Maximum likelihood estimation

Patrick Breheny

January 29

Introduction

- Generalized linear models – indeed, the vast majority of statistical methods that entered widespread use in the 20th century – are based on the idea of *likelihood*
- The idea of likelihood has a long history in statistics, but the formal, rigorous study of likelihoods and their properties as the foundation for inference is largely due to the work of R. A. Fisher in the 1920s
- Before we can study generalized linear models and their properties, we need to establish a basic foundation in likelihood theory

Definition

- Let $f(\mathbf{x}|\theta)$ denote the pdf of the sample $(X_1, X_2, \ldots, X_n)$. Then, given that sample $\mathbf{x}$ is observed, the *likelihood function* of θ is defined to be

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

- On the surface of things, this may seem uninteresting – the likelihood function is just the same thing as the pdf
- It is crucial to keep in mind, however, that $f(\mathbf{x}|\theta)$ is a function of $\mathbf{x}$ with θ known, while $L(\theta|\mathbf{x})$ is a function of θ with $\mathbf{x}$ known

Interpretation

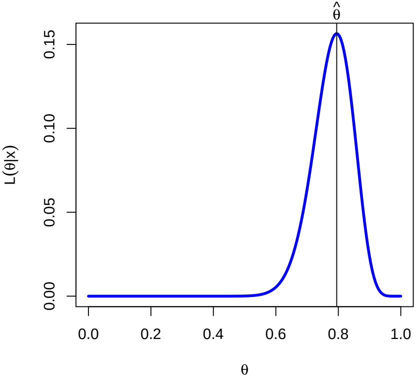
- Thus, while $f(\mathbf{x}|\theta)$ measures how probable various values of $\mathbf{x}$ are for a given value of θ, $L(\theta|\mathbf{x})$ measures *how likely the sample we observed was* for various values of θ
- So, if $L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x})$, this suggests that it is more plausible, in light of the data we have gathered, that the true value of θ is θ_1 than it is that θ_2 is the true value
- Of course, we need to address the question of how meaningful a given difference in likelihoods is, but certainly it seems reasonable to ask how likely it is that we would have collected the data we did for various values of the unknown parameter θ

Maximum likelihood estimation

- Perhaps the most basic question is: which value of θ maximizes $L(\theta|\mathbf{x})$?
- This is known as the *maximum likelihood estimator*, or MLE, and typically abbreviated $\hat{\theta}$ (or $\hat{\theta}_{\mathrm{MLE}}$ if there are multiple estimators we need to distinguish between)
- Provided that the likelihood function is differentiable and unimodal, we can obtain the MLE by taking the derivative of the likelihood and setting it equal to 0

MLE: Example

- For example, suppose X follows a binomial distribution with n trials and probability of success θ
- **Exercise:** Then $\hat{\theta} = x/n$, the sample proportion



Log-likelihood

- Note that, for identically and independently distributed (iid) data, we have:

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

- This tends to be difficult to take the derivative of, as it requires extensive use of the product rule
- An alternative that is almost always simpler to work with is to maximize the log of the likelihood, or *log-likelihood instead*:

$$\ell(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log\{f(x_i|\theta)\}$$

MLE: Example

- Note that the binomial example from before is a bit easier when working with the log-likelihood
- For other distributions, such as the normal, it is much easier than working with the likelihood directly
- **Exercise**: For the normal distribution with mean θ, $\hat{\theta} = \bar{x}$ regardless of the value of σ^2

Definition

- Given that so much of maximum likelihood estimation revolves around (a) working with the log-likelihood and (b) taking derivatives, it is perhaps unsurprising that the derivative of the log-likehood is given its own name: the *score*
- Formally, the score, commonly denoted U, is defined as

$$U_X(\theta) = \frac{d}{d\theta}\ell(\theta|X);$$

  note that U is a random variable, as it depends on X, and is also a function of θ – these will often be dropped for the sake of convenience in the notation
- Note that with iid data, the score of the entire sample is the sum of the scores for the individual observations:

$$U = \sum_i U_i$$

Role of the score in maximum likelihood estimation

- Thinking about maximum likelihood estimation in terms of the score, we see that the MLE is found by setting the sum of the observed scores equal to 0:

$$\sum_i U_i(\hat{\theta}) = 0$$

- **Exercise**: For the normal distribution,

$$\begin{aligned} U_i &= \frac{X_i - \theta}{\sigma^2} \\ \implies \hat{\theta} &= \bar{x} \end{aligned}$$

