Likelihood functions

Susan Thomas
IGIDR, Bombay

August 12, 2008

Recap

- Likelihood functions are based on the probability of observing the data.
- The first step is fixing a probability distribution, $f(\theta)$ where θ is the parameter defining the probability distribution.
- For a given dataset, $(Y_1, Y_2, \ldots, Y_N)$, the probability of observing the dataset, given θ is:

  $$f_\theta(Y_1, Y_2, \ldots, Y_N)$$

  This is a statement in outcome-space.
- The likelihood function turns this around:
  $L_{Y_1, Y_2, \ldots, Y_N}(\theta) = f_\theta(Y_1, Y_2, \ldots, Y_N)$
  L is a statement in parameter-space.

Example: likelihood for a Bernoulli distribution

- For eg., for a Bernoulli distribution, the probability distribution is:

$$f_\theta(y) = \theta^y(1-\theta)^{(1-y)}$$

- Given a sample of N observations, the joint distribution of $(Y_1, Y_2, \dots, Y_N)$ is:

$$\begin{aligned} f_\theta(\vec{Y}) &= \Pi_{i=1}^{i=N} f(Y_i = y_i) \\ &= \Pi_{i=1}^{i=N} \theta^{y_i}(1-\theta)^{(1-y_i)} \end{aligned}$$

- Example, suppose $(\vec{Y}) = (0, 0, 0, 1, 0, 0, 1, 1)$. What is $f_\theta(\vec{Y})$?

$$\begin{aligned} f_\theta(\vec{Y}) &= \Pi_{i=1}^{i=N} f(Y_i = y_i) \\ &= (1-\theta)^5\theta^3 \\ &= L_{(\vec{Y})}(\theta) \end{aligned}$$

Estimation using the likelihood approach

- The likelihood approach asks: *What value of θ makes the dataset ($\vec{Y}$) most probable?*

$$\hat{\theta} = \arg \max L_{(\vec{Y})}(\theta) = f(\theta; \vec{Y})$$

- Likelihood estimation: What value of θ makes ($\vec{Y}$) most probable?
- Usual approach: maximise the function wrt θ, set it to zero.
- This is the Maximum Likelihood Estimation of parameter θ. Usually referred to as **MLE**.

Example of MLE for a Bernoulli distribution

- Example, Bernoulli distribution. Dataset: $(\vec{Y}) = (0, 0, 0, 1, 0, 0, 1, 1)$.
- Here, the likelihood function, $L_{(\vec{Y})}(\theta) = (1-\theta)^5\theta^3$.
- Log is a monotonic transformation that makes it simpler to work with.
- Log likelihood function is $\log(L)_{(\vec{Y})}(\theta) = 5 * log(1-\theta) + 3 * log(\theta)$
- Maximise the function for θ – differentiating it wrt θ:

$$
\begin{aligned}
\log(L)_{(\vec{Y})}(\theta) &= 5 * log(1-\theta) + 3 * log(\theta) \\
\delta \log(L)/\delta\theta = 0 &= -5/(1-\hat{\theta}) + 3/\hat{\theta} \\
0 &= -5\hat{\theta} + 3(1-\hat{\theta}) \\
\hat{\theta} &= 3/8 = 0.375
\end{aligned}
$$

Generalising the MLE for the Bernoulli distribution

- Given a generic data set, $\vec{Y} = (Y_1 \leq y_1, \ldots, Y_N \leq y_N)$, given they are distributed bernoulli with parameter θ:

$$\begin{aligned}
L_{(Y_1 \leq y_1, \ldots, Y_N \leq y_N)}(\theta) &= \Pi_{i=1}^{i=N} P(Y_i \leq y_i) \\
&= \Pi_{i=1}^{i=N} \theta^{y_i} (1-\theta)^{(1-y_i)} \\
&= \theta^{\sum_1^N y_i} (1-\theta)^{\sum_1^N (1-y_i)} \\
\text{But } \bar{y} &= \frac{1}{N} \sum_1^N y_i \\
L_{(Y_1 \leq y_1, \ldots, Y_N \leq y_N)}(\theta) &= \theta^{n\bar{y}} (1-\theta)^{n(1-\bar{y})} \\
\text{Transforming into log space} & \\
\log L_{(Y_1 \leq y_1, \ldots, Y_N \leq y_N)}(\theta) &= n\bar{y} \log \theta + (1-\bar{y}) \log (1-\theta)
\end{aligned}$$

The MLE of the generic bernoulli distribution

- Maximising $\log L$ wrt θ gives:

$$\begin{aligned}
\delta \log L/\delta\theta &= n\left(\frac{\bar{y}}{\theta} - \frac{1-\bar{y}}{1-\theta}\right) \\
n\left(\frac{\bar{y}}{\hat{\theta}} - \frac{1-\bar{y}}{1-\hat{\theta}}\right) &= 0 \\
\hat{\theta} &= \bar{y}
\end{aligned}$$

(Cross-check that it is the maximum? Calculate the second derivative of $\log L(\theta)$ wrt θ and check that it is negative at $\hat{\theta}$.)

