Maximum Likelihood Estimation (MLE)

Week 4
POLS 8830: Advanced Quantitative Methods

Ryan Carlin
Georgia State University
rcarlin@gsu.edu

Maximum Likelihood Estimation (MLE)

- Take the classic linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

- Under all the assumptions of the CLRM, taking the partial derivative of equation [1] with respect to $\mathbf{x_k}$ yields:

$$\frac{\partial E(\mathbf{y}|\mathbf{X})}{\partial \mathbf{x_k}} = \frac{\partial \mathbf{X}\beta}{\partial \mathbf{x_k}} = \beta_k \tag{2}$$

Maximum Likelihood Estimation (MLE)

- In the CLRM, the partial derivative helps calculate the slope coefficient for *each* independent variable, holding everything else constant.
- Two important differences between LRM and non-linear models (such as MLE):
  - First, the partial derivative in equation [2] only depends on the value of β_k and nothing else
  - In non-linear models (such as MLE) $\frac{\partial E(\mathbf{y}|\mathbf{X})}{\partial \mathbf{x_k}}$ is influenced by the value of $\mathbf{x_k}$ and also the values of all the other independent variables in the model.

Maximum Likelihood Estimation (MLE)

- Second, in the CLRM, taking the partial derivative boils down to measuring the discrete change in $\mathbf{x_k}$ and the corresponding change in $\mathbf{y}$.
  - In non-linear models, $\frac{\partial E(\mathbf{y}|\mathbf{X})}{\partial \mathbf{x_k}}$ is not simply measuring the discrete change in $\mathbf{x_k}$ and the corresponding change in $\mathbf{y}$.
- Therefore, the major differences between OLS and MLE:
  - ML estimates do **NOT** reflect a deterministic behavior with an attached error term
  - Rather, ML estimates follow a distribution of possible behaviors
  - Determining the appropriate distribution for $\mathbf{y}$ (and by extension for ϵ) is critical to MLE, and is often highly subjective.
  - In other words, *it is a critical — and often unstated — assumption.*

Some Notes on Distributions

- Given the importance of selecting the appropriate distribution, the question becomes how to select from among the dozens of known statistical distributions
- Information about our dependent variable helps us narrow down our choices to a given family of distributions:
  - Is the dependent variable continuous or discrete?
  - Is the depend value truncated a a given value (e.g. 0)
- Our choice of distribution reflects (in part) our level of uncertainty about the functional form of the relationship between $\mathbf{X}$ and the $\mathbf{y}$.
- This is an important decision that requires careful thought, examination of various plots and other preliminary data analysis techniques, and knowledge of the nature of the dependent variable.

Bernoulli Distribution

- This is the simplest statistical distribution
- Represents the situation where a random variable (**y**) has only two possible event outcomes, each with a non-zero probability of occurrence
- Example: flipping a coin
- $\Pr(y_i = 1) = \pi$ and $\Pr(y_i = 0) = 1 - \pi$.
- Formally, we represent the distribution:

$$y_i \sim f_{Bern}(y_i|\pi) = \pi^{y(1-\pi)(1-y)}$$

Binomial Distribution

- This is a series of N Bernoulli random variables, where we only observe the sum of the observations
- The distribution is nonnegative and discrete (no fractions), with an upper bound of n
- Examples: the number of bills in a legislature, number of cases on a court's docket
- Mathematical specification:

$$f_{k,n,p} = \binom{n}{k} p^k (1-p)^{(n-k)}$$

- where:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Normal Distribution

- Most intuitively familiar distribution (pdf froms the familiar “bell-shaped” curve)
- Used in OLS regression models
- Somewhat difficult to employ in MLE, because it does not possess an analytic solution
  - Analytic solution requires computing integrals
  - Computationally, the mathematics underlying this distribution were too complex for early computers
- Mathematical specification:

$$y_i \sim \mathcal{N}(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{(y-\mu)^2}{\sigma^2}\right]}$$

Logistic Distribution

- Better adept at modeling probabilities for dichotomous outcomes than the Normal distribution
- Contains an analytic solution (e.g. is mathematical tractable)
- Low computational costs (can even be done by hand)
- Mathematical specification:

$$y_i \sim f_{Logistic}(y_i|\mathbf{X}\beta) = \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}}$$

Poisson Distribution

- Used when dependent variable is a count with no upper bound
- Key assumption: Occurrence of one event has no influence on the expected number of subsequent events (λ)
- Mathematical specification:

$$y_i \sim f_{Poisson}(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

- where $\lambda > 0$ and $y_i = 0, 1, 2, \ldots$

Negative Binomial Distribution

- Two key assumptions about the Poisson distribution are often problematic:
  - That events accumulating during observation period i are independent
  - Events have a constant rate of occurrence
- If either assumption is violated, then a new distribution is required because λ is no longer constant for all observations
  - Instead, must assume that λ itself varies across observations according to a particular probability distribution
  - The most popular distribution for λ is the gamma distribution
  - This involves calculating another parameter in the equation — the variance of the distribution

