Quant III

Lab 2: Maximum Likelihood Estimation

Junlong Aaron Zhou

September 17, 2020

Outline

- MLE
- Gradient Descent
- `optim` function

Likelihood

- What is likelihood?
- $L(\theta; x_i) = p_\theta(x_i)$
- NB: It's a function of θ instead of x_i.

Likelihood

- What is likelihood?
- $L(\theta; x_i) = p_\theta(x_i)$
- NB: It's a function of θ instead of x_i.
    - If θ is fixed and we change x_i, we have the density function of x_i
    - If x_i is fixed and we change θ, we are working on likelihood function.

Likelihood ctd.

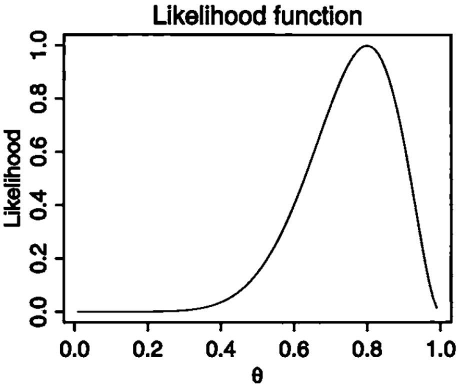

Figure 2.1: ***Likelihood function of the success probability θ in a binomial experiment with $n = 10$ and $x = 8$. The function is normalized to have unit maximum.***

Figure 1: Likelihood

Likelihood ctd..

- $L(\theta; X) = \prod_{i=1}^{n} p_\theta(x_i)$
- We want to know the θ that makes the data most likely to appear.
- Therefore: Maximal likelihood.
- It could be solved analytically.

Likelihood ctd. . .

Eg. Binomial Distribution

$$
\begin{aligned}
y_i &\sim Bernoulli(\theta) \\
L(p; y_i) &= p_\theta(y_i) = \theta^{y_i}(1-\theta)^{(1-y_i)} \\
L(p; Y) &= \prod_{i=1}^{n}(\theta^{y_i}(1-\theta)^{(1-y_i)}) \\
logL &= \sum_{i=1}^{n}(y_i log(\theta) + (1-y_i) log(1-\theta)) \\
F.O.C: \frac{\partial logL}{\partial \theta} = 0 &= \frac{\sum y_i}{\theta} - \frac{\sum(1-y_i)}{1-\theta} \\
\hat{\theta}_{MLE} &= \frac{\sum y_i}{n}
\end{aligned}
$$

GLM

- θ_i can vary across i
- We think $y_i|\theta_i \sim F(\theta_i)$
- $\theta_i = g^{-1}(\eta_i)$
- $\eta_i = X_i\beta$

GLM ctd.

- Why do we have link function?
- Data $X_i \to$ Linear predictor $\eta_i \to$ Link Function $\to \theta_i \to$ Response Y_i
- E.g. Normal Distribution: $y_i|X_i \sim N(X_i\beta, \sigma^2)$
- $X_i \to \eta_i = X_i\beta$
- Write $\eta_i = g(\mu_i)$
- Equivalent to: $\mu_i = g^{-1}(X_i\beta)$
- $y_i \sim N(\mu_i, \sigma^2)$

GLM ctd..

- E.g. In Bernoulli, $\eta_i = X_i\beta = ln(\frac{p_i}{1-p_i})$
- We define $g(p_i) = ln(\frac{p_i}{1-p_i}) = \eta_i$
- Rewrite: $p_i = \frac{1}{1+e^{-\eta_i}}$
- Equivalent to: $p_i = g^{-1}(\eta_i) = \frac{1}{1+e^{-\eta_i}}$

GLM ctd...

- Intuitively: we transform a linear combination $X_i\beta \in R$ to a parameter space ($(0,1)$ in Bernoulli case)
- Also: we are modeling log odds: $log\frac{Pr(Y_i=1|X_i)}{Pr(Y_i=0|X_i)} = X_i\beta$
- In theory: check [**"Exponential Family"**] . (Only when you are super interested)

Numerical Method: Gradient Descent

- How to get β in GLM case?

Numerical Method: Gradient Descent

- How to get β in GLM case?
- MLE!
- When we have multiple parameter, like β, we may not have analytical solution.
- We need numerical method.

Gradient Descent ctd.

- Minimize $f(\theta)$

Gradient Descent ctd.

- Minimize $f(\theta)$
- Taylor Series Expansion!
- $f(\theta) \approx f(\theta_0) + f'(\theta_0)(\theta - \theta_0)$
- Rewrite $\theta - \theta_0 = \alpha * v,\ \alpha > 0, |v| = 1$
- $f(\theta) - f(\theta_0) \approx f'(\theta_0)(\alpha * v)$
- We want $f(\theta) - f(\theta_0) < 0$ and $f(\theta)$ being as smaller as possible
- v points to the opposite direction of $f'(\theta_0)$!

