Упражнение 3. Задача 5: Аппроксимация Лапласа и BIC для Бернуллиевой модели

Даниил Ковех

2025-10-28

Содержание

1. Теория

Рассматриваем модель: $y_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0,1)$. Приор равномерный: $p(\theta) = 1$.

Логарифм правдоподобия:

$$\ell(\theta) = s \log \theta + (N-s)\log(1-\theta), \quad s = \sum_{i=1}^{N} y_i.$$

Максимум достигается в $\hat{\theta} = s/N$.

Наблюдаемая информация (вторая производная с минусом):

$$J(\theta) = -\ell''(\theta) = \frac{s}{\theta^2} + \frac{N-s}{(1-\theta)^2}.$$

Аппроксимация Лапласа:

$$p(y \mid M) \approx \exp\left(\ell(\hat{\theta})\right)\sqrt{\frac{2\pi}{J(\hat{\theta})}}.$$

Логарифмическая форма удобнее:

$$\log p(y \mid M) \approx \ell(\hat{\theta}) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log J(\hat{\theta}).$$

BIC для этой модели:

$$\text{BIC} = -2\ell(\hat{\theta}) + k \log N,$$

где $k = 1$ — число свободных параметров.

2. Жизненный пример

Маркетолог измеряет долю пользователей, которые нажали на кнопку "Купить" при новой верстке сайта. Нужно быстро сравнить гипотезы. Точный подсчёт маргинального правдоподобия требует интегрирования, но Лаплас даёт быструю оценку. BIC — ещё более грубая метрика, но популярная из-за простоты. Мы посчитаем обе величины и сравним.

3. Академическое решение

3.1. 1. Функции для логарифма правдоподобия, информации и аппроксимации

```
log_likelihood <- function(theta_hat, s, N) {
  s * log(theta_hat) + (N - s) * log(1 - theta_hat)
}

observed_information <- function(theta_hat, s, N) {
  s / theta_hat^2 + (N - s) / (1 - theta_hat)^2
}

laplace_log_marginal <- function(s, N) {
  theta_hat <- s / N
  ll <- log_likelihood(theta_hat, s, N)
  info <- observed_information(theta_hat, s, N)
  ll + 0.5 * log(2 * pi) - 0.5 * log(info)
}

bic_value <- function(s, N) {
  theta_hat <- s / N
  -2 * log_likelihood(theta_hat, s, N) + log(N)
}
```

Проверим на одном сценарии:

```
laplace_log_marginal(s = 10, N = 100)
```

```
## [1] -35.09592
```

```
bic_value(s = 10, N = 100)
```

```
## [1] 69.62176
```

3.2. 2. Таблица значений

Условия: вероятность успеха 0.1, значит $s = 0.1N$ (округлено).

```
scenarios <- data.frame(
  N = c(10, 100, 1000, 10000),
  s = c(1, 10, 100, 1000)
)
```

```r
scenarios$theta_hat <- with(scenarios, s / N)
scenarios$log_laplace <- mapply(laplace_log_marginal, s = scenarios$s, N = scenarios$N)
scenarios$neg2log_laplace <- -2 * scenarios$log_laplace
scenarios$BIC <- mapply(bic_value, s = scenarios$s, N = scenarios$N)
scenarios
```

```
##       N    s theta_hat  log_laplace neg2log_laplace         BIC
## 1    10    1       0.1    -4.687157        9.374313    8.804245
## 2   100   10       0.1   -35.095917       70.191833   69.621765
## 3  1000  100       0.1  -328.821885      657.643771  657.073702
## 4 10000 1000       0.1 -3255.719938     6511.439877 6510.869808
```

3.3. 3. Сравнение с точным значением

Используем формулу из задачи 4 для точного логарифма маргинального правдоподобия.

```r
marginal_log_likelihood <- function(s, N, alpha = 1, beta = 1) {
  lgamma(s + alpha) + lgamma(N - s + beta) - lgamma(N + alpha + beta) -
    (lgamma(alpha) + lgamma(beta) - lgamma(alpha + beta))
}

scenarios$log_exact <- mapply(marginal_log_likelihood, s = scenarios$s, N = scenarios$N)
scenarios$neg2log_exact <- -2 * scenarios$log_exact
scenarios$gap_laplace <- scenarios$neg2log_laplace - scenarios$neg2log_exact
scenarios$gap_bic <- scenarios$BIC - scenarios$neg2log_exact
scenarios
```

```
##       N    s theta_hat  log_laplace neg2log_laplace         BIC   log_exact
## 1    10    1       0.1    -4.687157        9.374313    8.804245    -4.70048
## 2   100   10       0.1   -35.095917       70.191833   69.621765   -35.09744
## 3  1000  100       0.1  -328.821885      657.643771  657.073702  -328.82204
## 4 10000 1000       0.1 -3255.719938     6511.439877 6510.869808 -3255.71995
##   neg2log_exact   gap_laplace    gap_bic
## 1      9.400961 -2.664763e-02 -0.5967162
## 2     70.194888 -3.054352e-03 -0.5731229
## 3    657.644084 -3.138210e-04 -0.5703824
## 4   6511.439908 -3.147152e-05 -0.5701000
```

