UNIVERSITY OF MANNHEIM
School of Social Sciences

Advanced Quantitative Methods in Political Science: A first peek at Maximum Likelihood

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova

Week 4 - 9 March 2022

Leftovers

Relaxing the iid assumption

What if iid (independent identically distributed) assumption is unrealistic?

- *Relax identical distribution assumption* ($\pi_i = \pi$) such that π is a random variable rather than being fixed, thus we need to find $P(\pi)$ and π falls in the interval [0,1].
  - Take *Beta distribution*, i.e., $P = B(\rho, \gamma)$, which can be very flexible (unimodal, bimodal, skewed). Also used to model proportions.
- One can show that *relaxing the independence assumption* by letting π vary according to the Beta distribution one gets the *extended Beta-Binomial distribution* P_{ebb}.
  - Combine (aka *compound*) Beta and Binomial distributions to get *extended Beta-Binomial distribution* $P_{ebb}(y_i, \pi|\gamma)$. γ represents the degree to which π varies across the unobserved realizations of the binary random variables. For $\gamma = 0$ one arrives at the binomial distribution again.
- Example: Lauderdale, Benjamin E. (2012). Compound Poisson-Gamma Regression Models for Dollar Outcomes That Are Sometimes Zero. *Political Analysis*, 20(3), 387–399.

Multinomial Distribution

First Principle:

- Characteristics about the DGP that generates $Y = (y_1, \ldots, y_k)' \sim Multinomial(n, \pi_1, \ldots, \pi_k)$:
    - n repeated, independent trials. Each trial has k mutually exclusive and exhaustive outcomes (say $\{1, \ldots, k\}$)
    - Probability that outcome j occurs is $\pi_j \in [0{,}1]$ and $\sum_{j=1}^{k} \pi_j = 1$
    - Let y_j be a random variable counting how often outcome j occurs, thus $\sum_{j=1}^{k} y_j = n$.
- The pmf is:

$$P((y_1, y_2, \ldots, y_k)') = P(y|n, \pi_1, \ldots, \pi_k) = \frac{n!}{y_1! y_2! \ldots, y_k!} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k}$$

- Example? How can it go wrong? What happens for $k = 2$?

- $E(Y_j) = n\pi_j$ and $Var(y_j) = n\pi_j(1 - \pi_j)$

Further Univariate Probability Distributions

There are many, many other distributions (and compounds of them) as you can imagine. Just to name a few ...

- Poisson; Negative binomial for modeling counts - discrete, countably infinite, nonnegative
- Normal - continuous, unimodal, symmetric, unbounded
- Log-Normal; Gamma - continuous, unimodal, skewed, bounded from below by zero
- Truncated-Normal - continuous, unimodal, symmetric, bounded from below or above (or both)
- Multinomial for modeling discrete outcomes - discrete, unordered

Remember: Pick (or construct) a probability distribution to define the stochastic component of your model that best describes the potential values of your outcome variable (i.e., the sample space).

Likelihood as a Model of Inference

The Problem of Inference

Does the number of appointed woman judges reflect descriptive representation?

Second Senate of the Federal Constitutional Court

- How can we answer this question?
- What is the DGP and what is Y?
- Which probability model (stochastic component)?
- Assumption 1: Decisions are made *independent* of every vacant position
- Assumption 2: Each decision has same underlying probability of choosing a women (*identically distributed*)

- The pdf of the Binomial: $P(Y=y|\pi) = \frac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y}$.
- Thus, if $\pi_0 = .5$, then: $P(\text{ No. of women} = 2|\pi_0 = .5) = \frac{8!}{2!6!} \cdot .5^2 \cdot .5^6 \approx .109$
- Is that really what we wanted to know? In fact, we do not know which π generated our data, thus we need to estimate it and see to what degree it is different from $\pi_0 = .5$.