Gradient Descent ctd.

- What is $f'(\theta)$ when θ is multi-dimensional vector $\vec{\theta}$?

Gradient Descent ctd.

- What is $f'(\theta)$ when θ is multi-dimensional vector $\vec{\theta}$?
- Gradient!

Gradient Descent ctd.

- What is $f'(\theta)$ when θ is multi-dimensional vector $\vec{\theta}$?
- Gradient!
- $f(\vec{\theta}) \approx f(\vec{\theta_0}) + \nabla f(\vec{\theta_0})(\vec{\theta} - \vec{\theta_0})$
- $f(\vec{\theta}) - f(\vec{\theta_0}) \approx \nabla f(\vec{\theta_0})(\alpha \vec{v})$
- $\vec{v}$ points to the opposite direction of $\nabla f(\vec{\theta_0})$!
- Intuition: $\vec{\alpha}\vec{\beta} = |\vec{\alpha}||\vec{\beta}| cos(\vec{\alpha}, \vec{\beta})$

GD & GA

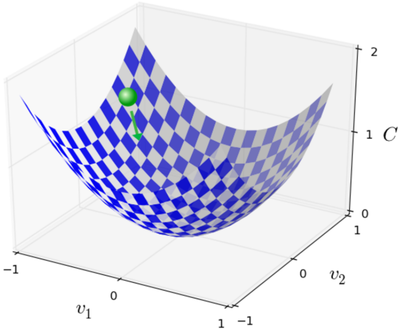

Figure 2: Gradient Descent

GD & GA

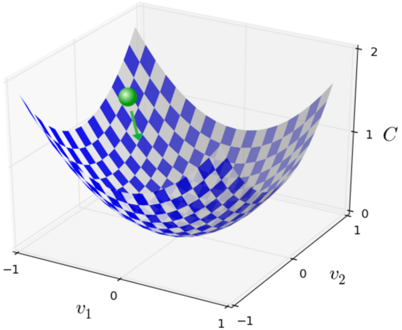

Figure 2: Gradient Descent

- Maximization: Follow the ~white rabbit~ gradient!
- Minimization: Go in the opposite direction!

GD & GA

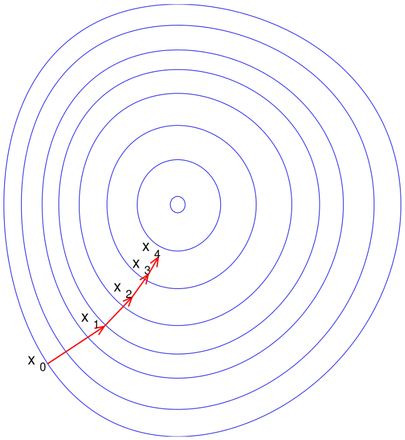

Figure 3: Gradient Descent

More GD

- In MLE $f(\theta)$ is likelihood function

More GD

- In MLE $f(\theta)$ is likelihood function
- But $f(\theta)$ can be any *loss function*
- In classification problem, you may see Loss $= \sum \mathbb{I}(\hat{y}_i \neq y_i)$
- In OLS, Loss $= \sum(\hat{y}_i - y_i)^2$
- GD theoretically can be used to solve all minimization/maximization problem, as long as single modal.

More GD

- In MLE $f(\theta)$ is likelihood function
- But $f(\theta)$ can be any *loss function*
- In classification problem, you may see Loss $= \sum \mathbb{I}(\hat{y}_i \neq y_i)$
- In OLS, Loss $= \sum(\hat{y}_i - y_i)^2$
- GD theoretically can be used to solve all minimization/maximization problem, as long as single modal.

- We need to calculate $L(x_i)$ for all i
- Batch GD
- Stochastic GD
- We don't worry about this in most of cases

(Quasi) Newton Methods

- For a well-behaved function, solving maximize $f(\theta)$ is same as solving $f'(\theta) = 0$
- Taylor Series Expansion again!
- $f'(\theta) \approx f'(\theta_0) + f''(\theta_0)(\theta - \theta_0)$
- $\theta = \theta_0 - \frac{f'(\theta_0)}{f''(\theta_0)}$

Newton's Method ctd.

