PoliSci 4782 Political Analysis II

Theories of Inference and Maximum Likelihood Estimation

Seamus Wagner

The Ohio State University

Inferences

- Descriptive inference
  - use samples to study population (e.g. census, public opinion polls)
- **Predictive inference**
  - use models/simulations to make forecasts
- **Causal inference**
  - use models/research designs to identify causal relationships

Model Estimation

- Model estimation is also a process of inference ($\boldsymbol{X}, \boldsymbol{Y} \rightsquigarrow \hat{\beta}, \hat{\eta}$).
- For linear models, ordinary least squares (OLS) is the best linear unbiased estimator (BLUE).
- For generalized linear models, OLS may not work.
  - The systematic component may not be linear, so we cannot fit a straight line to our data.
  - The probability density of the outcome variable may not be the normal, so the stochastic component may follow a different distribution.

Two Theories of Inference: Likelihood vs. Bayesian

Problems of Inference

- Here is probability by definition:

$$Pr(y|M) \equiv Pr(Data|Assumption) \equiv Pr(Known|Unknown)$$

- This is the goal of inference as inverse probability:

$$Pr(\theta|y, M^*) \equiv Pr(Unknown|Known)$$

where M^* is the assumed model M with unknown parameter(s) θ and y is data

- To be more succinct, inference is $Pr(\theta|y)$

Problem of Inference

According to the Bayes Theorem:

$$Pr(\theta|y) = \frac{Pr(\theta, y)}{Pr(y)}$$
$$= \frac{Pr(\theta)Pr(y|\theta)}{Pr(y)}$$

- $Pr(y|\theta)$ is probability density function, but what is the rest on the right side?
- Two groups of theorists, likelihoodists and Bayesians, provide different interpretations, which leads to different estimation approaches.

Interpretation 1: Likelihood Theory

- According to likelihood theorists (as different from Bayesian theorists), θ is fixed while y is random (*the laws have been written, but probability also plays a part*).
- As θ is fixed, $\frac{Pr(\theta)}{Pr(y)}$ is only a function of y, which we can rewrite as $k(y)$:

$$\frac{Pr(\theta)}{Pr(y)} \equiv k(y)$$

- Thus,

$$Pr(\theta|y) = k(y)Pr(y|\theta)$$

Likelihood Estimation: Genesis

- Given
$$Pr(\theta|y) = k(y)Pr(y|\theta)$$
- $k(y)$ is unknown but still a function of y
- So the targeted probability is proportional to $Pr(y|\theta)$:
$$Pr(\theta|y) = k(y)Pr(y|\theta) \propto Pr(y|\theta)$$
- We define the likelihood function L that gives the probability of any value of θ given y:
$$L(\theta|y) \propto Pr(y|\theta)$$

Likelihood

$L(\theta|y) \propto Pr(y|\theta)$:

- Likelihood is a relative measure of uncertainty. It changes with the data set y.
- Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.
- Comparing values of $L(\theta|y)$ across data sets is meaningless (just as you can't compare R^2 values across equations with different dependent variables).
- The likelihood principle: the data only affect inferences through the likelihood function.

(Log-)likelihood Estimation

- For algebraic simplicity and numerical stability, we use a natural log likelihood function to estimate θ.
    - $ln(A \times B) = ln(A) + ln(B)$
- The logarithmic transformation simplifies the shape of the function without changing the position of the maximum point.
- The estimation strategy is to find the maximum point (the most likely θ given our data).
- We will detail this method later.

Interpretation 2: Bayesian Theory

Recall:

$$
\begin{aligned}
Pr(\theta|y) &= \frac{Pr(\theta, y)}{Pr(y)} \\
&= \frac{Pr(\theta)Pr(y|\theta)}{Pr(y)}
\end{aligned}
$$

- According to Bayesian theorists, however, y is fixed while θ is random (*only what you see is certain*).
- Because y is fixed,

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

Bayesian Inference

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

- $Pr(\theta)$ is called "the prior (probability)," which distinguishes Bayesian inference from likelihood estimation
- $Pr(\theta|y)$ is called "the posterior (probability)," a probability density fucntion that takes both the prior and the probability density function of y into account.

The Prior $Pr(\theta)$

- It is a probability density that represents all prior evidence about θ.
- It provides an opportunity/requirement of getting other (theoretical/qualitative) information outside the data set into the inference.
- The philosophical assumption in behind is that nonsample information should matter (as it always does) and be formalized and included in all inferences.

The Posterior $Pr(\theta|y)$

- Like L, it is a summary estimator for all possible values of θ.
- It also obeys the principle that the data set only affects inferences through likelihood function.
- If $Pr(\theta) = 1$ (i.e., a uniform distribution in the relevant region), there is no difference between likelihood function and the Bayesian posterior probability function ($L(\theta|y) = Pr(\theta|y)$).
- "Likelihoodists are Bayesians who do not know their priors."

