Direct Sampling

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

03/22/2021

Simulating from pdf or pmf

- In statistics it is common to encounter integrals that are difficult to solve analytically.
  - Computing expected values or variances
  - Computing the normalizing constant for marginal distributions
  - Finding probabilities when the cdf is unknown
- Encounter these issues often in Bayesian statistics....
- Monte Carlo simulation, in which observations are randomly generated from the model, is helpful to overcome this.

Direct Monte Carlo integration: expected value

- Assume $\theta \sim p(\theta)$, and consider the functional form $g(\theta)$
  - e.g., p may be a posterior distribution $p(\theta \mid \mathbf{y})$
- Assume we can generate samples $\theta_1, \ldots, \theta_N \overset{iid}{\sim} p(\theta)$
  - Using R (and other software) it is easy to generate from known distributions such as the Beta, Binomial, Poisson, Normal, or Gamma.
- Then, by the Law of Large Numbers,

$$\int g(\theta)p(\theta)\, d\theta = Eg(\theta) \approx \frac{1}{N}\sum_{j=1}^{N} g(\theta_j)$$

for large N

Direct Monte Carlo integration: standard error

- Let $\hat{\gamma} = \frac{1}{N}\sum_{j=1}^{N} g(\theta_j)$.
- The variance of $g(\theta)$ under p can be approximated similarly:

$$\text{Var}(g(\theta)) \approx \frac{1}{N-1}\sum_{j=1}^{N}[g(\theta_j) - \hat{\gamma}]^2$$

- We can derive the standard error for $\hat{\gamma}$ using $\text{Var}(\hat{\gamma}) = \text{Var}(g(\theta))/N$:

$$\hat{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)}\sum_{j=1}^{N}[g(\theta_j) - \hat{\gamma}]^2}.$$

  - This does NOT represent uncertainty of $g(\theta)$ implied by p
  - It represents simulation uncertainty in our estimate for $Eg(\theta)$
  - Converges to 0 as we increase number of simulations N
  - By Central Limit Theorem, $\hat{\gamma} \sim N(E(g(\theta)), \hat{se}(\hat{\gamma})^2)$ for large N

Direct Monte Carlo integration

- Consider $g(\theta) = \mathbb{1}_{a<c(\theta)<b}$ for some function $c(\theta)$
- Then, $Eg(\theta) = P(a < c(\theta) < b) \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta_i)$
- We can estimate this probability under simulation:

$$\hat{p} = \frac{\text{number of } c(\theta_j)s \in (a, b)}{N}$$

  and our standard error for this estimate is $\sqrt{\hat{p}(1-\hat{p})/N}$.

- Implication: a histogram of the $c(\theta_j)s$ approximates the density for $c(\theta)$.

Example: Eye exam

- An eye patient is asked to identify shapes from a given distance
- She is shown $n = 20$ shapes
- Jeffreys beta-binomial model for number of shapes she identifies correctly:

$$y \sim \text{Binomial}(20, \theta),$$

$$\theta \sim \text{Beta}(0.5, 0.5).$$

- Assume she identifies 16 shapes correctly.
  - $p(\theta \mid y) = \text{Beta}(16.5, 4.5)$

Example: Eye exam

- Consider the logit transformation $g(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$.
- Approximate distribution of $g(\theta)$ induced by beta posterior.
- For $j = 1, \ldots, N$:
  - Simulate θ_j from $\text{Beta}(16.5, 4.5)$
  - Compute $g(\theta_j) = \log\left(\frac{\theta_j}{1-\theta_j}\right)$
- Consider distribution of $g(\theta_j)s$ for $N = 10000$ sims

