Lecture 7:
Estimation Methods
Maximum Likelihood & Bayesian Estimation

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 1, 2020

Recap

- Computation
- QR decomposition
- MapReduce and Spark
- Demo `Scraping`
- Message: web scraping involves as much art as it does science

Agenda

Motivation

- Maximum Likelihood is, by far, the most popular technique for deriving estimators
- Developed by Roanld A. Fisher (1890-1962)
- “If Fisher had lived in the era of “apps,” maximum likelihood estimation might have made him a billionaire” (Efron and Tibshiriani, 2016)
- Why? MLE gives “automatically”
  - Unbiasedness
  - Minimum variance

Maximum Likelihood Estimation

Let $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$, the likelihood function is defined by

$$L(\theta|x) = \Pi_{i=1}^{n} f(x_i|\theta) \tag{1}$$

A maximum likelihood estimator of the parameter θ:

$$\hat{\theta}^{MLE} = \underset{\theta \in \Theta}{argmax}\, L(\theta, x) \tag{2}$$

- Intuitively, MLE is a reasonable choice for an estimator.
- MLE is the parameter point for which the observed sample is most likely
- *It is kind of a 'reverse engineering' process: to generate random numbers for a certain distribution you first set parameter values and then get realizations. This is doing the reverse process: first set the realizations and try to get the parameters that are 'most likely' to have generated them*

Maximum Likelihood Estimation

Note that maximizing (1) is the same as maximizing

$$l(\theta|x) = \ln L(\theta|x) = \sum_{i=1}^{n} l_i(x|\theta) \tag{3}$$

Advantages of (3)

- It is easy to see that the **contribution** of observation i to the likelihood is given by $l_i(x|\theta) = \ln f(x_i|\theta)$
- Eq. (1) is also prone to underflow; can be very large or very small number that it cannot easily be represented in a computer.

Maximum Likelihood Estimation

If the likelihood function is differentiable (in θ) a possible candidate for the MLE are the values of θ that solve

$$\frac{\partial L(\theta|x)}{\partial \theta} = 0 \tag{4}$$

- These are only *possible candidates*, this is a necessary condition for a max
- Need to check SOC

Maximum Likelihood Estimation

Let $X_1, \ldots, X_n \sim N(\mu, 1)$. We want to estimate $\theta = \mu$
Here

$$L(\theta|x) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{5}$$

taking logs

$$l(\theta|x) = -\frac{n}{2} log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{6}$$

FOC

$$\frac{\partial l(\theta|x)}{\partial \mu} = 0 \tag{7}$$

Maximum Likelihood Estimation

$$\frac{\partial l\left(\theta|x\right)}{\partial\mu} = 2\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)}{2} = 0 \tag{8}$$

$$\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right) = 0 \tag{9}$$

then

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \tag{10}$$

The MLE is the sample mean. Next we check the SOC

$$\frac{\partial^2 l(\theta|x)}{\partial\theta^2} = -n < 0 \tag{11}$$

We are in a global maximum

Conditional Likelihood

Suppose now, that $f(y,x|\eta)$ is the joint density function of two variables X and Y. Then, it can be decomposed as

$$f(y,x|\eta) = f(y|x,\theta)f(x|\phi) \tag{12}$$

- θ, $\phi \subset \eta$
- The parameter vector of interest is θ
- Maximizing the joint likelihood is achieved through maximizing separately the conditional and the marginal likelihood
- The MLE of θ also maximizes the conditional likelihood
- We can obtain ML estimates by specifying the conditional likelihood only

Example 1

Let $y_i|X_i \sim_{iid} Bernoulli(p)$, where $p = PR(y = 1|X) = F(X\beta)$ and $F(.)$ normal cdf. Then the conditional likelihood is

$$L(\beta, Y) = \Pi_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{13}$$

The log likelihood is then

$$l(\beta, Y) = \sum_{i=1}^{n} (y_i \ln F(X_i\beta) + (1 - y_i)\ln(1 - F(X_i\beta))) \tag{14}$$