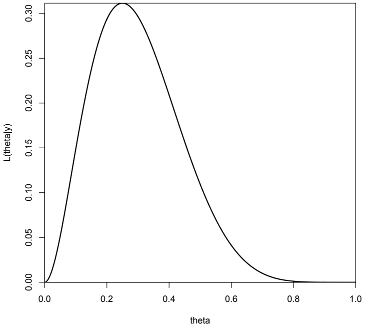
The Likelihood Theory of Inference

- Conditional Probability: $Pr(y|M) = Pr(known|unknown)$
- We actually care about the so-called *inverse probability*: $Pr(M|y) = Pr(unknown|known)$ (and $P(M|y)$ if data is continuous)
- Or at least about: $Pr(\theta|y, M^*) = Pr(\theta|y)$, if $M = \{M^*, \theta\}$ where M^* is assumed and θ to be estimated.
- The solution turns out to be the likelihood, $L(\theta|y)$, defined as values *proportional* to the traditional probability (density) distribution for different values of θ.

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

  - Second line is a more convenient way to express the first line without the constant.
  - $k(y)$ is a unknown function of the data, with θ fixed at its true value. It changes, if y changes.
  - $L(\theta|y)$ is a function. For observed (i.e. fixed) y it returns the *likelihood* of any value θ (that generated the data y assuming $M*$).

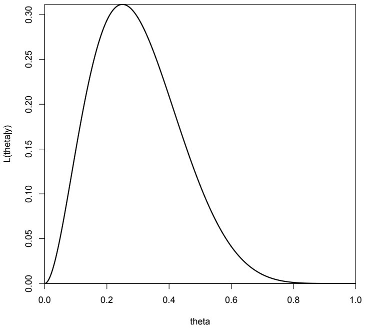
The Likelihood Theory of Inference



- When estimating competing models, the likelihood function gives us information about the *relative* plausibility of various parameter values conditional on the same observed data y
- Comparing the value of $L(\theta|y)$ for different θ's in one data set y makes sense.
- Comparing the value of $L(\theta|y)$ for different θ's across data sets is meaningless (similar to comparing R^2 across OLS regression models with different DVs).

- The likelihood principle: the data only affect inferences through the likelihood function.
- The likelihood function is a summary estimator of θ. Given the likelihood principle this means, that once plotted, we can discard the data (if the model is correct, i.e. inferences are still *model dependent*).

The Likelihood Theory of Inference



- The maximum is a one-point summary of the likelihood function and is called Maximum Likelihood estimate $\hat{\theta}_{ML}$.
- The uncertainty of this point estimate is represented by the curvature at the maximum.
- For analytical tractability or numerical stability the log-likelihood is typically used instead of the likelihood.
- The log-transformation changes the shape of the likelihood, however, the maximum will be the same.

- The value of θ for which the observed data y are most likely (i.e. have highest probability of being observed) is called the *maximum likelihood estimate*.
- In our (univariate) example $\theta = \pi$, thus $L(\theta|y) = L(\pi|y = 2, N = 8)$.

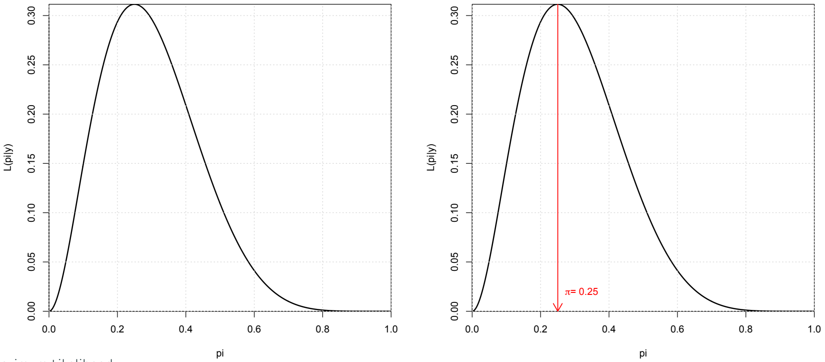
The Likelihood of Our Example

How does the likelihood function $L(\pi|y = 2, N = 8)$ of our example look like?