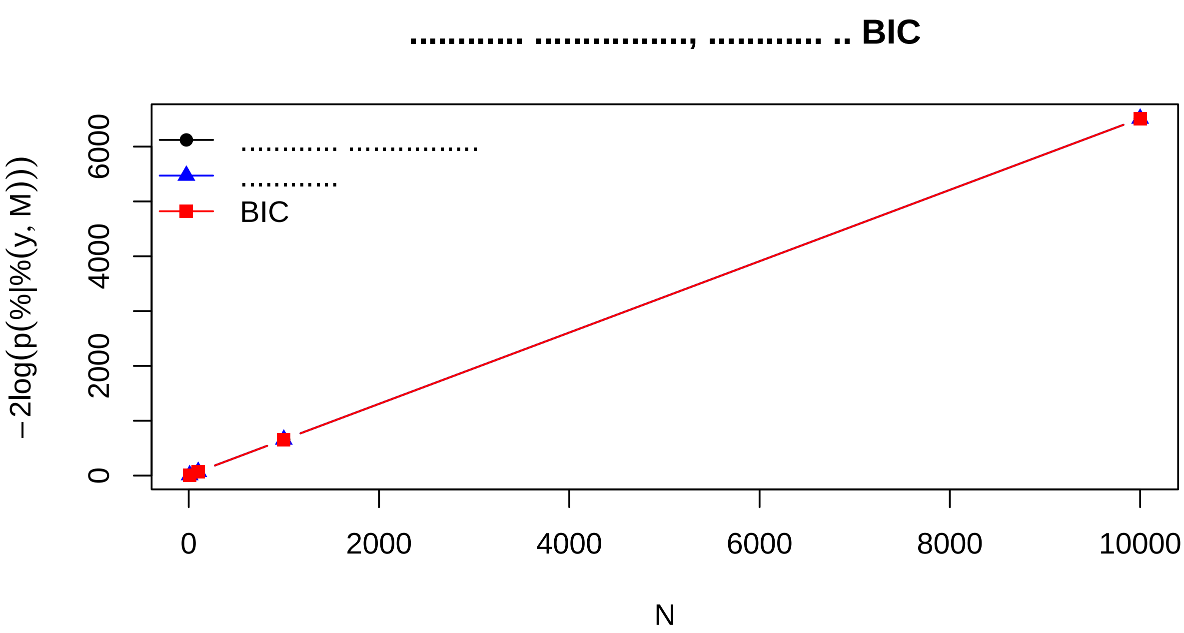
3.4. 4. Интерпретация

- Лаплас даёт очень точную аппроксимацию уже при $N = 100$.
- BIC даёт смещение: он завышает $-2 \log p(y \mid M)$, потому что добавляет штраф $\log N$ за параметр.
- Разница между Лапласом и точным значением убывает при росте N.

3.5. 5. График для визуализации

```r
plot(scenarios$N, scenarios$neg2log_exact,
     type = "b", pch = 16,
     ylim = range(c(scenarios$neg2log_exact, scenarios$neg2log_laplace, scenarios$BIC)),
     xlab = "N", ylab = expression(-2 * log(p(y %|% M))),
     main = "Точное значение, Лаплас и BIC")
lines(scenarios$N, scenarios$neg2log_laplace, type = "b", pch = 17, col = "blue")
lines(scenarios$N, scenarios$BIC, type = "b", pch = 15, col = "red")
legend("topleft",
```

```
       legend = c("Точное значение", "Лаплас", "BIC"),
       col = c("black", "blue", "red"),
       pch = c(16, 17, 15),
       lty = 1,
       bty = "n")
```



3.6. 6. Вывод

- Аппроксимация Лапласа практически совпадает с точным маргинальным правдоподобием.
- BIC — более грубая метрика, но простая в вычислении. Для небольших выборок штраф $\log N$ делает BIC заметно больше точного значения.
- В отличие от BIC, Лаплас учитывает кривизну логарифма правдоподобия (через $J(\hat{\theta})$), поэтому даёт более точный ответ.