Bayesian Inference

February 14, 2013

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y (the data).

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y (the data).
- Randomly (iid) sample y from some data generating process (DGP): $P(y|\theta)$

Inference

- We want to estimate parametric models of the form:

$$y \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y (the data).
- Randomly (iid) sample y from some data generating process (DGP): $P(y|\theta)$
- Goal is to estimate $P(\theta|y)$ given $P(y|\theta)$ and information about $P(\theta)$

Posterior Inference

- Use Bayes' theorem for $p(\theta|y)$:

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)}$$

Posterior Inference

- Use Bayes' theorem for $p(\theta|y)$:

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)}
\end{aligned}
$$

Posterior Inference

- Use Bayes' theorem for $p(\theta|y)$:

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)}
\end{aligned}
$$

- The Prior: $P(\theta)$
- The Sampling (Data): $P(y|\theta)$
- The Posterior: $P(\theta|y)$
- The Marginal: $P(y)$

Posterior Inference

- Use Bayes' theorem for $p(\theta|y)$:

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)} = \frac{P(\theta)P(y|\theta)}{P(y)}$$

- The Prior: $P(\theta)$
- The Sampling (Data): $P(y|\theta)$
- The Posterior: $P(\theta|y)$
- The Marginal: $P(y)$

The Bayesian Approach

- Use all available information about the model parameters
  - Evidence from the sampling process
  - All prior information

The Bayesian Approach

- Use all available information about the model parameters
  - Evidence from the sampling process
  - All prior information
- Why formalize prior information?
  - Inference drawn from the probability process
  - Incorporating the 'body of prior knowledge'
  - Integral to iterative data analysis and model checking

The Bayesian Approach

- Use all available information about the model parameters
  - Evidence from the sampling process
  - All prior information
- Why formalize prior information?
  - Inference drawn from the probability process
  - Incorporating the 'body of prior knowledge'
  - Integral to iterative data analysis and model checking
- Why to not formalize prior information?
  - Does not emerge from the randomization process
  - Prior information could be explicitly biased
  - Inferences are not invariant to reparameterization (or even different types of diffuse priors)
  - Computationally and mathematically complex

The Bayesian Approach

- "The posterior is proportional to the prior and the likelihood:"

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

The Bayesian Approach

- "The posterior is proportional to the prior and the likelihood:"
$$p(\theta|y) \propto p(\theta)p(y|\theta)$$
- Estimate $\hat{\theta}$ by randomly sampling from the posterior $p(\theta|y)$ distribution
  - Sample from a closed form of $p(\theta|y)$ based on the model and priors
  - Approximate $p(\theta|y)$ numerically or through sampling and monte carlo approaches

A Simple Example: The Prior

- Inference about genetic probability (Gelman et al 2004):
  - Hemophilia exhibits X-chromosomal recessive inheritance
  - A male with the carrier X-chromosome is affected, whereas a women with only one carrier X-chromosome is not

A Simple Example: The Prior

- Inference about genetic probability (Gelman et al 2004):
  - Hemophilia exhibits X-chromosomal recessive inheritance
  - A male with the carrier X-chromosome is affected, whereas a women with only one carrier X-chromosome is not
- Consider a woman whose brother is affected, but whose father is not
  - The woman is either a carrier ($\theta = 1$) or is not ($\theta = 0$)
  - Given no other information, what is the $Pr(\theta = 1)$ for the woman?

A Simple Example: The Data

- The data we collect to update this prior is the affliction status status of her sons
  - Define y_i be 1 if her ith son is afflicted, and 0 otherwise
  - Assume that y_i and y_j are exchangeable for all $i \neq j$, and are independent conditional on θ

A Simple Example: The Data

- The data we collect to update this prior is the affliction status status of her sons
  - Define y_i be 1 if her ith son is afflicted, and 0 otherwise
  - Assume that y_i and y_j are exchangeable for all $i \neq j$, and are independent conditional on θ
- We then have the following likelihoods (for two unafflicted sons):
  - $Pr(y_1 = 0, y_2 = 0|\theta = 1)$?
  - $Pr(y_1 = 0, y_2 = 0|\theta = 0)$?

