우도와 모수의 추론

Jong-June Jeon [1]

[1]Department of Statistics
University of Seoul

August 29, 2017

우도와 모수의 추론

우도함수: 우도함수는 데이터가 주어진 모수에 대한 함수다.

- $Y_i \underset{iid}{\sim} \text{Bernoulli}(\theta)$ 라고 가정하자. 우리는 Y_i의 pdf 가 다음과 같이 주어진다는 것을 알고 있다.

$$\Pr(Y_i = y) = f(y;\theta) = \theta^y (1-\theta)^{1-y}$$

- $Y_i = y_i$ $(i = 1, \cdots, n)$로 확률변수의 관측값이 주어져있다고 하자.
- 우도함수는 다음과 같이 주어진다.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(y_i;\theta) = \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i}$$

- 로그우도함수는 다음과 같이 정의한다.

$$l(\theta) = \log \mathcal{L}(\theta)$$

우도함수

- $Y_i \underset{iid}{\sim} N(\mu, \sigma^2)$ 이라고 가정하자. 우리는 Y_i의 pdf 가 다음과 같이 주어진다는 것을 알고 있다.

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

- $Y_i = y_i$ $(i = 1, \cdots, n)$로 확률변수의 관측값이 주어져있다고 하자.
- 편의상 (μ, σ^2)을 θ로 표시하면, 우도함수는 다음과 같이 주어진다.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

우도함수와 최대우도추정량

- 어떤 확률변수 X의 확률모형(확률밀도함수) $f(x;\theta)$가 주어진 경우, n개의 랜덤샘플 $x_1, \cdots, x_n$에 기반한 우도함수는 일반적으로

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

  로 주어진다.
- 최대우도추정량(maximum likelihood estimator:MLE)는 $\mathcal{L}(\theta)$를 최대화 하는 모수 θ값이다.
- MLE는 로그우도함수의 최대값과 같다.

예제

- 베르누이 실험에서 랜덤샘플 $(1, 1, 0, 1, 1)$을 관찰하였다.
- R을 이용하여 우도함수를 그려라.
- R을 이용하여 로그우도함수를 그려라.

Rcode

```
> y = c(1,1,0,1,1)
> dbinom(y, size = 1, prob = 0.5, log = TRUE )
> loglike  <- like <- c()
> theta.vec <- seq(0,1,length = 100)
> for (i in 1:100)
+ {
+   theta <- theta.vec[i]
+   like[i] <- prod(dbinom(y, size = 1, prob = theta,
+                 log = FALSE ))
+   loglike[i] <- sum(dbinom(y, size = 1, prob = theta,
+                 log = TRUE ))
+ }
> plot(theta.vec, loglike, type = 'l', col = 'blue')
> plot(theta.vec, like, type = 'l', col = 'blue')
```

예제

- 분산이 1인 정규분포를 따르는 확률변수의 랜덤샘플 $(0.1, 0.5, 0.3, 0.15, 0.2)$을 관찰하였다.
- R을 이용하여 우도함수를 그려라.
- R을 이용하여 로그우도함수를 그려라.

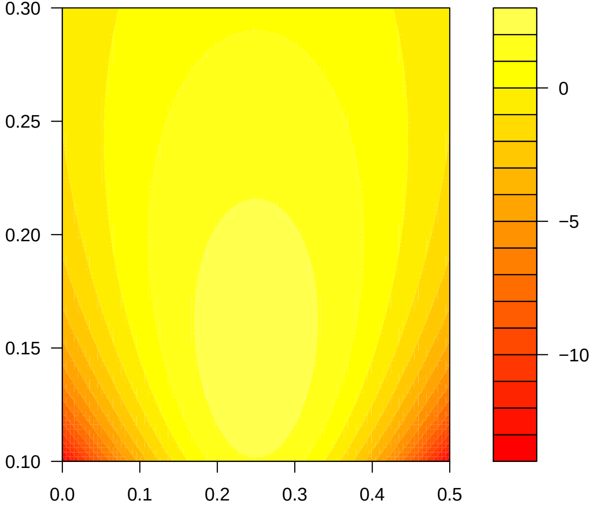
예제

- 분산이 알려지지 않은 정규분포를 따르는 확률변수의 랜덤샘플 $(0.1, 0.5, 0.3, 0.15, 0.2)$을 관찰하였다.
- R을 이용하여 로그우도함수를 그려라.