- What is $f''(\theta_0)$ when θ is multi-dimensional vector $\vec{\theta}$?
- Hessian!
- $f(\vec{\theta}) \approx f(\vec{\theta_0}) + \nabla f(\vec{\theta_0})(\vec{\theta} - \vec{\theta_0}) + \frac{1}{2}(\vec{\theta} - \vec{\theta_0})^T H_{\theta_0}(\vec{\theta} - \vec{\theta_0})$
- $\nabla f(\vec{\theta}) \approx \nabla f(\vec{\theta_0}) + H_{\theta_0}(\vec{\theta} - \vec{\theta_0})$
- $\vec{\theta} = \vec{\theta_0} - H_{\theta_0}^{-1} \nabla f(\vec{\theta_0})$

Quasi-Newton Method

- Hessian may be hard to calculated
- We approximate Hessian under quasi-Newton condition
- Other optimization algorithm are available.
- Google them if you want

Optim Function

```r
optim(par,  fn, gr = NULL, ...,
  method = c("Nelder-Mead", "BFGS", "CG",
  "L-BFGS-B", "SANN", "Brent"),
  lower = -Inf, upper = Inf,
  control = list(), hessian = FALSE)
```

Optim ctd

- Write the function you want to minimize (*fn*)
- Initiate parameters (*par*)
- Return a list

Fisher Information

- $L(\theta; X) = \prod f_\theta(x_i)$
- Log likelihood: $l(\theta; X) = \sum logf_\theta(x_i)$
- Score function: $S(\theta; x) = \frac{\partial l(\theta; x)}{\partial \theta}$
- Define Fisher Information $I(\theta) = E(S(\theta)^2)$

Fisher Information ctd.

- Some conditions
- $E(S(\theta)) = 0$
- $E(\frac{\partial^2 l(\theta;x)}{\partial\theta^2}) = -I(\theta)$
- NB: Left hand side is expected Hessian Matrix

Condition 1

Under some regularity conditions:

$$
\begin{aligned}
E(S(\theta)) &= \int \frac{\partial l(\theta; x)}{\partial \theta} f(x) dx \\
&= \int \frac{\partial log f(x)}{\partial \theta} f(x) dx \\
&= \int \frac{\partial f(x)}{\partial \theta} \frac{1}{f(x)} f(x) dx \\
&= \int \frac{\partial f(x)}{\partial \theta} dx \\
&= \frac{\partial}{\partial \theta} \int f(x) dx \\
&= \frac{\partial}{\partial \theta} 1 = 0
\end{aligned}
$$

Condition 2

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}E(S) = \frac{\partial}{\partial\theta}\int \frac{\partial l}{\partial\theta}f(x)dx \\
&= \int \frac{\partial}{\partial\theta}\{\frac{\partial l}{\partial\theta}f(x)\}dx = \int\{\frac{\partial^2 l}{\partial\theta^2}f(x) + \frac{\partial l}{\partial\theta}\frac{\partial f(x)}{\partial\theta}\}dx \\
&= \int\{\frac{\partial^2 l}{\partial\theta^2}f(x) + \frac{\partial l}{\partial\theta}\frac{\partial L(x)}{\partial\theta}\}dx \\
&= \int\{\frac{\partial^2 l}{\partial\theta^2}f(x) + \frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta}f(x)\}dx \\
&= \int\{\frac{\partial^2 l}{\partial\theta^2}f(x) + \int \frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\theta}f(x)\}dx \\
&= E(\frac{\partial^2 l}{\partial\theta^2}) + E((\frac{\partial l}{\partial\theta})^2)
\end{aligned}
$$

$$
E(\frac{\partial^2 l(\theta;x)}{\partial\theta^2}) = -I(\theta)
$$

Variance of MLE Estimator

- $S(\hat{\theta_{MLE}}) \approx S(\theta) + S'(\theta)(\hat{\theta_{MLE}} - \theta)$
- $\hat{\theta_{MLE}} - \theta \approx \frac{-S_n(\theta)}{H_n(\theta)}$
- Plug in the conditions
- $\sqrt{n}(\hat{\theta_{MLE}} - \theta) \sim N(0, \frac{1}{I(\theta)})$
- $Var(\hat{\theta_{MLE}}) = \frac{1}{nI(\theta)} = \frac{1}{-H_n(\theta)}$
- NB: Difference between expected Hessian and observed Hessian.
- Variance of each parameter is the inverse of $diag(-H_n)$

Some Calculation

$$\hat{\theta_{MLE}} - \theta \approx \frac{-S(\theta)}{H(\theta)}$$

$$\text{By CLT: } S(\theta) \xrightarrow{d} N(E(S(\theta)), I(\theta))$$

$$\text{By LLN: } H(\theta) \xrightarrow{p} -I(\theta)$$

$$\therefore \text{By Slutsky's theorem: } \frac{-S(\theta)}{H(\theta)} \xrightarrow{d} N(0, \frac{1}{I(\theta)})$$