A Simple Example: The Data

- The data we collect to update this prior is the affliction status status of her sons
  - Define y_i be 1 if her ith son is afflicted, and 0 otherwise
  - Assume that y_i and y_j are exchangeable for all $i \neq j$, and are independent conditional on θ
- We then have the following likelihoods (for two unafflicted sons):
  - $Pr(y_1 = 0, y_2 = 0|\theta = 1)$?
  - $Pr(y_1 = 0, y_2 = 0|\theta = 0)$?
- In this process:
  - What is $Pr(y_1 = 1|\theta = 1)$?
  - What is $Pr(y_1 = 1|\theta = 0)$?

A Simple Example: The Data

- The data we collect to update this prior is the affliction status status of her sons
  - Define y_i be 1 if her ith son is afflicted, and 0 otherwise
  - Assume that y_i and y_j are exchangeable for all $i \neq j$, and are independent conditional on θ
- We then have the following likelihoods (for two unafflicted sons):
  - $Pr(y_1 = 0, y_2 = 0|\theta = 1)$?
  - $Pr(y_1 = 0, y_2 = 0|\theta = 0)$?
- In this process:
  - What is $Pr(y_1 = 1|\theta = 1) = 1/2$
  - What is $Pr(y_1 = 1|\theta = 0) = 0$

A Simple Example: The Data

- The data we collect to update this prior is the affliction status status of her sons
  - Define y_i be 1 if her ith son is afflicted, and 0 otherwise
  - Assume that y_i and y_j are exchangeable for all $i \neq j$, and are independent conditional on θ
- We then have the following likelihoods (for two unafflicted sons):
  - $Pr(y_1 = 0, y_2 = 0|\theta = 1) = (1/2)(1/2) = 1/4$
  - $Pr(y_1 = 0, y_2 = 0|\theta = 0) = (1)(1) = 1$
- In this process:
  - What is $Pr(y_1 = 1|\theta = 1) = 1/2$
  - What is $Pr(y_1 = 1|\theta = 0) = 0$

A Simple Example: The Posterior

- Combine the prior and the evidence about her sons:

$$\begin{aligned} Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)Pr(\theta = 1)}{p(y|\theta = 1)Pr(\theta = 1) + p(y|\theta = 0)Pr(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.2 \end{aligned}$$

- A woman with no affected sons is less likely to be afflicted, 'correcting' the prior in the direction of the data

A Simple Example: The Posterior

- Combine the prior and the evidence about her sons:

$$
\begin{aligned}
Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)Pr(\theta = 1)}{p(y|\theta = 1)Pr(\theta = 1) + p(y|\theta = 0)Pr(\theta = 0)} \\
&= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.2
\end{aligned}
$$

- A woman with no affected sons is less likely to be afflicted, 'correcting' the prior in the direction of the data
- Add more 'data' with a third unaffected son, using the previous posterior as the new prior:

$$
Pr(\theta = 1|y_1, y_2, y_3) = \frac{(0.5)(0.2)}{(0.5)(0.2) + (1.0)(0.8)} = 0.111
$$

A Simple Example: The Posterior

- Combine the prior and the evidence about her sons:

$$
\begin{aligned}
Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)Pr(\theta = 1)}{p(y|\theta = 1)Pr(\theta = 1) + p(y|\theta = 0)Pr(\theta = 0)} \\
&= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.2
\end{aligned}
$$

- A woman with no affected sons is less likely to be afflicted, 'correcting' the prior in the direction of the data
- Add more 'data' with a third unaffected son, using the previous posterior as the new prior:

$$
Pr(\theta = 1|y_1, y_2, y_3) = \frac{(0.5)(0.2)}{(0.5)(0.2) + (1.0)(0.8)} = 0.111
$$

- A third son who IS affected? What is the ML estimate?

Normal Data with Known Variance

- We randomly sample (iid) n units, and observe y_i. Assuming $y \sim N(\beta, \sigma^2)$:

$$
\begin{aligned}
P(y|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\
P(y|\beta,\sigma) &= \prod_{i=1}^{n} N(y_i|\beta,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y_i-\beta}{\sigma}\right)^2\right\}
\end{aligned}
$$

Normal Data with Known Variance

- To estimate β using $p(\beta|y)$, we need to choose a prior distribution $p(\beta)$
- A natural choice is a Gaussian prior: $p(\beta) \sim N(\mu_0, \tau_0^2)$, where μ_0, τ_0^2 are known hyperparemeters. Thus:

$$p(\beta) \quad \propto \quad \exp\left\{-\frac{1}{2\tau_0^2}\left(\beta-\mu_0\right)^2\right\}$$