The Likelihood of Our Example

How does the likelihood function $L(\pi|y = 2, N = 8)$ of our example look like?

$$L(\pi|y = 2, N = 8) = \frac{8!}{2!6!}\pi^2(1 - \pi)^6$$



Maximizing the Likelihood

How to maximize the likelihood function $L(\pi|y = 2, N = 8)$?

Maximizing the Likelihood

How to maximize the likelihood function $L(\pi|y = 2, N = 8)$?

$$
\begin{aligned}
L(\pi|y = 2, N = 8) &= \frac{8!}{2!6!}\pi^2(1-\pi)^6 = 28\pi^2(1-\pi)^6 \\
\frac{\partial L(\pi)}{\partial \pi} &= 56\pi \cdot (1-\pi)^6 - 168\pi^2 \cdot (1-\pi)^5
\end{aligned}
$$

Maximizing the Likelihood

How to maximize the likelihood function $L(\pi|y = 2, N = 8)$?

$$
\begin{aligned}
L(\pi|y = 2, N = 8) &= \frac{8!}{2!6!}\pi^2(1-\pi)^6 = 28\pi^2(1-\pi)^6 \\
\frac{\partial L(\pi)}{\partial \pi} &= 56\pi \cdot (1-\pi)^6 - 168\pi^2 \cdot (1-\pi)^5
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L(\pi)}{\partial \pi} &= 0 \iff \\
56\pi \cdot (1-\pi)^6 - 168\pi^2 \cdot (1-\pi)^5 &= 0
\end{aligned}
$$

After some tedious algebra one obtains $\hat{\pi}_{ML} =$

Maximizing the Likelihood

How to maximize the likelihood function $L(\pi|y=2, N=8)$?

$$\begin{aligned} L(\pi|y=2,N=8) &= \frac{8!}{2!6!}\pi^2(1-\pi)^6 = 28\pi^2(1-\pi)^6 \\ \frac{\partial L(\pi)}{\partial \pi} &= 56\pi \cdot (1-\pi)^6 - 168\pi^2 \cdot (1-\pi)^5 \end{aligned}$$

$$\begin{aligned} \frac{\partial L(\pi)}{\partial \pi} &= 0 \iff \\ 56\pi \cdot (1-\pi)^6 - 168\pi^2 \cdot (1-\pi)^5 &= 0 \end{aligned}$$

After some tedious algebra one obtains $\hat{\pi}_{ML} = .25$ (...tada!).

Easier Way: Maximizing the Log-Likelihood

How to find the maximum of the log-likelihood function $log(L(\pi|y = 2, N = 8))$?

$$
\begin{aligned}
log(L(\pi|y = 2, N = 8)) &= log(28\pi^2(1 - \pi)^6) \\
&= log(28) + 2log(\pi) + 6log(1 - \pi)
\end{aligned}
$$

Easier Way: Maximizing the Log-Likelihood

$\hat{\pi}_{ML}$ fulfills the first-order condition of the log-likelihood

$$
\begin{aligned}
\frac{\partial logL(\pi)}{\partial\pi} = \frac{\partial\left(log(28) + 2log(\pi) + 6log(1-\pi)\right)}{\partial\pi} &= 0 \iff \\
\frac{2}{\pi} - \frac{6}{1-\pi} &= 0 \\
\frac{2}{\pi} &= \frac{6}{1-\pi} \\
2(1-\pi) &= 6\pi \\
2 &= 8\pi \\
1/4 &= \pi
\end{aligned}
$$

Thus, one obtains the same $\hat{\pi}_{ML} = .25$ through maximizing the log-likelihood.

Back to our substantive (univariate) example

Does the number of appointed woman judges reflect descriptive representation?

Second Senate of the Federal Constitutional Court

- Take a look at the *likelihood ratio*, which corresponds to the ratio of the traditional probabilities (Why?)
- Recall:

$$\frac{L(\pi_0 = .5|y = 2, N = 8)}{L(\hat{\pi}_{ML}|y = 2, N = 8)} \approx \frac{.109}{.311} \approx .35$$

- The likelihood for gender reflection $L(\pi_0)$ is 35 percent of the maximum $L(\hat{\pi}_{ML})$.
- Thus, it seems very unlikely that the appointment process is driven exclusively by concerns of descriptive representation.

