STA 250 Lecture 3

Rick Wang

Oct.7, 2013

1. Preliminary Knowledge

Data $x \in \chi, x_j \in \chi_j$
Parameter $\theta \in \Theta$
Joint pdf:

$$p(x_1, x_2|\theta)$$

"Marginal" pdf for $x_1|\theta$:

$$p(x_1|\theta) = \int_{\chi_2} p(x_1, x_2|\theta) dx_2$$

"Conditional" pdf for $x_1|x_2, \theta$:

$$p(x_1|x_2, \theta) = \frac{p(x_1, x_2|\theta)}{p(x_2|\theta)} = \frac{p(x_1, x_2|\theta)}{\int p(x_1, x_2|\theta) dx_1}$$

$$\Rightarrow p(x_1, x_2|\theta) = p(x_2|\theta)p(x_1|x_2, \theta) = p(x_1|\theta)p(x_2|x_1, \theta)$$

Furthermore,

$$p(x_1, \cdots, x_n|\theta) = \begin{cases} \Pi_{i=1}^n p(x_i|\theta), \text{ if } x_1, \cdots, x_n \text{ are independent} \\ \Pi_{i=1}^n p(x_i|x_{[0:i-1]}, \theta), \text{ in general} \\ \Pi_{i=1}^n p(x_i|x_{i-1}, \theta), \text{ if } x_1, \cdots, x_n \text{ is Markov} \end{cases}$$

where $x_{[i:j]} = (x_i, x_{i+1}, \cdots, x_j)$.

Example
$Y_{ij}|\lambda_i \overset{indep.}{\sim} Poisson(e_{ij}\lambda_i)$, for $i = 1, \cdots, K, j = 1, \cdots, n_i$; $\lambda_i \overset{iid}{\sim} Gamma(\alpha, \beta)$

Observations: $\{y_{ij}\}$
Unknown: $\{\lambda_i, \alpha, \beta\}$
Known constants: $\{e_{ij}\}$
Model

$$\Pi_{i=1}^{K}\{p(\lambda_i|\alpha,\beta)\Pi_{j=1}^{n_i}p(y_{ij}|\lambda_i)\} = p(\mathbf{y},\boldsymbol{\lambda}|\alpha,\beta)$$

marginalize if only interested in α, β:

$$p(\mathbf{y}|\alpha,\beta) = \int p(\mathbf{y},\boldsymbol{\lambda}|\alpha,\beta)d\boldsymbol{\lambda}$$

2. Maximum Likelihood Estimation

$\hat{\theta}$ is said to be the MLE of θ if

$$\hat{\theta} = \text{argmax}_{\theta\in\Theta}p(y|\theta) = \text{argmax}_{\theta}p(data|parameters)$$

i.e. value of the parameter that makes the data "most likely".
Practically we use $\hat{\theta}_n = \text{argmax}_{\theta}l_n(\theta)$, where $l_n(\theta) = \log p(y_1,\cdots,y_n|\theta)$.
Properties:
(1) Let $y_i{'}s \overset{iid}{\sim} p(y|\theta_0)$, then $\hat{\theta}_n \xrightarrow{P} \theta_0$ as $n\to\infty$, i.e. θ_0 is the "true value" of the parameter, the MLE converges to the true parameter as $n\to\infty$.
(2) $\sqrt{n}(\hat{(\theta)}_n - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta))$, where $I_1^{-1}(\theta) = E[-\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)|\theta]$.
Example:
$Y_1,\cdots,Y_n \overset{iid}{\sim} N(\mu,\sigma^2)$

$$p(y_1,\cdots,y_n|\mu,\sigma^2) = \Pi_{i=1}^{n}p(y_i|\mu,\sigma^2) = \Pi_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

Take log, then take derivatives:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1} ny_i = \bar{y}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1} n(y_i-\bar{y})^2 = \frac{n-1}{n}s^2$$

where s^2 is the sample variance.
Suppose the data were generated from $N(0,1)$, i.e. $\mu_0 = 0, \sigma_0^2 = 1$, then as $n \to 0, \hat{\mu} \to 0, \hat{\sigma}^2 \to 1$.

Confidence Intervals
$C^{1-\alpha}(\mathbf{y})$ is a $100(1-\alpha)\%$ CI for θ if

$$P_\theta(\theta \in C^{1-\alpha}(\mathbf{y})) = 1 - \alpha, \forall\ \theta \in \Theta$$

i.e. under repeated sampling of datasets, $100(1-\alpha)\%$ of intervals will contain the true value of the parameter.
Example: $Y_1, \cdots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$
To estimate μ, we use $\hat{\mu} = \bar{y}$.
A $100(1-\alpha)\%$ CI for μ turns out to be

$$(\bar{y} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{y} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$$

where $t_{n-1,1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})^{th}$ percentile of a t-distribution with $n-1$ degrees of freedom.

Model Misspecification
We use a density $p(\mathbf{y}|\theta)$ to model our data, but what happens if the data comes from a different density, say, g, i.e. the model is wrong. Then

$$\hat{\theta}_n \to \theta^* \text{ as } n \to \infty$$

where θ^* generates the member of $p(\mathbf{y}|\theta)$ that is "closest" to g. Also

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, J_1^{-1}(\theta) V_1(\theta) J_1^{-1}(\theta))$$

where

$$J_1(\theta) = E[-\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)|\theta],\ V_1(\theta) = Var[\frac{\partial}{\partial\theta}\log p(y|\theta)|\theta]$$

If the model is true, $V_1(\theta) = J_1(\theta)$, we get usual result; when model is wrong, we have extra J terms either side, which leads to the so-called "sandwich estimate for the variance of $\hat{\theta}$".
Note:

$$V_n(\theta) = Var[\frac{\partial}{\partial\theta}\log p(y_1, \cdots, y_n|\theta)|\theta]$$

2. Bootstrap

The bootstrap is a method to obtain standard errors for parameter estimates, and it is a very general methodology.
Let $Y_1, \cdots, Y_n \overset{iid}{\sim} F$, we want to estimate $\theta = T(F)$ (population quantity), T is a function applied on F. We use the "plug-in" estimate $\hat{\theta}_n = T(\hat{F}_n) = t(\chi_n)$, where $\hat{F}_n$ is the empirical distribution (CDF) of the data, $\chi_n = (Y_1, \cdots, Y_n)$.
Example:
To estimate median

$$\theta = F^{-1}(0.5) = \text{population median}$$

$$\hat{\theta}_n = \hat{F}_n^{-1}(0.5) = \text{sample median}$$

If we have an estimate $\hat{\theta}_n$, we want to estimate its distribution or specifically standard error.
Idea: resample from empirical distribution to approximate the distribution of $\hat{\theta}_n$ under the true model.
Algorithm: (pseudo R code)

```
for (b in 1:B) {
```

```
    # B is large
    # sample with replacement (size n) from data
    bdata = sample(data, replace = TRUE)
    # compute estimate of θ for the bootstrap dataset
    theta_hat = estimate(bdata)
    estimate_vec[b] = theta_hat
}
```

To estimate standard deviation of $\hat{\theta}_n$, we use

$$SD(\{\hat{\theta}^*_{n,b}, b = 1, \cdots, B\})$$

where $\hat{\theta}^*_{n,b}$ is the estimate of θ from the b^{th} bootstrap dataset.