Lecture 6:
MLE
Intro To Scraping

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 26, 2021

Recap

CEF

$$\hat{y} = X\beta$$

$$y = \underbrace{X\beta}_{\hat{y}} + u$$

$$\beta = (X'X)^{-1}X'y$$

- Least Square Estimator
- Problem Set ✓
- Github Repos
- Bonus: Contribuir con los tutorials ya sea en `Python` o `R`: 0.25 bonus de la nota final (max 1 punto)

25%

Agenda

1. Motivation
2. Maximum Likelihood Estimation
3. Profile Likelihood
4. Example
5. Computation Newton-Method
6. Motiviation Web Scraping
7. Further Readings

Motivation

$y = X\beta + u$ MCO MM MLE

- Maximum Likelihood is, by far, the most popular technique for deriving estimators
- Developed by Ronald A. Fisher (1890-1962)

the signal & the noise Nate Silver

- "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 $E(\hat{\beta}) = \beta$
  - Minimum variance $V(\tilde{\beta}) - V(\hat{\beta}) \geq 0$

Maximum Likelihood Estimation vs Density

realización

parámetro

- Let X be a random variable with density $f(x,\theta)$
- And $X_1,\dots,X_n$ and iid sample $sim_{iid} f(x|\theta)$
- The likelihood function for X is

\sim

iid < indep ident dist

$f(x|\theta) \to \theta \to$ dimensión K

$$\mathcal{L}(\theta|x):\underbrace{\mathbb{R}^K}_{\text{espacio parámetros}}\to\mathbb{R} \to r \qquad (1)$$

$x \sim N(\mu, \sigma^2)$

$f \sim$ densidad normal

$\theta = \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}$

$K = 2$

- Note the following,

  1. In the density function, θ is taken as given, and x varies
  2. In the likelihood function, x is taken as given, and θ varies

Maximum Likelihood Estimation

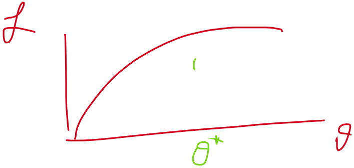
Let $X_1, \ldots, X_n \sim_{iid} f(x|\theta)$, the likelihood function is defined by

$$\mathcal{L}(\theta|x) = \Pi_{i=1}^{n} \mathcal{L}(\theta|x) = \Pi_{i=1}^{n} \underline{f(x_i|\theta)} \tag{2}$$

A maximum likelihood estimator of the parameter θ:

$$\boxed{\hat{\theta}^{MLE}} \neq \underset{\theta \in \Theta}{argmax}\, \underline{\mathcal{L}(\theta|x)}(\cancel{\theta, x}) \tag{3}$$

not $\mathcal{L}$



Maximum Likelihood Estimation

norder (5)

Example

- Let $X \sim N(\mu, \sigma^2)$
- Note that $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}_{2\times 1}$
- Then
  - $f(x|\theta) : \mathbb{R} \to \mathbb{R}$ and $f(x, \theta) = \underbrace{\frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$
  - $\mathcal{L}(\theta|x) : \mathbb{R}^2 \to \mathbb{R}$ and $\mathcal{L}(\theta|x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$
- Intuition?

Maximum Likelihood Estimation

- 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*

$x \leftarrow rnorm(\overset{100}{n}, \mu, \sigma^2)$

x_1
x_n
?
(μ, σ^2)

Maximum Likelihood Estimation

$x_1, \quad \underline{x_1} \sim_{iid} f(x, \theta)$

Note that maximizing

$$\underline{\mathcal{L}(\theta|x)} = \underline{\Pi_{i=1}^n f(x_i|\theta)} \tag{4}$$

is the same as maximizing

$$\underline{l(\theta|x)} = \ln \mathcal{L}(\theta|x) \boxed{= \sum_{i=1}^n l_i(x|\theta)} \tag{5}$$

✓

Advantages of (5)

- 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. (4) 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

$$\boxed{\frac{\partial \mathcal{L}(\theta|x)}{\partial\theta} = 0} \tag{6}$$

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

Maximum Likelihood Estimation

Let $X_1, \dots, X_n \underset{iid}{\sim} N(\mu, 1)$. We want to estimate $\theta = \mu$
Here