Example 1

FOC

$$\frac{\partial l(\beta|y,X)}{\partial\beta}=0 \tag{15}$$

$$\sum_{i=1}^{n} y_i \frac{1}{F(X_i\beta)} f(X_i'\beta)X_i' + \sum_{i=1}^{n}(1-y_i)\frac{1}{(1-F(x_i'\beta))} - f(X_i'\beta)X_i' = 0 \tag{16}$$

$$\vdots$$

$$\sum_{i=1}^{n}\frac{(y_i - F(X_i'\beta))f(X_i'\beta)x_i}{F(X_i'\beta)(1-F(X_i'\beta))} = 0 \tag{17}$$

Note:

- This is a system of K non linear equations with K unknown parameters.
- We cannot explicitly solve for $\hat{\beta}$

Example 2: Linear Regression

Now consider the following linear model

$$y = X\beta + u \quad u \sim_{iid} N(0, \sigma^2 I) \tag{18}$$

Note that $y_i|X_i \sim N(X_i\beta, \sigma^2)$ thus the pdf of $y_i|X$

$$f_i(y_i|\beta, \sigma, X_i) = \frac{1}{(\sqrt{2\pi\sigma^2})} e^{-\frac{1}{2\sigma^2}(y_i - X_i\beta)^2} \tag{19}$$

Example 2: Linear Regression

The contribution to the log likelihood from observation i

$$l_i(y_i|\beta,\sigma,X_i) = -\frac{1}{2}log2\pi - \frac{1}{2}log\sigma^2 - \frac{1}{2\sigma^2}(y_i - X_i\beta)^2 \tag{20}$$

Since we assumed that obs are *iid*, then the log likelihood

$$l(y|\beta,\sigma,X) = -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 \tag{21}$$

$$= -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'y - X\beta) \tag{22}$$

The ML estimators for β and σ result from maximizing this last line

Example 2: Linear Regression

The first step in maximizing $l(y|\beta, \sigma, X)$ is to **concentrate** it with respect to σ

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{2\sigma} - \frac{1}{\sigma^3}(y - X\beta)'y - X\beta) = 0 \tag{23}$$

Solving for σ^2

$$\hat{\sigma}^2(\beta) = \frac{1}{n}(y - X\beta)'y - X\beta) \tag{24}$$

Example 2: Linear Regression

Replacing this in the log likelihood we get the concentrated (profile) likelihood

$$l^c(y|\beta, X) = -\frac{n}{2}log2\pi - \frac{n}{2}log\left(\frac{1}{n}(y - X\beta)'y - X\beta)\right) - \frac{n}{2} \tag{25}$$

1. Get $\hat{\beta}$
2. Replace β in $\hat{\sigma}^2(\beta) = \frac{1}{n}(y - X\beta)'y - X\beta) \rightarrow$ get $\hat{\sigma}^2$

This is not the only way, you could concentrate relative to β first and solve for σ^2

Bayesian Estimation

- The Bayesian approach to stats is fundamentally different from the classical approach we have been taking
- In the classical approach, the parameter θ is thought to be an unknown, but fixed quantity, e.g., $X_i \sim f(\theta)$
- In the Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution (*prior distribution*)
- Then a sample is taken from a population indexed by θ and the prior is updated with this sample
- The resulting updated prior is the *posterior distribution*

Bayesian Estimation

For this updating we use *Bayes Theorem*

$$\pi(\theta|X) = \frac{f(X|\theta)p(\theta)}{m(X)} \tag{26}$$

with $m(X)$ is the marginal distribution of X, i.e.

$$m(X) = \int f(X|\theta)p(\theta)d\theta \tag{27}$$

Bayesian Linear Regression

Consider

$$y_i = \beta x_i + u_i \;\; u_i \sim_{iid} N(0, \sigma^2 I) \tag{28}$$

The likelihood function is

$$f(y|\beta, \sigma, x) = \Pi_{i=1}^{n} \frac{1}{(\sqrt{2\pi\sigma^2})} e^{-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2} \tag{29}$$