- **Exercise**: For the Poisson distribution,

$$\begin{aligned} U_i &= \frac{X_i}{\theta} - 1 \\ \implies \hat{\theta} &= \bar{x} \end{aligned}$$

Mean

- We now turn our attention to the theoretical properties of the score
- It is worth noting that there are some regularity conditions that $f(x|\theta)$ must meet in order for these theorems to work; for the purposes of this class we will assume that we are working with a distribution for which these hold (this is the case for the vast majority of common situations, although important exceptions do arise)
- **Theorem**: $\mathrm{E}(U) = 0$

Variance

- The variance of U is given a special name in statistics: it is called the *Fisher information* or simply the *information*
- In this class we will denote the information with a V to emphasize this relationship
- Like the score, the information is a function of θ, although unlike the score, it is not random, as the random variable X has been integrated out
- **Exercise**: For the normal distribution, $V = \frac{1}{\sigma^2}$

Information

- The name "information" is apt: the amount of information that each observation from a normal distribution contains is inversely proportional to how noisy the data is
- As another example, note that $V = \sum_i V_i$: for iid data, each observation contains the same amount of information, and they add up to the total information in the sample
- For a sample from the normal distribution, $V = \frac{n}{\sigma^2}$
- Note that the expression V_i is perhaps a bit strange, in that V is not random and thus does not vary from observation to observation; nevertheless we will continue to use V_i from time to time when it is necessary to distinguish the total information from the information in a single observation

Another information identity

- Another property of scores which is often useful is the following:

$$V = -\mathrm{E}\{U'\}$$

- **Exercise**: For the normal distribution, $U_i' = -1/\sigma^2$
- **Exercise**: For the Poisson distribution, $U_i' = -X/\theta^2$

Asymptotic distribution

One final, very important theoretical result for the score may be obtained by applying the central limit theorem:

$$\sqrt{n}\{\bar{U} - \mathrm{E}(U)\} \xrightarrow{\mathsf{d}} N(0, V_i),$$

or equivalently,

$$\frac{1}{\sqrt{n}} U \xrightarrow{\mathsf{d}} N(0, V_i),$$

where the expression $\xrightarrow{\mathsf{d}}$ means that the quantity on the left "converges in distribution" to the distribution on the right as the sample size n goes to ∞

Asymptotic distribution (cont'd)

- Other classes will cover the formal meaning of various kinds of asymptotic convergence
- For this class, we may simply take the result on the previous slide to mean that

  $$U \dot{\sim} N(0, V),$$

  where $\dot{\sim}$ means "approximately distributed as"; this approximation may be poor for small n, but will get better as n gets larger

Multiple parameters

- The preceding results all describe a situation in which we are interested in a single parameter θ
- It is often the case (and always the case in regression modeling) that $f(x)$ depends on multiple parameters
- All of the preceding results can be extended to the case where we are interested in a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$

Multivariate extensions

- The score is now defined as

$$U(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}|\mathbf{x}),$$

  where $\nabla \ell(\boldsymbol{\theta}|\mathbf{x})$ is the *gradient* of the log-likelihood, and has elements $\frac{\partial}{\partial \theta_1}\ell(\boldsymbol{\theta}|\mathbf{x}), \frac{\partial}{\partial \theta_2}\ell(\boldsymbol{\theta}|\mathbf{x}), \ldots$
- Note that the score is now a $p \times 1$ vector; to denote this I will often write the score vector as $\mathbf{u}$
- The MLE is now found by setting each component of the score vector equal to zero; *i.e.*, solving the linear system of equations $\mathbf{u} = \mathbf{0}$, where $\mathbf{0}$ is a $p \times 1$ vector of zeros

Multivariate extensions (cont'd)

- The score still has mean zero: $\mathrm{E}(\mathbf{u}) = \mathbf{0}$
- The variance of the score is still the information, $\mathrm{Var}(\mathbf{u}) = \mathbf{V}$, although the information $\mathbf{V}$ is now a $p \times p$ covariance matrix
- It is still true that, for iid data, $\mathbf{u} = \sum_i \mathbf{u}_i$ and $\mathbf{V} = \sum_i \mathbf{V}_i$
- We again have that $\mathbf{V} = -\mathrm{E}(\nabla \mathbf{u})$, where $\nabla \mathbf{u}$ is a $p \times p$ matrix of second derivatives with i, jth element $\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta} | \mathbf{x})$; this matrix is sometimes referred to as the *Hessian* matrix
- Finally, it is still true that $\mathbf{u} \dot{\sim} N(0, \mathbf{V})$