MLE and the Linear Regression Model

- Suppose we have observed independently the following government approval ratings:

$$Y = \{54, 53, 49, 61, 58, \cdots\}$$

- *First step*: How is the DGP and how is Y distributed? Suppose:

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) && \text{stochastic} \\ \mu_i &= X_i\beta && \text{systematic} \end{aligned}$$

- We have some observations (assuming *iid*) Y and we want to estimate μ_i and σ^2.
- *Second step*: Choose a parametrization of the stuff you would like to estimate. For now we model only μ_i (see above) with covariates. However, we will also (next week!) parameterize the variance to model heteroskedasticity.
- *Third step*: *Maximum Likelihood Estimation* implies that we need to find those parameter values (β, σ^2) of our chosen (assumed) stochastic component that *maximizes* the respective *likelihood function* conditional on the data we have.
- Thus, lets construct the respective *likelihood function*.

How does the Likelihood function look like?

- We assumed that Y_i is distributed normal ($Y_i \sim f_N(y_i|\mu_i, \sigma^2)$), hence for ith observation y_i we get

$$Pr(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right)$$

- Recall that we also assumed Y_i to be iid, thus for instance

$$Pr(Y_1 = 54, Y_2 = 53) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \times exp\left(-\frac{(54 - \mu_1)^2}{2\sigma^2}\right) \times exp\left(-\frac{(53 - \mu_2)^2}{2\sigma^2}\right)$$

- Thus, for N realizations (observations) of iid random variables we get

$$Pr(Y_1, \cdots, Y_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right)$$

How does the Likelihood function look like?

- Applying our parameterization for μ_i the likelihood of the entire sample is

$$L(\beta, \sigma^2 | y, X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right)$$

- Or equivalently in matrix notation

$$L(\beta, \sigma^2 | y, X) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} exp(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta))$$

How does the Log-Likelihood function look like?

- Now taking the logs (rather $ln(\cdot)$) yields

$$
\begin{aligned}
L(\beta, \sigma^2 | y, X) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right) \\
lnL(\beta, \sigma^2 | y, X) &= \sum_{i=1}^{N} ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right)\right] \\
&= -\frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i\beta)^2 \\
&= (\cdot) + (\cdot)\beta - \left(\frac{\sum_{i=1}^{N} x_i^2}{2\sigma^2}\right)\beta^2
\end{aligned}
$$

- Or equivalently in matrix notation

$$
lnL(\beta, \sigma^2 | y, X) = -\frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)
$$

Finding the ML Estimator

- While oftentimes not possible (numerical solutions have to be used instead) in this case we can find a closed form solution ($\hat{\theta}_{ML} = (\hat{\beta}_{ML}, \hat{\sigma}_{ML})'$) of the parameters that most-likely generated the data.
- We start with taking the log-likelihood in matrix notation. By expanding the last term we get

$$lnL(\beta, \sigma^2|y, X) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(y'y - 2y'X\beta + \beta'X'X\beta)$$

- Now we need to take the (partial) derivatives of lnL with respect to β and σ^2 (the entries of the so-called *gradient vector*) and set them equal to zero.

Finding $\hat{\beta}_{ML}$

Taking the derivative of the log-likelihood with respect to β yields

$$\begin{aligned}
\frac{\partial lnL}{\partial \beta} &= -\frac{1}{2\sigma^2}\frac{\partial(y'y - 2y'X\beta + \beta'X'X\beta)}{\partial \beta} \\
&= -\frac{1}{2\sigma^2}(-2X'y + 2X'X\beta) \\
&= \frac{1}{\sigma^2}(X'y - X'X\beta)
\end{aligned}$$

We now set this equal to zero:

$$\begin{aligned}
\frac{1}{\sigma^2}(X'y - X'X\beta) &= 0 \\
X'X\beta &= X'y \\
\hat{\beta}_{ML} &= (X'X)^{-1}X'y
\end{aligned}$$

This is the familiar formula we know from the OLS coefficient vector. Thus, $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$.