Example: Eye exam

- The estimated posterior expected value is

$$\frac{1}{10000}\sum_{j=1}^{N} g(\theta_j) = 1.386$$

- The estimated posterior variance is

$$\frac{1}{N-1}\sum_{j=1}^{N}[g(\theta_j) - \hat{\gamma}]^2 = 0.310$$

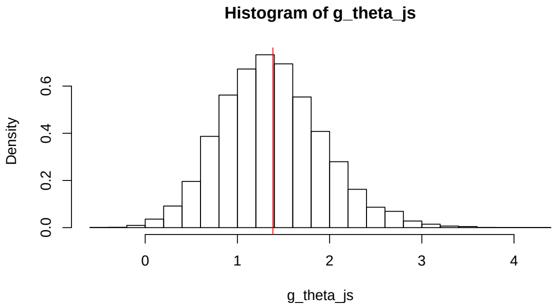
- The standard error for *our estimate* of $Eg(\theta)$ is

$$\sqrt{0.310/10000} = 0.006$$

  - Approximate 95% confidence interval: $1.386 \pm 2 \cdot 0.006$

Example: Eye exam

- Estimated density of $g(\theta)$:



http://www.ericfrazerlock.com/Direct_Sampling_Rcode1.r

Direct sampling: multivariate density

- Let $\theta = (\theta_1, \ldots, \theta_k)$. $p(\theta_1, \theta_2) = p(\theta_2 \mid \theta_1) \cdot p(\theta_1)$
- Note that

$$p(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{k} p(\theta_i \mid \theta_{i-1}, \ldots, \theta_1)$$

$\rightarrow (\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$

- We can draw direct samples $\tilde{\theta}$ from $p(\theta_1, \ldots, \theta_k)$ as follows:
  - Draw $\tilde{\theta}_1$ from $p(\theta_1)$,
  - Draw $\tilde{\theta}_2$ from $p(\theta_2 \mid \tilde{\theta}_1)$,
  - ..., draw $\tilde{\theta}_k$ from $p(\theta_k \mid \tilde{\theta}_{k-1}, \ldots, \tilde{\theta}_1)$.
- Useful for simulating difficult marginal densities
  - $p(\theta_1)$ and $p(\theta_2 \mid \theta_1)$ may be easy to obtain, but not $p(\theta_2)$
  - Here $\tilde{\theta}_2$ is a draw from $p(\theta_2)$.

Example: Body Fat (cont.)

- Recall: % body fat ($BF\%$) measured for 100 adult males.
- Also measured 9 predictor variables
  - *Age*, *Weight*, *Height*; circumference of *neck*, *chest*, *abdomen*, *ankle*, *bicep*, and *wrist*.
- Consider the model

$$\mathbf{y} = \boldsymbol{\beta} X + \boldsymbol{\epsilon}$$

  where
  - $\mathbf{y}$ is population-centered $BF\%$
  - X is the standardized matrix of predictor variables

Example: Body Fat (cont.)

- Used iid normal prior for $\beta_i's$:

$$\beta \sim \text{Normal}(0, 0.62\sigma^2 I)$$

  with $\hat{\tau}^2 = 0.62$ estimated empirically.

- $IG(3, 20)$ prior for σ^2
- Gives the conditional posterior

$$p(\beta \mid \mathbf{y}, \sigma^2) = \text{Normal}\left(\tilde{\beta}, \sigma^2 V_\beta\right)$$

  where $\tilde{\beta} = (X^T X + \frac{1}{0.62})^{-1}(X^T \mathbf{y})$
  and $V_\beta = (X^T X + \frac{1}{0.62})^{-1}$

Example: Body Fat (cont.)

- The marginal posterior for σ^2 is

$$p(\sigma^2 \mid \mathbf{y}) = IG\left(a_n, b_n\right)$$

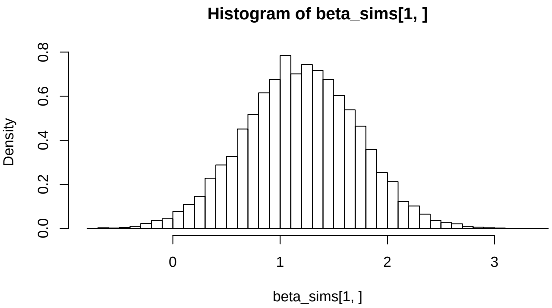
where $a_n = 3 + \frac{100}{2}$ and
$b_n = 20 + \frac{1}{2}[\mathbf{y}^T\mathbf{y} - \tilde{\beta}^T(X^TX + \frac{1}{0.62}I)^{-1}\tilde{\beta}]$