Comparison between Likelihood and Bayesian

- Likelihood is more mainstream and mathematically easier to comprehend (thus becomes our focus).
- Because of technological development (e.g. better computational capacity in PC and MCMC algorithms), Bayesian has growing attractions as it includes more information (the prior).
- Huge philosophical differences yet minor practical differences.

Maximum Likelihood Estimation

Likelihood Function

- Recall

$$L(\theta|y) \propto Pr(y|\theta)$$

- Now $f(y|\theta)$ is the probability density function of y given parameters θ.
- If our data $y_i \in (y_1, y_2, ...y_n)$ are *independently and identically* distributed—in order words, follow the same probability distribution and are mutually exclusive,

$$f(y_1, y_2, ...y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \mathcal{L}(\theta|y)$$

Log Likelihood Function

- $\prod_{i=1}^{n} f(y_i|\theta)$ is algebraically difficult.
- Given that $\ln(xy) = \ln(x) + \ln(y)$, we switch to log likelihood for computational simplicity:

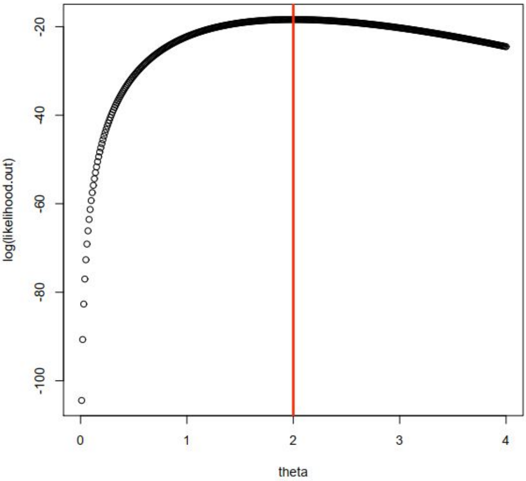
$$\ln \mathcal{L}(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

More precisely,

$$\ln \mathcal{L}(\theta|\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}, \boldsymbol{x_i})$$

Solving Log Likelihood Function

The goal is to find the $\hat{\theta}$ that maximizes the likelihood score—this is why this method is called maximum likelihood estimation (MLE):



Maximum and Curvature of the Likelihood

- If the log-likelihood is well approximated by a quadratic function, we need at least two quantities to represent it:
  - The location of the maximum (which indicates estimated value)
  - The curvature at the maximum (which indicates estimation uncertainty)
- Define the **score function** $S(\theta)$ as the first derivative of the log-likelihood: $S(\theta) \equiv \frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta)$.
- At the maximum, the score function equals to 0 and its shape curves downward so the second derivative will be negative.
- Define the curvature at $\hat{\theta}$ as $I(\hat{\theta})$, where $I(\hat{\theta}) \equiv -\frac{\partial^2}{\partial\theta^2} \ln L(\hat{\theta})$ (observed Fisher information)
  - A large curvature is associated with a tight peak (less uncertainty about θ)

Standard Errors in MLE

- Given the estimated $\hat{\theta}$ and the observed Fisher information $I(\hat{\theta})$ as the curvature of the score function, we can compute standard error by

$$\mathrm{se}(\hat{\theta}) = I^{-1/2}(\hat{\theta})$$

- We report $\hat{\theta}\,[\mathrm{se}(\hat{\theta})]$ as estimation results

Steps of MLE

- Write down log likelihood function.
- Take the first derivative (score function).
- Set the score function equal to zero (in order to find the maximum or minimum).
- Solve for θ and label it $\hat{\theta}$.
- Make sure that it is the maximum, not the minimum (by checking sign of the second derivative is negative)
- Compute standard error with observed Fish Information

Generalizing to Multiple Parameters

When we have multiple parameters θ (e.g. effect parameters) to estimate:

- The score function is the same (first derivative w.r.t. θ), but now we have a vector of first derivatives
- The **score vector** (sometimes called the **gradient vector**) is a vector of length k

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln L(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_k} \ln L(\boldsymbol{\theta}) \end{bmatrix} = 0$$

Generalizing to Multiple Parameters

- For the second derivatives, we now end up with an **information matrix** (**Hessian matrix**)
- A $k \times k$ matrix of second partial derivatives of the log-likelihood w.r.t. the parameters:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1 \theta_k} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2 \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k^k} \end{bmatrix}$$

- Standard errors for $\hat{\boldsymbol{\theta}}$ are given by the square roots of the diagonal elements of $\boldsymbol{I}^{-1}(\hat{\theta})$

Properties of MLE

Under a few mild regularity conditions:

- Consistency: as $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value.
- Asymptotic normality: as $n \to \infty$, the distribution of MLE/se(MLE) converges to the normal distribution.
- Asymptotic efficiency: as $n \to \infty$, the MLE contains as much information as can be packed into a point estimator.

In short, **the larger dataset, the better MLE performs**.

OLS vs. MLE

- For outcomes variables following other probability distributions than the normal, MLE works whereas OLS does not.
- When outcome variables follow the normal, MLE is equivalent to OLS if sample size is sufficiently large.
- Under a set of conditions specified by the Gauss-Markov theorem, OLS is the best.

