Maximum Likelihood Estimation

Fall 2017

Outline

Maximum Likelihood Estimation (MLE)

- Why use maximum likelihood estimation?
  - General purpose tool - works in many situations (data can be censored, truncated, include covariates, time-dependent, and so forth)
  - It is "optimal," the best, in the sense that it has the smallest variance among the class of all unbiased estimators. (Caveat: for large sample sizes)
- A drawback: Generally, maximum likelihood estimators are computed iteratively, no closed-form solution
  - For example, you may recall a "Newton-Raphson" iterative algorithm from calculus
  - Iterative algorithms require starting values. For some problems, the choice of a close starting value is critical

Likelihood and Log-Likelihood Functions

- Let $\mathrm{f}(\cdot;\boldsymbol{\theta})$ be the probability mass function if X is discrete or the probability density function if it is continuous
- The likelihood is a function of the parameters ($\boldsymbol{\theta}$) with the data ($\mathbf{x}$) fixed rather than a function of the data with the parameters fixed
- Define the *likelihood function*,

$$L(\boldsymbol{\theta}) = L(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathrm{f}(x_i;\boldsymbol{\theta}),$$

  evaluated at a realization $\mathbf{x}$
- Define the *log-likelihood function*,

$$l(\boldsymbol{\theta}) = l(\mathbf{x};\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \mathrm{f}(x_i;\boldsymbol{\theta}),$$

  evaluated at a realization $\mathbf{x}$
- In the case of independence, the joint density function can be expressed as a product of the marginal density functions and, by taking logarithms, we can work with sums

Example: Pareto Distribution

- Suppose that $X_1, \ldots, X_n$ represent a random sample from a single-parameter Pareto with cumulative distribution function:

$$\mathrm{F}(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500$$

- In this case, the single parameter is $\theta = \alpha$
- The corresponding probability density function is $\mathrm{f}(x) = 500^{\alpha}\alpha x^{-\alpha-1}$ and the logarithmic likelihood is

$$l(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln \mathrm{f}(x_i; \alpha) = n\alpha \ln 500 + n \ln \alpha - (\alpha + 1) \sum_{i=1}^{n} \ln x_i .$$

Maximum Likelihood Estimators

- The value of $\boldsymbol{\theta}$, say $\hat{\theta}_{MLE}$, that maximizes $L(\boldsymbol{\theta})$ is called the *maximum likelihood estimator*
- Maximum likelihood estimators are values of the parameters $\boldsymbol{\theta}$ that are "most likely" to have been produced by the data
- Because $\ln(\cdot)$ is a one-to-one function, we can also determine $\hat{\theta}_{MLE}$ by maximizing the log-likelihood function, $l(\boldsymbol{\theta})$

Example. Course C/Exam 4. May 2000, 21. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with cumulative distribution function:

$$\mathrm{F}(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

Calculate the maximum likelihood estimate of the parameter α

Likelihood Ratio Test

One important type of inference is to select one of two candidate models, where one model (reduced model) is a special case of the other model (full model)

In a **Likelihood Ratio Test**, we conduct the following hypothesis test:

- H_0: The reduced model is correct
- H_1: The full model is correct

To conduct the Likelihood Ratio Test:

- Determine the maximum likelihood estimator for the full model, $\hat{\theta}_{MLE}$
- Now assume that p restrictions are placed on the parameters of the full model to create the reduced model; determine the maximum likelihood estimator for the reduced model, $\hat{\theta}_{Reduced}$
- The statistic, $LRT = 2\left(l(\hat{\theta}_{MLE}) - l(\hat{\theta}_{Reduced})\right)$, is called the likelihood ratio (a difference of the logs is the log of the ratio. Hence, the term "ratio.")
- The critical value for the likelihood ratio test is a percentile from a chi-square distribution with degrees of freedom equal to p
- This allows us to judge which of the two models is correct. If the statistic LRT is large relative to the critical value, then we reject the reduced model in favor of the full model

Information Criteria

- The following statistics can be used when comparing several candidate models that are not necessarily nested (as in the Likelihood Ratio Test) One picks the model that maximizes the criterion
- *Akaike's Information Criterion*

$$AIC = l(\hat{\theta}_{MLE}) - (number\ of\ parameters)$$

  - The additional term $(number\ of\ parameters)$ is a penalty for the complexity of the model
  - Other things equal, a more complex model means more parameters, resulting in a smaller value of the criterion
- *Bayesian Information Criterion*

$$BIC = l(\hat{\theta}_{MLE}) - (0.5)(number\ of\ parameters)\ln(number\ of\ observations)$$

  - This measure gives greater weight to the number of parameters, resulting in a larger penalty
  - Other things being equal, BIC will suggest a more parsimonious model than AIC