Now consider that the prior for β is $N(\beta_0, \tau^2)$

$$p(\beta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\beta - \beta_0)^2} \tag{30}$$

Bayesian Linear Regression

The Posterior distribution then

$$\pi(\beta|y,x) = \frac{f(y,x|\beta)p(\beta)}{m(y,x)} \tag{31}$$

$$= \frac{f(y|x,\beta)f(x|\beta)p(\beta)}{m(y,x)} \tag{32}$$

by assumption $f(x|\beta) = f(x)$

$$= f(y|x,\beta)p(\beta)\frac{f(x)}{m(y,x)} \tag{33}$$

$$\propto f(y|x,\beta)p(\beta) \tag{34}$$

Bayesian Linear Regression (Detour)

Useful Result:

Suppose a density of a random variable θ is proportional to

$$exp\left(\frac{-1}{2}(A\theta^2 + B\theta)\right) \tag{35}$$

Then $\theta \sim N(m, V)$ where

$$m = \frac{-1B}{2A} \quad V = \frac{1}{A} \tag{36}$$

Bayesian Linear Regression (we are back)

$$P(\beta|y,X) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(\frac{-1}{2\sigma^2}\sum(y_i-\beta x_i)^2\right) exp\left(\frac{-1}{2\tau^2}(\beta-\beta_0)^2\right) \quad (37)$$

$$\propto exp\left[\frac{-1}{2}\left(\frac{1}{\sigma^2}\sum(y_i-\beta x_i)^2+\frac{-1}{\tau^2}(\beta-\beta_0)^2\right)\right] \quad (38)$$

Bayesian Linear Regression (we are back)

Using the previous detour

$$A = \frac{1}{\sigma^2}\sum x_i^2 + \frac{1}{\tau^2} \tag{39}$$

$$B = -2\frac{1}{\sigma^2}\sum y_i x_i + \frac{1}{\tau^2}\beta_0 \tag{40}$$

Then $\beta \sim N(m, V)$ with

$$m = \frac{\frac{1}{\sigma^2}\sum y_i x_i + \frac{1}{\tau^2}\beta_0}{\left(\frac{1}{\sigma^2}\sum x_i^2 + \frac{1}{\tau^2}\right)} \tag{41}$$

$$V = \frac{1}{A} \tag{42}$$

Bayesian Linear Regression (we are back)

$$m = \left(\frac{\frac{\sum x_i^2}{\sigma^2}}{\frac{\sum x_i^2}{\sigma^2} + \frac{1}{\tau^2}} \right) \frac{\sum x_i y_i}{\sum x_i^2} + \left(\frac{\frac{1}{\tau^2}}{\frac{\sum x_i^2}{\sigma^2} + \frac{1}{\tau^2}} \right) \beta_0 \tag{43}$$

$$m = \omega \hat{\beta}_{MLE} + (1 - \omega)\beta_0 \tag{44}$$

Remarks

- If prior belief is strong $\tau \downarrow 0 \rightarrow \omega \downarrow 0 \implies m = \beta_0$
- If prior belief is weak $\tau \uparrow \infty \rightarrow \omega \uparrow 1 \implies m = \beta_{MLE}$

Review & Next Steps

- Maximum Likelihood Estimation
- Conditional Maximum Likelihood Estimation
- Bayesian Estimation
- **Next Class:** Cont. Bayesian Stats.
- Questions? Questions about software?

Further Readings

- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.
- Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods (Vol. 5). New York: Oxford University Press.
- Efron, B., & Hastie, T. (2016). Computer age statistical inference (Vol. 5). Cambridge University Press.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Hayashi, F. (2000). Econometrics. 2000. Princeton University Press. Section, 1, 60-69.