- $\hat{\theta}$ is the value of the distribution parameter that maximises the value of the likelihood function.

$$\hat{\theta}_{\mathrm{mle}} = \bar{y}$$

- This expression for θ is called the **estimator**.
- The specfic value of $\hat{\theta}$ for the given sample is called the **estimate**.

Point 1 to remember about the likelihood function

- The MLE does not give the "most probable" value of θ. It gives the under which the sample is the most likely. Ie, the likelihood is maximised.
- MLE is not magic: all the problems of inference from sample remain with us.
- For example: I tossed a coin 10 times and got 9 heads. Using this data, the MLE gives $\hat{p} = 0.9$.
  MLE does not eliminate sampling noise, or give us the truth. It's just a decent estimator.

Point 2 to remember about the likelihood function

- Since $f()$ is a joint probability, we will always have $\log L(\theta : X_i) > 0$.
  But we **can** have $\log L(\theta : X_i) > 1$.
- Remember that $f(x)$ is a pdf, but $g(\theta)$ is not! Specifically, integrating over parameter space,

$$\int_{-\infty}^{\infty} L(\theta) d\theta \neq 1!$$

Back to the econometrician's checklist of tasks

- In defining and applying the likelihood approach, we have executed Step 1: ie, *estimated* the economic model.
- Step 2 is validating the hypothesis: ie, *inference*.
- For example, in the economic problem using the number of girls vs. boys among newly borns, the hypothesis was that the probability of a girl being born is 50%. Ie, $\theta = 0.5$.
- In our dataset, $\bar{y} = 48.74\%$
- Inference asks the question: is the sample esimate statistically different from the hypothesis?

Approach to implementing statistical inference

- Consider a "restricted" model estimation: we set

$$\theta = 0.5$$

  Under $\theta = 0.5$ we can calculate the joint probability of observing $\vec{Y}$. This becomes the likelihood value of the "restricted" model.
- We have already calculated the likelihood of the "unrestricted" model – which is

$$\hat{\theta} = \bar{y}$$

- Statistically test whether the value of "unrestricted" model likelihood is significantly different from the "restricted" model likelihood.

Approach to implementing statistical inference

- A popularly used test is called the "log-likelihood" ratio test, or the **LR** test statistic:

$$\mathrm{LR} = -2 \log \left(L_{\mathrm{restricted}} / L_{\mathrm{unrestricted}} \right)$$

- We can calculate the value for both.
- Question: what do we expect it to be?
- In our dataset of fraction of girl vs. boy newborns, the likelihood values are:

$$\begin{aligned} \log L_{\mathrm{R}} &= -496290.6 \\ \log L_{\mathrm{U}} &= -496033.8 \\ \mathrm{LR} &= 513.6 \end{aligned}$$

- Questions: is this a large difference?
- The answer comes from theorems on what distributions we can expect for likelihood statistics.

Distributions of sample estimates

Population parameters and sample estimates

- Given a population distribution, $f(x)$ and a sample from that population, we know:
  - $f(x)$ is a deterministic function of the PD/PDF.
    But $\hat{f}(x)$ is a random variable, which is the function of the sample!
  - $f(x)$ is always the same for a given x.
    $\hat{f}(x)$ varies depending upon the sample.
- This is also true for moments of the population and the sample.
- For instance, the first moment of a distribution is $E(x)$.
  $E(x) = \mu$ is a deterministic function of the PD/PDF.
  $E(\vec{x}) = \hat{\mu}$ varies in value from sample to sample.

Population parameters and sample estimates

- Therefore, a sample moment is an **estimate**, which is a random variable.
- Like all rv, every estimate has to have a *expected value* and a *variation* around the expected value.
- This is unlike the case of the population distribution, which has a well-defined *expected value*, and therefore, *no variance*.

Population parameters and sample estimates

- Example of the fraction of girl vs. boy births,
  - $E(y) = \hat{\mu}_y = \sum_{i=1}^{N} y_i = 0.4876$
  - Variance of y = $E(y - \hat{\mu}_y)^2 = E(y)^2 - E(\hat{\mu}_y)^2$
    This works out to be $0.4876 - 0.4876^2 = 0.25$
- This is interpreted as:
  Across different samples of size N, we expect that the mean $E(y)$ will be 0.4876.
  But since $E(y)$ will be different for different samples, there will be a range of values of $E(y)$ around 0.4876, which is determined by $\sigma = 0.5$
- This implies that the expected fraction of girl to boy births in the population distribution could be different from the estimate from any one sample.
- However, there **is** a link between population moments and sample moments, despite sampling uncertainty.
  This link is derived using *asymptotic theory* or the theory of large-samples.