Finding $\hat{\sigma}^2_{ML}$

Taking the derivative of the log-likelihood with respect to σ^2 yields

$$\begin{aligned} lnL &= -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \\ \frac{\partial lnL}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{aligned}$$

We now set this equal to zero:

$$\begin{aligned} -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) &= 0 \\ \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) &= \frac{N}{2\sigma^2} \\ \frac{1}{\sigma^2}(y - X\beta)'(y - X\beta) &= N \end{aligned}$$

Since we have already $\hat{\beta}$, we can substitute this in ($\beta = \hat{\beta}$) and solve for σ^2:

$$\begin{aligned} \frac{1}{\sigma^2}(e'e) &= N \\ \hat{\sigma}^2_{ML} &= \frac{e'e}{N} \end{aligned}$$

Comparing $\hat{\sigma}^2_{ML}$ with $\hat{\sigma}^2_{OLS}$

- While $\hat{\sigma}^2_{ML} = \frac{e'e}{N}$, recall that the OLS estimate of the variance, $\hat{\sigma}^2_{OLS} = \frac{e'e}{N-(k+1)}$, is unbiased.
- Thus, $\hat{\sigma}^2_{ML} \neq \hat{\sigma}^2_{OLS}$
- Moreover, $\hat{\sigma}^2_{ML}$ is biased downwards in small samples.
- However, $\hat{\sigma}^2_{ML}$ and $\hat{\sigma}^2_{OLS}$ are asymptotically equivalent, i.e., they converge as N goes to infinity.

MLE and Statistical Inference

Properties of the Maximum (i.e. of $\hat{\theta}_{ML}$)

Small Sample Properties

- Invariance to reparameterization
  - Rather than estimating a parameter $\hat{\theta}_{ML}$, one can first estimate a function $g(\hat{\theta}_{ML})$, which is also a ML estimator.
  - In a second step, recover $\hat{\theta}_{ML}$ from $g(\hat{\theta}_{ML})$.
  - Very useful because $g(\hat{\theta}_{ML})$ might be easier derived, or has an more intuitive interpretation (see e.g., King & Browning's 1987 *APSR*)
  - Allows for transformation of parameters (logit transformation of probabilities; logarithmic transformation of variances; Fisher z-transformation of correlations)
- Invariance to sampling plans
  - Information about how data is collected (e.g., sample size) that does *not* affect the likelihood is irrelevant.
  - OK to look at results while deciding how much (further) data to collect.
  - Allowed to pool data (if independent, just add LL to the existing one!) to get more precise estimates
- Minimum Variance Unbiased Estimator (MVUE)
  - A single unbiased estimator with smallest variance (not necessarily linear!).

Properties of the Maximum (i.e. of $\hat{\theta}_{ML}$)

Asymptotic Properties (think of *repeated sampling*, i.e., let $\{\hat{\theta}_N\}$ be a sequence of estimators calculated in the same way from larger and larger samples of size N. For each sample size, $\hat{\theta}_N$ has a *sampling distribution*)

- Consistency
  - From the *Law of Large Numbers*, as $N \to \infty$, the sampling distribution of $\hat{\theta}_{ML}$ collapses to a spike over the (true) parameter value θ.
- Asymptotic normality
  - From the *Central Limit Theorem*, as $N \to \infty$, the sampling distribution of $\hat{\theta}_{ML}/se(\hat{\theta}_{ML})$ converges to the normal distribution (Mean?, Variance?).
  - No matter what distribution we assumed in the model for θ itself!
  - Allows us to do hypothesis testing and to construct confidence intervals.
- Asymptotic efficiency
  - Among all consistent, asymptotically normal distributed estimators, $\hat{\theta}_{ML}$ has the smallest variance.
  - $\hat{\theta}_{ML}$ contains as much information as can be packed into a point estimator.