- This is a conjugate prior, i.e., in the same probability distribution family

Normal Data with Known Variance

- Use this prior to identify an analytical form for $p(\beta|y)$

$$
\begin{aligned}
P(\beta|y) &\propto \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\beta)^2\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\tau_0^2}(\beta-\mu_0)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}(y_i-\beta)^2}{\sigma^2}+\frac{(\beta-\mu_0)^2}{\tau_0^2}\right]\right\}
\end{aligned}
$$

Normal Data with Known Variance

- Combine and factor out terms that do not depend on β

$$
\begin{aligned}
\propto & \quad \exp\left\{-\frac{1}{2}\left[\frac{\sum y_i^2 - 2\beta\sum y_i + n\beta^2}{\sigma^2} + \frac{\beta^2 - 2\beta\mu_0 + \mu_0^2}{\tau_0^2}\right]\right\} \\
\propto & \quad \exp\left\{-\frac{\beta^2}{2\tau_0^2} + \frac{\beta\mu_0}{\tau_0^2} + \frac{\beta n\bar{y}}{\sigma^2} - \frac{n\beta^2}{2\sigma^2}\right\} \\
\propto & \quad \exp\left\{-\frac{\beta^2}{2}\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right) + \beta\left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\right\} \\
\propto & \quad \exp\left\{-\frac{\beta^2}{2\tau_*^2} + \frac{2}{2}\cdot\frac{\beta\mu_*}{\tau_*^2}\right\} = \exp\left\{-\frac{1}{2\tau_*^2}\left(\beta^2 - 2\beta\mu_*\right)\right\}
\end{aligned}
$$

Normal Data with Known Variance

- Combine and factor out terms that do not depend on β

$$
\begin{aligned}
&\propto \quad \exp\left\{-\frac{1}{2}\left[\frac{\sum y_i^2 - 2\beta\sum y_i + n\beta^2}{\sigma^2} + \frac{\beta^2 - 2\beta\mu_0 + \mu_0^2}{\tau_0^2}\right]\right\} \\
&\propto \quad \exp\left\{-\frac{\beta^2}{2\tau_0^2} + \frac{\beta\mu_0}{\tau_0^2} + \frac{\beta n\bar{y}}{\sigma^2} - \frac{n\beta^2}{2\sigma^2}\right\} \\
&\propto \quad \exp\left\{-\frac{\beta^2}{2}\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right) + \beta\left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\right\} \\
&\propto \quad \exp\left\{-\frac{\beta^2}{2\tau_*^2} + \frac{2}{2}\cdot\frac{\beta\mu_*}{\tau_*^2}\right\} = \exp\left\{-\frac{1}{2\tau_*^2}\left(\beta^2 - 2\beta\mu_*\right)\right\}
\end{aligned}
$$

- Completing the square:

$$
\begin{aligned}
&\propto \quad \exp\left\{-\frac{1}{2\tau_*^2}\left(\beta - \mu_*\right)^2 - \mu_*^2\right\} \\
p(\beta|y) &\propto \quad \exp\left\{-\frac{1}{2\tau_*^2}\left(\beta - \mu_*\right)^2\right\}
\end{aligned}
$$

Normal Data with Known Variance

- We should notice that this is implies $\beta|y \sim N(\mu_*, \tau_*^2)$, where μ_* and τ_*^2 are parameterized as:

$$\begin{aligned}\mu_* &= \tau_*^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right) \\ \frac{1}{\tau_*^2} &= \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)\end{aligned}$$

- We see the posterior mean μ_* is a weighted average of the prior mean μ_0 and the data $n\bar{y}$, normalized by the posterior *precision* τ_*^2
- Should also observe that as number of samples n gets large, μ_* is increasingly weighted towards $\bar{y}$