Negative Binomial Distribution

- Mathematical specification:

$$y_i \sim f_{nb}(y_i|\lambda, \sigma^2) = \frac{\Gamma\left(\frac{\lambda}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\lambda}{\sigma^2-1}\right)} \left(\frac{\sigma^2 - 1}{\sigma^2}\right) y_i (\sigma^2)^{\frac{-\lambda}{\sigma^2-1}}$$

- where $\lambda > 0$ and $\sigma^2 > 0$.
- Note: the more events within observation i that are positively related, the larger σ^2 becomes. Also, as σ^2 approaches 0, the negative binomial distribution collapses into the Poisson distribution

Calculating a Maximum Likelihood

- Calculating a maximum likelihood refers to the joint probability that the observations included in a dataset could have been selected randomly *given the true state of the world*
- Stated another way, the likelihood involves estimating the chance that our dataset would have been selected, as opposed to another dataset with different observations *given the true state of the world*
- Similar to calculating specific probabilities but with increased uncertainty
- Remember, we are estimating the likelihood of the entire dataset, not a single observation
- Assuming that the observations are independently and identically distributed (i.i.d.) then the probability of a joint event is the product of the probability of each single event

An Example

- Assume that we observe one of two possible events: an individual turned out to vote or not
- Voter turnout $1 =$ yes and $0 =$ no
- To calculate the maximum likelihood we must first select the appropriate probability distribution
- In this case the Binomial distribution is appropriate because we have multiple observations (or trials) of a dependent variable with only two outcomes
- We can refer to turnout with the parameter p
- $1 = p$ and $0 = 1 - p$

An Example

- We can calculate the maximum likelihood of observing 3 individuals, 2 of whom turned out to vote
- Observed data is 1,1,0
- One method of calculation is a grid search
- $L \propto \Pr(y|\Theta)$
- where L = likelihood and Θ = parameter of interest
- This is read as "the likelihood of observing our data is proportional to the probability of y given Θ"

An Example

- Calculating a grid search
- Arbitrarily select values for the unknown parameter and calculate the joint probability of observing the data
- Refine calculations until maximum probability is determined

An Example

p	L(1, 1, 0)
0.0	
0.1	
0.2	
0.3	
0.4	
0.5	
0.6	
0.7	
0.8	
0.9	
1.0	

An Example

p	L(1, 1, 0)
0.0	0
0.1	.027
0.2	.096
0.3	.189
0.4	.288
0.5	.375
0.6	.432
0.7	.441
0.8	.384
0.9	.243
1.0	0

An Example

p	L(1, 1, 0)
0.0	0
0.1	.003
0.2	.096
0.3	.189
0.4	.288
0.5	.375
0.6	.432
0.7	.441
0.8	.384
0.9	.243
1.0	0

Finding an Analytical Solution

- The theory of MLE rests on the ability to estimate the probability that a given population (reflected in assumptions regarding the distribution) produced the matrix of observations

$$
\begin{aligned}
L(\Theta|y) &= k(y)\Pr(y|\Theta) \\
&\propto \Pr(y|\Theta)
\end{aligned}
$$

- where k is a constant which translates into the likelihood function measuring relative uncertainty

Finding an Analytical Solution

- Example: calculate the likelihood of observing presidential vetoes of legislation
- Step 1: select the appropriate probability distribution
    - In this case, since we are dealing with count data, the Poisson distribution is appropriate
- The Poisson distribution

$$y_i \sim f_{Poisson}(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

- Remember that the likelihood involves calculating the joint probability
    - Assuming the data are i.i.d. then this involves the product of all individual probabilities

Finding an Analytical Solution

$$L \propto \prod_{i=1}^{N} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \qquad (3)$$

- How do you evaluate the product of observations?
- Calculating the mathematics of a product is extremely complicated
- However, one can take the log of equation [3] and calculate the log-likelihood, which simplifies the mathematics
- **Note: taking the log also means that the β coefficients are calculated from the log-likelihood which is not interpretable**

Finding an Analytical Solution

- Step 1: Distribute:

$$\frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod y_i!} \tag{4}$$

- Step 2: Take the natural log:

$$\ln L = -N\lambda + \sum y_i \ln \lambda - \sum \ln y_i! \tag{5}$$

- Step 3: Take the partial derivative with respect to the single unknown parameter (λ):

$$\frac{\partial \ln L}{\partial \lambda} = -N + \frac{\sum y_i}{\lambda} \tag{6}$$

- Step 4: Set equal to 0 and solve:

$$\lambda = \frac{\sum y_i}{N} \tag{7}$$

Notes on Analytical Solutions

- The first derivative provides information about the slope of a line running tangential to the likelihood curve at its most sensitive location
- The second derivative provides information on how fast the slope of that tangential line is changing along the curve
- The second derivative is known as the Hessian Matrix, the inverse of which is used to calculate standard errors