- Direct sampling scheme:
  - For samples $j = 1, \ldots, N = 10000$:
  - Draw σ_j from $p(\sigma^2 \mid \mathbf{y})$
  - Draw β_j from $p(\beta \mid \mathbf{y}, \sigma_j^2)$

http://www.ericfrazerlock.com/More_on_Direct_Sampling_Rcode1.r

Example: Body Fat (cont.)

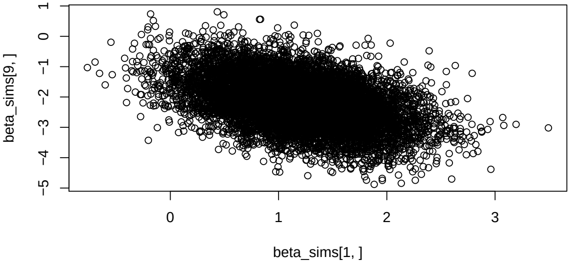
- Simulated marginal posterior of *Age* coefficient β_1:



- 95% credible interval from simulations: $(0.152, 2.220)$
- 95% credible interval from t-dist: $(0.157, 2.229)$

Example: Body Fat (cont.)

- Scatterplot of simulated coefficients for *Age* β_1 and *Wrist* β_2:



http://www.ericfrazerlock.com/More_on_Direct_Sampling_Rcode1.r

- Assume an individual has the following standardized measurements:
  - *Age*: 1.20, *Weight*:0.5, *Height*:-0.4; circumference of *neck*:2.2, *chest*:0.3, *abdomen*:0.8, *ankle*:0.2, *bicep*:0.3, and *wrist:0.6*.
- Denote his vector of predictors as $\mathbf{x}_{101}$
- Simulate from the posterior predictive for y_{101}:

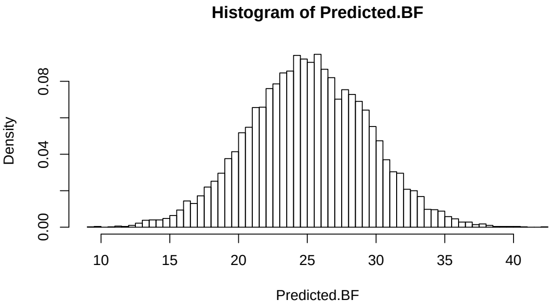
  $P(y_{101} \mid \vec{y}, X_{101})$

  - Draw σ_j, β_j as before
  - Draw $y_{101,j}$ from Normal$(\mathbf{x}_{101}\beta_j, \sigma_j)$

$P(y_{101}, \sigma^2, \beta \mid \vec{y}) = P(\sigma^2 \mid \vec{y}) \cdot P(\beta \mid \sigma^2, \vec{y}) \cdot P(y_{101} \mid \sigma^2, \beta$

Example: Body Fat (cont.)

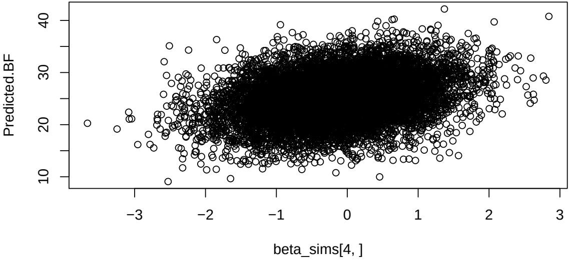
- Histogram of predicted body fat % for individual 101:



http://www.ericfrazerlock.com/More_on_Direct_Sampling_Rcode1.r

Example: Body Fat (cont.)

- Scatterplot of simulated predicted body fat for individual 101, and *Neck* coefficient β_4:



http://www.ericfrazerlock.com/More_on_Direct_Sampling_Rcode1.r