Lecture 2 – Introduction to maximum likelihood

Is the distribution of the sexes among random newborns even? To answer this question, we need a way to model the sex of a random newborn from some population. This means that we need to be able to assign probabilities to possible outcomes of sex for a random newborn. Define Y_i to be the sex of the ith newborn, where $i = 1, \ldots, n$.

Questions: Suppose you have not gathered data yet. What are the possible values for Y_i? How is this variable different from other variables you may have encountered before?

Let us make the following assumptions:

1. Independence: $Y_1, \ldots, Y_n$ are mutually independent.

2. Identical distribution: All newborns are drawn from the same population.

3. Known distribution: $Y_i \sim \text{Bernoulli}(\theta)$

4. Parameter space: $\theta \in (0, 1)$ (Note: We exclude boundary points.)

The objective is then to figure out θ, which is unknown, given a sample. What we are going to do is very different from the case where we calculate probabilities of observing outcomes given that we know θ.

Questions: Suppose $X \sim \text{Bernoulli}(\theta)$, where θ is known to be equal to 0.3. What is the probability of observing the event that $X = 1$? Try graphing this. Now, suppose $X_1, X_2 \sim \text{Bernoulli}(\theta)$, where θ is once again known to be equal to 0.3. Assume independence. What is the probability of observing the event that $X_1 = 1$ and $X_2 = 1$? Try graphing the joint distribution of (X_1, X_2).

The basic machinery of maximum likelihood The maximum likelihood (ML) principle is based on the idea that we want to maximize the probability of observing the data. Here, what we observe are outcomes and we are recovering the parameters using these outcomes.

Question: Now, write down the probability of observing $(Y_1, \ldots, Y_n)$ under the assumptions made earlier. If you treat the result as a function of θ, you have what we call a likelihood function.

To apply ML, suppose we gathered data and got a sample that looks like $(Y_1 = y_1, \ldots, Y_n = y_n)$. This is a bit abstract but imagine numbers for these y_i's. It is a bit difficult to maximize this likelihood function directly. So, we take logarithms, giving us the log-likelihood function.

Question: Why would taking logarithms not alter the maximizer of the likelihood function? So think about what a maximum means and what happens when you take the logarithm of a function. After convincing yourself of this, find the maximum likelihood estimator (MLE).

At this point, you already have a basic idea of doing ML estimation. Now, we explore some of its properties.

Questions: What if instead of $Y_i \sim \text{Bernoulli}(\theta)$, we have $Z_i \sim \text{Bernoulli}(\eta)$. In this case, $Z_i = 1 - Y_i$. What is happening here? How is the MLE affected?

It would help at this point to know the following results from statistics:

1. (Strong) law of large numbers (SLLN): Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $\mathbb{E}(X_1) < \infty$. Then, as $n \to \infty$,

$$\overline{X} = \sum_{i=1}^{n} X_i \stackrel{a.s.}{\to} \mathbb{E}(X_1).$$

2. Central limit theorem (CLT): Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with mean μ and finite positive variance σ^2. Let $\overline{X} = \sum_{i=1}^{n} X_i$. Then, as $n \to \infty$,

$$\Pr\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z),$$

   for all $x \in \mathbb{R}$. Here $\Phi(z)$ is the standard normal cumulative distribution function (cdf) evaluated at z.

Questions: What is the distribution of the MLE? There is a small-sample distribution and a large-sample distribution. Is the MLE consistent for θ? Is the large-sample distribution normal?

In general, we have the following result for the MLE in the case of a one-dimensional parameter.

Theorem 1 Suppose $X_1, \ldots, X_n$ are iid with probability distribution governed by parameter θ. Let θ_0 be the true value of θ. Under regularity conditions, the MLE for θ is asymptotically normal. The asymptotic mean of the MLE is θ_0. The asymptotic variance can be computed in three ways: (i) I_θ^{-1}/n, (ii) $I_{\hat{\theta}}^{-1}/n$, (iii) the inverse of the negative of the second derivative of log-likelihood function at the MLE.

Fisher information The Fisher information I_θ is the negative of the expected value of the second derivative of the log-likelihood function, for a sample size of 1.

Exercises that you should try at this point (when you get home) Check out Exercise Set A in Chapter 7 of Freedman's book. Try #1, #2a, #3, #4, #9, #10, and #12. The others are a bit difficult. We will come back to #8 on May 13.

Returning to our very first question Is the distribution of sexes among random newborns even? How do we answer this question, now that we have the machinery? For the moment, data from the UK during the year 2004 indicates that there were 367586 male and 348410 female newborns.

Questions: What is the MLE given our earlier discussion?

How do we use the MLE to figure out whether the distribution of sexes among random newborns even or uneven? This is a classic test of hypothesis. The likelihood framework gives us three ways of testing the said hypothesis:

1. Look at the MLE itself and compare with the hypothesized value. This leads to the Wald (W) statistic.

2. Look at the log-likelihood values at the MLE and at the hypothesized value. This leads to the likelihood ratio (LR) statistic.

3. Look at the gradient or score of the log-likelihood at the hypothesized value. This leads to the score (LM) statistic.

Questions: Why is it a good idea to look at these statistics? Do they lead to the same conclusions? What are the distributions of these statistics at the null? Perhaps this is a good time to actually think through what you are doing when you do hypothesis testing.

To give you a taste of some theory in econometrics, we will try to derive the distribution of the LR statistic from scratch, at least for this simple model. These are the steps:

1. Show that
$$LR = -2n\left[\overline{Y}\log\frac{\theta}{\overline{Y}} + \left(1-\overline{Y}\right)\log\frac{1-\theta}{1-\overline{Y}}\right].$$

2. Show that
$$LR = -2n\left[\overline{Y}\log(1+x) + \left(1-\overline{Y}\right)\log(1-y)\right],$$
where $x = \left(\theta - \overline{Y}\right)/\overline{Y}$ and $y = \left(\theta - \overline{Y}\right)/\left(1-\overline{Y}\right)$.

3. Here we take note of a Taylor series expansion of the logarithmic function, i.e.,
$$\log(1+x) = x - \frac{x^2}{2} + \frac{(x^*)^3}{3},$$
for all $|x| \leq 1$ and for some $|x^*| \leq |x|$. Show that
$$LR = \frac{n\left(\theta - \overline{Y}\right)^2}{\overline{Y}\left(1-\overline{Y}\right)} + Z^*,$$
where Z^* is some remainder term.

4. Now, using the LLN and CLT, show that $LR \xrightarrow{d} \chi^2_1$. Here χ^2_1 is the chi-squared distribution with 1 degree of freedom (df).

Extensions It is possible to extend the ideas here to the case where we have a k-dimensional parameter θ. It is also possible to drop the assumption of identical distribution from the iid assumption. These two extensions will suffice for the cross-sectional cases we are considering.