OLS vs. MLE: Regression Results

	OLS	MLE 1	MLE 2
(Intercept)	13.03	13.03	4.04
	(15.87)	(15.87)	(6.39)
sex	-24.34^{**}	-24.34^{**}	-21.63^{**}
	(8.13)	(8.13)	(6.81)
status	-0.15	-0.15	
	(0.24)	(0.24)	
income	4.93^{***}	4.93^{***}	5.17^{***}
	(1.04)	(1.04)	(0.95)
R^2	0.51		
Num. obs.	47	47	47
RMSE	22.92		
AIC		433.59	432.01
BIC		442.84	439.41
Log Likelihood		-211.79	-212.00
Deviance		22579.50	22781.32

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

OLS vs. MLE

- OLS and MLE have identical estimates and standard errors, when the same model specification is run.
- R^2 and $RMSE$ are unavailable in MLE.
- Log likelihood, deviance, AIC, BIC are added.

Log Likelihood Score

- It is the result of your log likelihood function with estimated parameters.
- Its value can be negative, because a logarithmic function can generate negative values.
- If two models are estimated by MLE with the same set of data, you can evaluate the performance of the two by comparing their log likelihood scores (the larger, the better).

Nested Models and Likelihood Ratio Test

- For nested models, we can further do a likelihood ratio test to decide if their difference in the likelihood score is significant or not.
- Two models are considered nested, if the "longer" model contains all the explanatory variables of the "shorter" one.
- The shorter model with fewer variables is referred to as the "*restricted*" model, while the longer one as the "*unrestricted*."
- If the test concludes with statistical significance, we say that the unrestricted model is significantly better.

Likelihood Ratio Test

- Let $\hat{L}_U$ and $\hat{L}_R$ be likelihoods for the unrestricted and restricted model respectively.
- Their difference in the number of parameters is k.
- We can compute the log likelihood ratio by

$$LR = -2\ln(\frac{\hat{L}_R}{\hat{L}_U}) = -2[\ln(\hat{L}_R) - \ln(\hat{L}_U)]$$

  which follows χ^2 distribution with k as the degree of freedom.
- We then conduct a significance test on likelihood ratio to decide whether the difference between the two is significant.

Likelihood Ratio Test in R

```r
m1 <- lm(gamble ~ sex + status + income, data = teengamb)

m2 <- glm(gamble ~ sex + status + income, data = teengamb, family = gaussian)

m3 <- glm(gamble ~ sex + income, data = teengamb, family = gaussian)

texreg(list(m1, m2, m3), no.margin = T)


##################################################################
### likelihood ratio test

# log likelihood reported by regression tables

l.r <- -212.00   # the restricted model
l.u <- -211.79   # log likelihood of the unrestricted model

# clearly the latter is larger, but the question is whether the difference is meaningful enough

library(lmtest)  ### we can use lrtest() in "lmtest" package
lrtest(m2, m3)  ### directly enter two models
```

31:1 (Untitled) R Script

Console ~/

```
>
>
> library(lmtest)  ### we can use lrtest() in "lmtest" package
> lrtest(m2, m3)  ### directly enter two models
Likelihood ratio test

Model 1: gamble ~ sex + status + income
Model 2: gamble ~ sex + income
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -211.79
2   4 -212.00 -1 0.4182     0.5178
>
```

The p value is much larger than 0.05, so the difference is not statistically significant. Therefore, M_2 is not significantly better than M_3.

Deviance

- It is a measure of "error", so the smaller, the better
- Intuition: it is a "likelihood ratio" between our model and the ideal model (saturated model)
- $D = 2\ln L(y|y) - 2\ln L(\hat{\theta}|y)$
- Since $L(y|y) = 1$ and $\ln L(y|y) = \ln 1 = 0$,

$$D = -2(\ln L(\hat{\theta}|y))$$

- When an meaningful explanatory variable is added, the deviance decreases by more than one unit (adding irrelevant variables to a model can still reduce its deviance).

Akaike's Information Criteria (AIC)

- We want a better measure of error than deviance, since even random noise can make deviance decrease.

- So we add a penalty for the model parsimony:

$$AIC = -2 \ln L(\hat{\theta}|y) + 2p = D + 2p$$

  where D is the deviance and p is the number of parameters being estimated.

- The smaller AIC, the better the model.

Bayesian Information Criteria (BIC)

- An alternative to AIC.
- Implement an even harsh penalty with a nonlinear component

$$BIC = -2 \ln L(\hat{\theta}|y) + k \ln(n) = D + k \ln(n)$$

where D is the deviance, k is the number of parameters being estimated, and n is the total number of data points.

- The smaller, the better.

Coming Up

- We already understand the theoretical bases of generalized linear regression and its estimation.
- In Week 5-8, we will discuss a series of generalized linear models in detail:
  - binary outcome models
  - count outcome models
  - categorical outcome models
  - duration outcome models