

Maximum Likelihood Inference:  
Principles, Methods, & Examples using Logistic Regression

([gustavoc@msu.edu](mailto:gustavoc@msu.edu))

Maximum Likelihood (ML) is one of the most commonly used estimation methods in statistics. The ML framework, originally developed by R.A. Fisher, is a very general procedure that can be used for estimation, inference, and hypothesis testing in almost any problem for which we can propose a probability model for the conditional distribution of the data. Point estimates are derived by maximizing the likelihood function, which is the joint distribution of the data, viewed as a function of the model parameters. A very elegant theory provides the basis for (large sample) inference. Under regularity conditions, in large samples, ML estimates follow multivariate normal distributions with mean equal to the true parameter values (i.e., ML estimates are asymptotically unbiased) and a sampling (co)variance matrix equal to the inverse of Fisher's information. This provides a framework for inference, including the derivation of SEs, confidence intervals, as well as for hypothesis testing. Reviewing ML theory is beyond the scope of the course and these notes. Good references for that are *Theory of Point Estimation* (by E.L. Lehmann and G. Casella) and *In all likelihood: statistical modelling and inference using likelihood* (by Y. Pawitan).

In this note, I illustrate maximum likelihood estimation, inference, and hypothesis testing using Logistic Regression; however, as noted, likelihood methods can be used with any problem for which we can propose a model for the sampling distribution of the data.

### **Bernoulli model**

Many outcomes of interest are binary, implying that they can take two values (say, 0/1). Disease is a typical example of this.

Binary random variables follow Bernoulli distributions:  $p(Y_i = 1) = \theta$  or  $p(Y_i = 0) = 1 - \theta$ , or,

$$p(Y_i = y_i | \theta) = \theta^{y_i} (1 - \theta)^{1-y_i}$$

(Note: above,  $Y_i$  denotes the random variable and  $y_i$  represents the realized value)

The odds of success are defined as  $\frac{p(Y_i=1)}{p(Y_i=0)} = \frac{\theta}{1-\theta}$ .

### **Maximum likelihood estimation of the success probability in the IID Bernoulli model**

The likelihood function is the joint probability of the data given the parameters, evaluated at the observed values of the data  $S = \{Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n\}$  viewed as a function of the

parameters ( $\theta$ ). In the case of a random sample, the joint probability of the data is simply the product of the probability of each of the data points, thus

$$\begin{aligned} p(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \theta) &= p(Y_1 | \theta) \times p(Y_2 | \theta) \times \dots \times p(Y_n | \theta) \\ &= \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \\ &= \theta^{\sum_i y_i} (1 - \theta)^{\sum_i 1-y_i} = \theta^{n\bar{y}} (1 - \theta)^{n(1-\bar{y})} \end{aligned}$$

Thus, the likelihood function is

$$L(\theta | y_1, \dots, y_n) = \theta^{n\bar{y}} (1 - \theta)^{n(1-\bar{y})}$$

The Maximum Likelihood estimator (MLE) is obtained by maximizing  $L(\theta | y_1, \dots, y_n)$  with respect to  $\theta$ ; the same estimate can be obtained by maximizing the log-likelihood

$$l(\theta | y_1, \dots, y_n) = \log\{L(\theta | y_1, \dots, y_n)\} = n\bar{y} \log(\theta) + n(1 - \bar{y}) \log(1 - \theta)$$

Differentiating with respect to  $\theta$  we get

$$\frac{dl(\theta | y_1, \dots, y_n)}{d\theta} = \frac{n\bar{y}}{\theta} - \frac{n(1 - \bar{y})}{(1 - \theta)}$$

Setting the derivative equal to zero we get maximum the MLE

$$\begin{aligned} \frac{n\bar{y}}{\hat{\theta}} &= \frac{n(1 - \bar{y})}{(1 - \hat{\theta})} \\ \frac{\bar{y}}{(1 - \bar{y})} &= \frac{\hat{\theta}}{(1 - \hat{\theta})} \quad (\text{assuming } \bar{y} \neq 1) \\ \hat{\theta} &= \bar{y} \end{aligned}$$

Thus, the MLE of the success probability is simply the sample mean of the data, which is not surprising considering that  $E[Y_i] = \theta$ .

## Logistic Regression

We are often interested on learning the effects of some factors (e.g., sex) and covariates (e.g., age) on the probability of a binary outcome ( $\theta$ , e.g., a disease probability). In the previous example this probability was assumed to be the same for all individuals. To model effects of covariates on  $\theta$ , in logistic regression, we make  $\theta$  a function of covariates.