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

\mathcal{L}

$$\mathcal{L}(\theta|x) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{7}$$

taking logs

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

FOC

$$\boxed{\frac{\partial l\left(\theta|x\right)}{\partial \mu} = 0} \tag{9}$$

Maximum Likelihood Estimation

$$\frac{\partial l(\theta|x)}{\partial \mu} = \cancel{2}\frac{\sum_{i=1}^{n}(x_i-\mu)}{\cancel{2}} = 0 \tag{10}$$

$$\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0 \tag{11}$$

$$\Sigma x_i - \Sigma \mu = 0$$

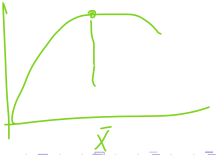
$$\Sigma x_i - \underbrace{n\mu}_{} = 0$$

then

$$\boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}} = \boxed{\bar{x}} \tag{12}$$

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

$$\frac{\partial^2 l(\theta|x)}{\partial\theta^2} = \underbrace{-n}_{} < 0 \tag{13}$$



We are in a global maximum

Concentrated/Profile 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

chain rule di prob

$$y \rightarrow \theta$$
$$X \rightarrow \phi$$

$$\boxed{f(y,x|\eta)} = \underbrace{f(y|x,\boxed{\theta})}_{\text{cond}}\underbrace{f(x|\phi)}_{\text{marginal}} \tag{14}$$

- $\underbrace{\theta, \phi}\subset \eta$
- The parameter vector of interest is θ and θ and ϕ are functionally unrelated
- 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, reducing computational burden of the original problems

Example 1: Linear Regression

$x_1, \quad x_n \sim N(\mu, 1)$

Now consider the following linear model

X, y [illegible]

$E(u|X) = 0$

$V(u|X) = \sigma^2 I$

$r(X) = K$

N

$$y = X\beta + u \tag{15}$$

$$u \sim_{iid} N(0, \sigma^2 I) \tag{16}$$

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

$u \sim N(0, \sigma^2 I)$

$y \sim N(X\beta, \sigma^2 I)$

$$f_i(y_i|\beta,\sigma,X_i) = \frac{1}{(\sqrt{2\pi\sigma^2})} e^{-\frac{1}{2\sigma^2}(y_i - \overset{\mu}{X_i\beta})^2} \tag{17}$$

u_i

$u = y - X\beta$

$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(u_i)^2}$

Example 1: Linear Regression

$\sigma^2 = 1$

The contribution to the log likelihood from observation *i*

$$l_i(y_i|\beta,\sigma,X_i) = -\frac{1}{2}log2\pi - \underbrace{\frac{1}{2}log\sigma^2} - \frac{1}{2\sigma^2}(y_i - X_i\beta)^2 \tag{18}$$

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

$$l(y|\beta,\sigma,X) = -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{n}(y_i - X_i\beta)^2} \tag{19}$$

$$= -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}\underbrace{(y - X\beta)'(y - X\beta)}_{SSR} \tag{20}$$

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

Example 1: Linear Regression

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

$$\frac{\partial l}{\partial\sigma} = -\frac{n}{2\sigma} - \frac{1}{\sigma^3}(y-X\beta)'y-X\beta) = 0 \qquad (21)$$

Solving for σ^2

$$+\frac{n}{2\sigma} = \frac{1}{\sigma^3}(y-X\beta)'(y-X\beta)$$

$$\boxed{\hat{\sigma}^2(\beta) = \frac{1}{n}(y-X\beta)'y-X\beta)} \qquad (22)$$

$$\sigma^2 = \frac{1}{n}\sum \underbrace{(y_i - x_i\beta_2)}_{u_i}^2$$

$$\sigma^2 = \frac{1}{(n-1)}\sum(\;)^2$$

Example 1: Linear Regression

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

$$l^{c}(y|\beta, X) = -\frac{n}{2}log2\pi - \frac{n}{2}log\left(\underbrace{\frac{1}{n}(y - X\beta)'(y - X\beta)}\right) - \underline{\frac{n}{2}} \tag{23}$$

$\sim SSR$

$\hat{\beta} = (X'X)^{-1}X'y$

1. Get $\underline{\hat{\beta}}$
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

$\beta(\sigma)$ H/W

Example 2

$\begin{cases} 0 \\ 1 \end{cases}$

F( ) Logistico → Logit

F( ) Normal → Probit

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

$= \prod_{i/1} p_i^{y_i} \prod_{i/0} (1-p_i)^{(1-y_i)}$ t/ w

$$L(\beta, Y) = \Pi_{i=1}^{n} p_i^{y_i}(1 - p_i)^{\underline{1-y_i}} \tag{24}$$

The log likelihood is then

$p = F(X\beta)$

$$l(\beta, Y) = \underbrace{\sum_{i=1}^{n} (y_i \ln \underline{F(X_i\beta)} + (1 - y_i) \ln(1 - F(X_i\beta)))}_{\text{contrib del elemento } i} \tag{25}$$

