3, when estimating confidence intervals for the odds ratio, we have used the log transformation.

- In the case of logistic regression, the coefficients are already on the log scale, obviating the need for a transformation.

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

- A two sided $100(1-\alpha)\%$ confidence interval for β_1 (a similar interval holds for any other coefficient) is given by

$$\left(\widehat{\beta}_1 - z_{\alpha/2}\widehat{\mathrm{SE}}(\widehat{\beta}_1), \widehat{\beta}_1 + z_{\alpha/2}\widehat{\mathrm{SE}}(\widehat{\beta}_1)\right),$$

  where $\widehat{\mathrm{SE}}(\widehat{\beta}_1)^2$ is the variance of the maximum likelihood estimate $\widehat{\beta}_1$ and can be extracted from the inverse of the observed Fisher information matrix.

- Because β_1 represents the log odds ratio associated with a unit increase in X, it is preferable to report this in terms of e^{β_1}, the odds ratio for a unit increase in X.

- This is trivially obtained by exponentiating the limits of the above confidence interval.

Estimation of logistic regression model parameters

The likelihood function based on a (simple) logistic regression model

- To illustrate these calculations, we consider fitting the simple logistic regression model to the full sample from the Western Collaborative Group Study data we have previously 'analysed', with the risk factor X representing body weight (in lb), measured on a continuous scale. The outcome D is the presence of CHD.

- There is no grouping now imposed on body weight; in fact, there are 124 distinct values of body weight observed amongst the 3154 sampled individuals, ranging from 78 to 320 lb.

- In the Supplementary Materials file we show how to obtain the maximum likelihood estimates of β_0 and β_1. These are: $\widehat{\beta}_0 = -4.215$ and $\widehat{\beta}_1 = 0.010$.

- Note that here would make sense to, at least, recentre the weight!

Estimation of logistic regression model parameters

Logistic regression with case-control data

- Let us now suppose we use a case-control design to study a population for which we assume a logistic regression model describes how the risk of disease depends on an exposure variable X.

- Again, for simplicity, we suppose that there is only one exposure variable as this illustrates both the issues and solution to the analysis of data arising from a case-control study.

- At a first glance, this sampling method appears to be problematic because conditional probabilities such as $\Pr(D \mid X = x)$ cannot be calculated from case-control data due to the design's manipulation of the frequency of cases and controls.

- Remember that for a binary exposure, the impact of case-control sampling turned out not to matter because of the properties of the odds ratio.

Estimation of logistic regression model parameters

Logistic regression with case-control data

- ↪ Case-control sampling leads to a distortion of the relative frequency of cases and controls in the sample at each level of the exposure X.
- ↪ Let us look more closely at the distortion introduced by case-control sampling.
- ↪ Let π_{case} denote the probability of an individual being sampled given that they are a case, i.e., $\pi_{\text{case}} = \Pr(\text{sampled} \mid D)$.
- ↪ Similarly, define $\pi_{\text{control}} = \Pr(\text{sampled} \mid \text{not } D)$, i.e., the probability of an individual being sampled given that they are a control.
- ↪ In case-control sampling these selection probabilities, or sampling fractions, are not related to the exposure level; that is, they do not depend on X.

Estimation of logistic regression model parameters

Logistic regression with case-control data

- The subpopulation who share exposure level $X = x$ can be divided into four distinct groups based on (i) their disease status and (ii) whether they are included in the sample or not.

- We can also measure the relative frequency of these groups in this subpopulation.

- For example, the probability of being a case and being sampled, amongst individuals with exposure $X = x$ is:

$$\begin{aligned}\Pr(D\& \text{ sampled} \mid X = x) &= \Pr(D \mid X = x)\Pr(\text{sampled} \mid D, X = x) \\ &= p_x \pi_{\text{case}}.\end{aligned}$$

- Note that because in a case-control design the individuals are sampled irrespectively of their exposure status, we have that $\Pr(\text{sampled} \mid D, X = x) = \Pr(\text{sampled} \mid D) = \pi_{\text{case}}$.

Estimation of logistic regression model parameters

Logistic regression with case-control data

- The four groups and their respective proportions within the subpopulation with exposure $X = x$ are then:
  - Individual is a case and is sampled: proportion = $\pi_{\text{case}} p_x$.
  - Individual is a case and is not sampled: proportion = $(1 - \pi_{\text{case}}) p_x$.
  - Individual is a control and is sampled: proportion = $\pi_{\text{control}}(1 - p_x)$.
  - Individual is a control and is not sampled: proportion = $(1 - \pi_{\text{control}})(1 - p_x)$.
- Obviously, we only observe sampled individuals. The relative probability of being a case at exposure level $X = x$ and conditional on being sampled is simply

$$\Pr(D \mid X = x, \text{ sampled}) = \frac{\pi_{\text{case}} p_x}{\pi_{\text{case}} p_x + \pi_{\text{control}}(1 - p_x)}.$$

Estimation of logistic regression model parameters

Logistic regression with case-control data

- The expression in the previous slide indicates directly how case control sampling modifies the 'apparent' risk at exposure level $X = x$, as $\Pr(D \mid X = x, \text{ sampled}) \neq p_x$, unless $\pi_{\text{case}} = \pi_{\text{control}}$, that is, unless cases and controls are sampled at the same frequency.

- We also have that

$$1 - \Pr(D \mid X = x, \text{ sampled}) = \frac{\pi_{\text{control}}(1 - p_x)}{\pi_{\text{case}} p_x + \pi_{\text{control}}(1 - p_x)}.$$

- Recall that in the population we assume the logistic regression model to hold.

Estimation of logistic regression model parameters

Logistic regression with case-control data

↪ Thus, for sampled individuals we have

$$\begin{aligned}
\log\left(\frac{\Pr(D \mid X = x,\ \text{sampled}}{1 - \Pr(D \mid X = x,\ \text{sampled}}\right) &= \log\left(\frac{\pi_{\text{case}} p_x}{\pi_{\text{control}}(1 - p_x)}\right) \\
&= \log\left(\frac{\pi_{\text{case}}}{\pi_{\text{control}}}\right) + \log\left(\frac{p_x}{1 - p_x}\right) \\
&= \log\left(\frac{\pi_{\text{case}}}{\pi_{\text{control}}}\right) + \beta_0 + \beta_1 x \\
&= \beta_0^* + \beta_1 x,
\end{aligned}$$

where $\beta_0^* = \log(\pi_{\text{case}}/\pi_{\text{control}}) + \beta_0$.

Estimation of logistic regression model parameters

Logistic regression with case-control data

- ↪ The case-control sampling has caused the log odds of D in the sample to differ from what it is in the population.
- ↪ Generally, $\pi_{\text{case}}/\pi_{\text{control}}$ would not be known, and so we are not able to estimate β_0.
- ↪ However, only the intercept changes, the slope term β_1 is exactly the same.
- ↪ The fact that the slope coefficient β_1 is not affected by case-control sampling is an extension of what we already observed about the odds ratio, namely, that this measure is not distorted by case-control sampling.
- ↪ We can therefore apply the same estimation technique we used for population-based and cohort designs, and hypotheses and parameters can be investigated as long as they do not directly involve the intercept term β_0.
- ↪ Because we are often most interested in relative comparisons, such as odds ratios, this restriction is 'still manageable'.