Since  $\theta \in [0,1]$  we cannot model  $\theta$  directly using linear regression because a linear function can take any value in the real line. To deal with this problem we introduce a “link” function (e.g., probit, logit). A link function maps from the real line onto the  $[0,1]$ . The most commonly used link is the logit which is the logarithm of the odds of success, that is:  $\log\left(\frac{\theta_i}{1-\theta_i}\right)$ . This function can take values in the real line, thus, we can model the logit using linear methods

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mu + X_{i1}\beta_1 + \dots + X_{ip}\beta_p. \quad [1]$$

Note that the above regression is a regression for the probability, not for the data, thus, it typically does not include an error term (in some over-dispersed models it may contain an error).

### From regression to probabilities

Solving [1] for  $\theta_i$  gives

$$\theta_i = \frac{\exp\{\mu + X_{i1}\beta_1 + \dots + X_{ip}\beta_p\}}{1 + \exp\{\mu + X_{i1}\beta_1 + \dots + X_{ip}\beta_p\}}. \quad [2]$$

Letting the right-hand side of [1], i.e., the regression function, be  $\eta_i = \mu + X_{i1}\beta_1 + \dots + X_{ip}\beta_p$  then we have:  $\theta_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$ .

**Odd-ratios:** From expression [1] we have that  $\frac{\theta_i}{1-\theta_i} = \exp\{\mu + X_{i1}\beta_1 + X_{i2}\beta_2 + \dots + X_{ip}\beta_p\}$ .

Suppose that  $X_{i1}$  is a dummy variable defining a group (e.g., treatment,  $X_{i1} = 1$  versus control,  $X_{i1} = 0$ ). The odds of success for treatment and controls are:

Treatment ( $X_{i1} = 1$ ):  $\exp\{\mu + \beta_1 + X_{i2}\beta_2 + \dots + X_{ip}\beta_p\}$ , and  
Control ( $X_{i1} = 0$ ):  $\exp\{\mu + X_{i2}\beta_2 + \dots + X_{ip}\beta_p\}$ , respectively.

Therefore, the treatment/control odds-ratio is:

$$\frac{\left[\frac{\theta_i}{1-\theta_i} \mid \text{Treatment}\right]}{\left[\frac{\theta_i}{1-\theta_i} \mid \text{Control}\right]} = \frac{\exp\{\mu + \beta_1 + X_{i2}\beta_2 + \dots + X_{ip}\beta_p\}}{\exp\{\mu + X_{i2}\beta_2 + \dots + X_{ip}\beta_p\}} = \exp\{\beta_1\}$$

The above result provides a clear interpretation of  $\exp\{\beta_1\}$  in terms of odds ratios for coefficients linked to dummy variables.

## Likelihood function for the logistic regression model

The likelihood function is the probability of the data given the parameters. As before, we will assume conditional independence, meaning that

$$p(Y_1, Y_2, \dots, Y_n | \mu, \beta_1, \dots, \beta_p, X) = p(Y_1 | \mu, \beta_1, \dots, \beta_p, X) \times p(Y_2 | \mu, \beta_1, \dots, \beta_p, X) \times \dots \times p(Y_n | \mu, \beta_1, \dots, \beta_p, X)$$

The probability of the  $i^{th}$  data-point is:

$$p(Y_i = 1) = \theta_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} ; p(Y_i = 0) = 1 - \theta_i = 1 - \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{\eta_i}}$$

$$\text{or, } p(Y_i = y_i) = \left[ \frac{e^{\eta_i}}{1+e^{\eta_i}} \right]^{y_i} \left[ \frac{1}{1+e^{\eta_i}} \right]^{1-y_i}$$

Therefore, assuming conditional independence, the joint likelihood becomes

$$\textbf{Likelihood: } L(\mu, \beta_1, \dots, \beta_p | Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n \left[ \frac{e^{\eta_i}}{1+e^{\eta_i}} \right]^{y_i} \left[ \frac{1}{1+e^{\eta_i}} \right]^{1-y_i} \quad [3]$$

Note that above: (i)  $y_i$  is a realized value of the corresponding Bernoulli random variable ( $Y_i$ ), therefore,  $y_i$  can take values either 0 or 1. (ii)  $\eta_i = \mu + X_{i1}\beta_1 + \dots + X_{ip}\beta_p$  is a function of both covariates ( $X_{ij}, j = 1, \dots, p$ ) and parameters ( $\mu, \beta_j$ ).

Therefore, the log-likelihood function is

$$l(\mu, \beta_1, \dots, \beta_p | y_1, y_2, \dots, y_n) = \sum_{i=1}^n y_i \log(\theta_i) + (1 - y_i) \log(1 - \theta_i) \quad [4]$$

$$\text{where } \theta_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \text{ and } 1 - \theta_i = \frac{1}{1+e^{\eta_i}}.$$

The entry [logisticRegression.md](#) in our gitHub repository implements this function in R.