PS 3

Example 2

FOC

[redacted] (26)

$$\frac{\partial l(\beta|y,X)}{\partial \beta} = 0 \tag{27}$$

→ gradiente

$$\sum_{i=1}^{n} y_i \frac{1}{F(X_i\beta)} \underbrace{f(X_i'\beta)} X_i' + \sum_{i=1}^{n} (1-y_i) \frac{1}{(1-F(x_i'\beta))} - f(X_i'\beta) X_i' = 0 \tag{28}$$

$\vdots$

Example 2

$x_i = (x_i \quad x_k)$

HW

$\vdots$

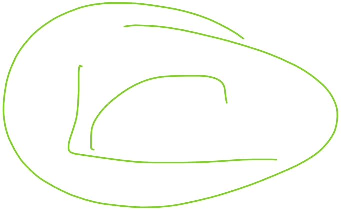
$$\boxed{\sum_{i=1}^{n} \frac{(y_i - F(X_i'\beta))f(X_i'\beta)\cancel{x_i}'}{F(X_i'\beta)(1-F(X_i'\beta))} = 0} \qquad (29)$$

$$\frac{(y_i - F(x_i'\beta))\, f(x_i'\beta)}{F(x_i\beta)\,(1-F(x_i'\beta))}\, x_i' \quad k$$

Note:

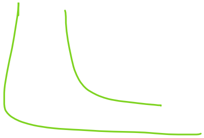
- This is a system of K non linear equations with K unknown parameters.
- We cannot explicitly solve for $\hat{\beta}$

Computation Newton-Method



- Here we can use Newton's Method

$$\theta' = \theta - \underbrace{H^{-1}}_{\substack{\varepsilon \\ \eta}}\underbrace{\nabla_\theta f(\theta)} \tag{30}$$



- Note the difference with Gradient Descent
- H^{-1} should be negative definite, why?

Computation Newton-Method

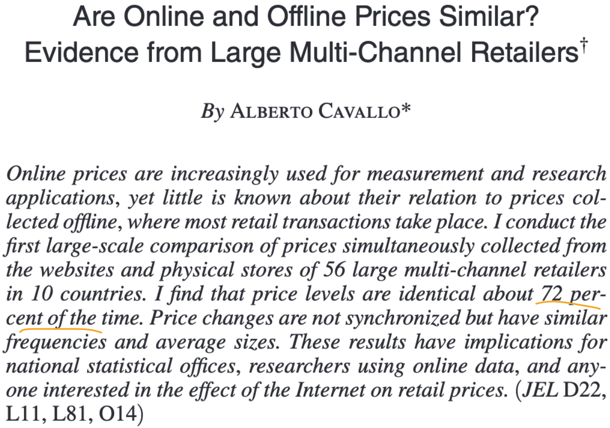
- Newton's Method usually does not work well, when H^{-1} is not n.d.
- Quasi-Newton Method

$$\boxed{\theta' = \theta + \eta D^{-1} \nabla_\theta f(\theta)} \tag{31}$$

- η is an scalar determined at each step Davidson & McKinnon
- $D \approx -H$ so it's always positive definite Hukl
- When the loglikelihood function is globally concave and not too flat, maximizing it is usually quite easy.
- When the loglikelihood function has several local maxima, doing so can be very difficult

Motiviation Web Scraping

billion price project

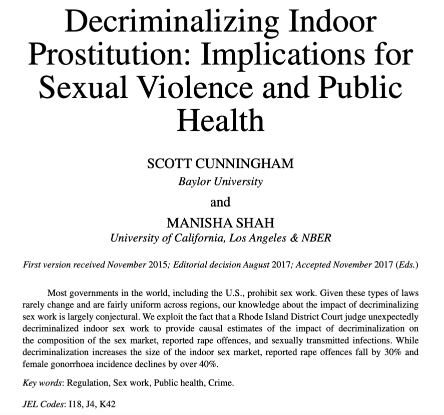


Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers†

By Alberto Cavallo*

Online prices are increasingly used for measurement and research applications, yet little is known about their relation to prices collected offline, where most retail transactions take place. I conduct the first large-scale comparison of prices simultaneously collected from the websites and physical stores of 56 large multi-channel retailers in 10 countries. I find that price levels are identical about 72 percent of the time. Price changes are not synchronized but have similar frequencies and average sizes. These results have implications for national statistical offices, researchers using online data, and anyone interested in the effect of the Internet on retail prices. (*JEL* D22, L11, L81, O14)