Rcode

```
> n = 100
> y<- c(0.1,0.5,0.3,0.15,0.2)
> sigma.vec <- seq(0.1,0.3, length = n)
> mu.vec <-  seq(0,0.5, length = n)
> z <- matrix(0,n,n)
> for (i in 1:n)
+   for (j in 1:n)
+     z[i,j] <- sum(dnorm(y,
+                     mu.vec[i], sigma.vec[j],
+                     log = T), na.rm = T)
> filled.contour(mu.vec, sigma.vec, z, nlevels = 20,
+                  col = heat.colors(20))
>
```

Rcode



회귀분석과 우도함수

조건부 분포와 우도함수

- 먼저 Y_i는 분산이 1인 정규분포를 따르는 확률변수라고 하자.
- $X_i \in \mathbb{R}$ 이 주어진 경우 Y_i의 평균은 $\beta_0 + \beta_1 X_i$라고 하자. 즉, X_i 가 주어진 경우 Y_i의 분포는 다음과 같이

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, 1)$$

 주어진다. 여기서 $\mu_i = \beta_0 + \beta_1 X_i$ 라고 하자.
- 랜덤 벡터 (Y_i, X_i)에 대한 관측값 (y_i, x_i)가 주어진 경우 우도함수는 다음과 같이 주어진다.

$$\prod_{i=1}^{n} f(y_i, x_i; \mu_i) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i)^2 \right)$$

조건부 분포와 우도함수

- μ_i 대신 $\beta_0 + \beta_1 x_i$ 를 대입하면 우도함수 식이 아래처럼 표현된다.

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

- 우도함수가 (β_0, β_1) 에 대한 식이므로 우도함수를 아래와 같이 표현할 수 있다.

$$\mathcal{L}(\beta_0, \beta_1) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

- 로그우도함수는 아래와 같이 주어진다.

$$l(\beta_0, \beta_1) = \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{constant}} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

조건부 분포와 우도함수

- $l(\beta_0, \beta_1)$을 최대화 하는 (β_0, β_1)의 값이 MLE다. 한편 MLE는 우도함수의 특성에 의해

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

을 최소화하는 추정량 (LSE)와 같음을 알 수 있다.

예제

Y_i가 규모모수가 1인 라플라스 분포를 따를 때, 회귀분석 모형을 유도해보고 이 모형의 MLE가 절대편이최소추정량 (Leat Absolute Deviation estimator)와 같음을 설명하여라.

로지스틱 회귀분석

- Y_i의 분포가 정규분포가 아닌 베르누이 분포를 따른다고 하자.
- 앞 선 예제의 경우 $E(Y_i|X_i) \in \mathbb{R}$에 대한 모형화를 하였다.
- Y_i가 베르누이 분포를 따른다면 $E(Y_i|X_i) = \Pr(Y_i = 1|X_i) \in [0, 1]$ 이 성립한다. 다시 말해 $E(Y_i|X_i) = \beta_0 + \beta_1 X_i \in \mathbb{R}$ 와 다른 형태의 연결고리를 생각해야 할 것이다.
- 연결함수(link function): $g(E(Y_i|X_i)) = \beta_0 + \beta_1 X_i \in \mathbb{R}$ 관계를 만들어 주는 함수. 우리 경우에는 logit 연결함수를 고려하자.

$$g(x) = \log\left(\frac{x}{1-x}\right)$$

즉,

$$\log\left(\frac{E(Y_i|X_i)}{1-E(Y_i|X_i)}\right) = \log\left(\underbrace{\frac{\Pr(Y_i = 1|X_i)}{1-\Pr(Y_i = 1|X_i)}}_{\text{odds}}\right) = \beta_0 + \beta_1 X_i$$

로지스틱 회귀분석

- 정리하면

$$E(Y_i|X_i) = \Pr(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

- 여기서 $\theta(X_i) = \Pr(Y_i = 1|X_i)$ 이라 놓고 베르누이 pdf를 다시 쓰면

$$\Pr(Y_i = y|X_i = x) = \theta(x)^y (1 - \theta(x))^{1-y}$$

- 관측이 (y_i, x_i) $(i = 1, \cdots, n)$이 주어진 경우 베르누이분포의 loglikelihood는 다음과 같이 써진다.

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp(\beta_0 + \beta_1 x_i)))$$

로지스틱 회귀분석

- 로지스틱 회귀분석의 MLE $= l(\beta_0, \beta_1)$를 최대로 만들어 주는 (β_0, β_1)
- 위의 문장을 다음과 같이 쓴다.

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmax}}\ l(\beta_0, \beta_1)$$

포아송 회귀분석

- Y_i의 분포가 포아송분포를 따른다고 하자
- $\mathrm{E}(Y_i|X_i) \in [0, \infty)$ 이 성립한다.
- 우리 경우에는 log 연결함수를 고려하자.

$$g(x) = \log(x)$$

즉,

$$\log\left(\mathrm{E}(Y_i|X_i)\right) = \beta_0 + \beta_1 X_i$$

- 관측이 (y_i, x_i) $(i = 1, \cdots, n)$이 주어진 경우 포아송 loglikelihood는 다음과 같이 써진다.

$$l(\beta_0, \beta_1) =$$

R 코드

```
> # gaussian
> glm(formula, family = gaussian, data)
> # logistic
> glm(formula, family = binomial(), data)
> # possison
> glm(formula, family = poisson(), data)
```