## Maximum Likelihood estimation

Maximum likelihood estimates are obtained by maximizing [4] with respect to the parameters ( $\mu, \beta_1, \dots, \beta_p$ ). The function `glm` in R fits logistic regression via maximum likelihood. We can also fit a logistic regression using a general-purpose optimization algorithm (e.g., `optim` in R). The entry [logisticRegression.md](#) in our gitHub repository shows how to fit logistic regression using `glm` and `optim`.

## Inference

Under regularity conditions, in large samples, maximum likelihood estimates (MLEs) follow multivariate normal distribution with mean equal to the true parameter values (i.e., MLEs are asymptotically unbiased<sup>1</sup>) and variance-covariance matrix equal to the inverse of Fisher's Information Matrix, that is

$$\text{Cov}(\hat{\theta}) = I(\theta)^{-1} = \left[ E \left\{ \frac{\partial -\log \text{Lik}(\theta)}{\partial \theta \partial \theta'} \right\} \right]^{-1}$$

In practice we approximate Fisher's Information matrix with the Observed Information matrix, which is the matrix of 2<sup>nd</sup> order derivatives of the negative-log-likelihood evaluated at the MLE (H=Hessian):

$$\text{Cov}(\hat{\theta}) = V \approx \left[ \left\{ \frac{\partial -\log \text{Lik}(\theta)}{\partial \theta \partial \theta'} \right\}_{\theta=\hat{\theta}} \right]^{-1} = H^{-1}$$

Where, above,  $H = \left\{ \frac{\partial -\log \text{Lik}(\theta)}{\partial \theta \partial \theta'} \right\}_{\theta=\hat{\theta}}$ . We can obtain this matrix by numerical evaluation of the 2<sup>nd</sup> order derivatives of the log-likelihood (aka the Hessian matrix) at the MLE. We present examples of this in the entry [logisticRegression.md](#).

### Standard errors and single-parameter confidence intervals

Once we evaluate  $V$ , standard errors for each of the parameters can be obtained by taking the square-root of the diagonal elements of  $V$ , that is  $SE(\hat{\theta}_j) = \sqrt{V_{jj}}$ ; and, assuming normality, we can form a CI with an approximate coverage of  $1-\alpha$  using

$$qnorm(\text{mean} = \hat{\theta}_j, \text{sd} = \sqrt{V_{jj}}, \text{prob} = c(\alpha/2, 1 - \alpha/2)).$$

### Hypothesis testing

For **1-DF test**, in large samples, we can form a z-statistic by taking the ratio between the parameter estimate and the SE  $z_j = \hat{\theta}_j / \sqrt{V_{jj}}$ ; then, a two-sided p-value for the test  $H_0: \theta_j = 0$  Vs  $H_0: \theta_j \neq 0$ , can be obtained using

$$2 \times pnorm(\text{mean} = 0, \text{sd} = 1, q = \text{abs}(z_j), \text{lower.tail} = \text{FALSE}).$$

For tests involving more than 1DF we can use Wald's test or Likelihood Ratio tests.

---

<sup>1</sup> Note that in finite samples ML estimates are not always unbiased; however, under regularity conditions, if the estimator is bias, its bias diminishes with sample size.

## Wald's test

To implement Wald's test, we follow these steps:

- Let  $\theta$  be the parameter vector under  $H_a$ ,
- Write the null hypothesis in linear form,  $T\theta = 0$ ,
- Above,  $T$  is a matrix of contrasts (the number of rows is the number of restrictions imposed by  $H_0$  on  $H_a$ ).
- Compute Wald's test-statistic:  $W = \hat{\theta}'[TVT']^{-1}\hat{\theta}$ , where  $\hat{\theta}$  is the MLE estimate from  $H_a$  and  $V$  is the sampling (co)variance matrix of  $\hat{\theta}$ .
- Compute a p-value using `pchisq(q=W, lower.tail=FALSE, df=nrow(T))`.

## Likelihood ratio test

For likelihood ratio test, we fit the model under  $H_0$  and  $H_a$ , for each of the fitted models we evaluate the log-likelihood at the ML estimates. Then, we evaluate the following Likelihood-Ratio-Test statistic

$$LRT = -2\{\log\text{Lik}(H_0) - \log\text{Lik}(H_a)\}$$

In large samples, under the null, and assuming regularity conditions, the  $LRT \sim \chi^2_{df}$  where here  $df$  is the number of restrictions imposed by  $H_0$ , for standard nested models, this is the difference in the number of parameters between the null and the alternative.