Motiviation Webscraping



Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health

SCOTT CUNNINGHAM
Baylor University

and

MANISHA SHAH
University of California, Los Angeles & NBER

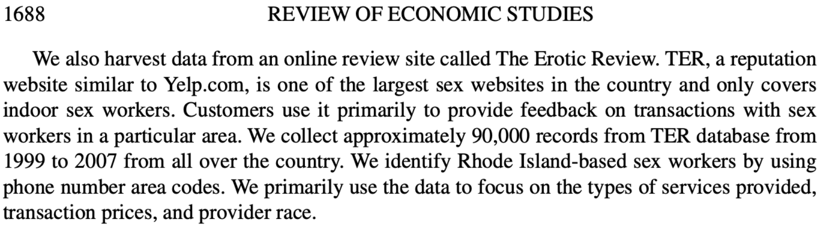
First version received November 2015*; Editorial decision August* 2017*; Accepted November* 2017 (*Eds.*)

Most governments in the world, including the U.S., prohibit sex work. Given these types of laws rarely change and are fairly uniform across regions, our knowledge about the impact of decriminalizing sex work is largely conjectural. We exploit the fact that a Rhode Island District Court judge unexpectedly decriminalized indoor sex work to provide causal estimates of the impact of decriminalization on the composition of the sex market, reported rape offences, and sexually transmitted infections. While decriminalization increases the size of the indoor sex market, reported rape offences fall by 30% and female gonorrhoea incidence declines by over 40%.

Key words: Regulation, Sex work, Public health, Crime.

JEL Codes: I18, J4, K42

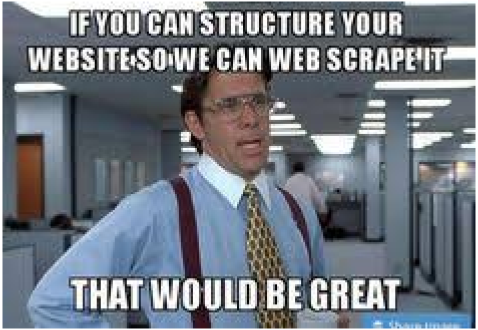
Motiviation Webscraping



1688 REVIEW OF ECONOMIC STUDIES

We also harvest data from an online review site called The Erotic Review. TER, a reputation website similar to Yelp.com, is one of the largest sex websites in the country and only covers indoor sex workers. Customers use it primarily to provide feedback on transactions with sex workers in a particular area. We collect approximately 90,000 records from TER database from 1999 to 2007 from all over the country. We identify Rhode Island-based sex workers by using phone number area codes. We primarily use the data to focus on the types of services provided, transaction prices, and provider race.

Webscraping basics



Webscraping basics

- How to get data, or ”content”, off the web and onto our computers.
- If you see it in your browser it exists somewhere
- To be “successful” one must have a working knowledge on:
  - how web pages display content (Hyper Text Markup Language or HTML)
  - where is the content “located”
    1. Server side
    2. Client side
  - The good news is that both server-side and client-side websites allow for web scraping

Caveat: ethical and legal limitations

- Just because you *can* scrape it, doesn't mean you *should*.
- Check `The Robots Exclusion Protocol` of a website, adding ``/robots.txt'' to the website's URL
  1. User-agent: the type of robots to which the section applies
  2. Disallow: directories/prefixes of the website not allowed to robots
  3. Allow: sections of the website allowed to robots
- `robots.txt` is de facto standard (see `http://www.robotstxt.org`)
- Also always check the terms and conditions and what they say about scraping
- Remember the immortal words of uncle Ben: "with great power comes great responsibility"

Review & Next Steps

- Maximum Likelihood Estimation
- Conditional Maximum Likelihood Estimation
- Intro to Web Scraping
  - *web scraping involves as much art as it does science*
- **Next Class:** 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.
- Web Scrapping slides from Fernandez Villaverde J., Guerrón P. & Zarruk Valencia, D.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- `Webscraping` tutorial from Prof. Grant McDermott.