Posterior Estimation of $\hat{\beta}$

- We can estimate $\hat{\beta}$ simply by sampling from $\beta|y$, and taking the expectation $E[\beta|y]$

```
mu0 <- 1.2; tau0 <- 1.3; sigma <- 1.5; N <- 10
set.seed(1004)
y <- rnorm(mean=2.3,sd=sqrt(sigma),n=N)

tau_star <- function(n,tau0,sigma){
     return(1/((1/tau0^2)+(n/sigma^2)))}

mu_star <- function(y,n,mu0,tau0,sigma,tau_star){
     return(tau_star*((mu0/tau0^2)+
       (n*mean(y)/sigma^2)))}

ts <- tau_star(n=N,tau0,sigma)
ms <- mu_star(y,n=N,mu0,tau0,sigma,tau_star=ts)
bhat <- mean(rnorm(mean=ms,sd=sqrt(ts),n=1000))
```

Conjugacy

- Recall that we said:

$$\begin{aligned} y|\beta &\sim N(\beta, \sigma^2) \\ \beta &\sim N(\mu_0, \tau_0^2) \end{aligned}$$

  - whose product is the posterior:

$$\beta|y \sim N(\mu_*, \tau_*^2)$$

  - so the evidence, prior and posterior are all Normal

- This feature is called *conjugacy*: all model components are from the same class of probabability distributions
  - Analytic and computational gains from considering conjugate models
  - Sampling techniques and powerful computers lessen the need for conjugacy

Diffuse Priors

- Consider the Normal model with a vague (uniform within a deviation) prior about β:

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Diffuse Priors

- Consider the Normal model with a vague (uniform within a deviation) prior about β:

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

- The joint posterior distribution is proportional to the product of the likelihood and $1/\sigma^2$:

$$\begin{aligned} p(\beta, \sigma^2 | y) &\propto \sigma^{-n-2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta)^2 \right\} \\ &\propto \sigma^{-n-2} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (y_i - \bar{y}) + n(\bar{y} - \beta) \right] \right\} \\ &\propto \sigma^{-n-2} \exp\left\{ -\frac{1}{2\sigma^2} \left[(n-1)s^2 + n(\bar{y} - \beta) \right] \right\} \end{aligned}$$

Diffuse Priors

- Consider the Normal model with a vague (uniform within a deviation) prior about β:

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

- The joint posterior distribution is proportional to the product of the likelihood and $1/\sigma^2$:

$$\begin{aligned}
p(\beta, \sigma^2|y) &\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2\right\} \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})+n(\bar{y}-\beta)\right]\right\} \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\bar{y}-\beta)\right]\right\}
\end{aligned}$$

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

Diffuse Priors

- To sample, we can factor the joint posterior distribution $p(\beta, \sigma^2|y)$ into its component conditional $p(\beta|\sigma^2, y)$ and marginal $p(\sigma^2|y)$ posterior distributions

Diffuse Priors

- To sample, we can factor the joint posterior distribution $p(\beta, \sigma^2|y)$ into its component conditional $p(\beta|\sigma^2, y)$ and marginal $p(\sigma^2|y)$ posterior distributions
- We can show the conditional posterior is $\beta|\sigma^2, y \sim N(\bar{y}, \frac{\sigma^2}{n})$
- Summing over β in the joint posterior, we can find the marginal posterior $\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2)$

Diffuse Priors

- To sample, we can factor the joint posterior distribution $p(\beta, \sigma^2|y)$ into its component conditional $p(\beta|\sigma^2, y)$ and marginal $p(\sigma^2|y)$ posterior distributions
- We can show the conditional posterior is $\beta|\sigma^2, y \sim N(\bar{y}, \frac{\sigma^2}{n})$
- Summing over β in the joint posterior, we can find the marginal posterior $\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2)$
- Thus an alternative form of $p(\beta, \sigma^2|y)$ is:

$$N(\bar{y}, \frac{\sigma^2}{n}) \times \text{Inv-}\chi^2(n-1, s^2)$$

- Suggests we can iteratively sample σ^2 from the marginal posterior $\sigma^2|y$, then using this σ^2 draw, can sample from $\beta|\sigma^2, y$ to obtain a posterior estimates for β and σ^2

Estimation of $\hat{\beta}$ with Diffuse Priors

- The above suggests we sample σ_t^2 from $p(\sigma^2|y)$, and then sample from $p(\beta|\sigma_t^2, y)$ using this sample value of σ_t^2 to find β_t

```
sigma <- 1.5; N <- 10
set.seed(1004)
y <- rnorm(mean=2.3,sd=sqrt(sigma),n=N)
s2 <- 1/(N-1)*sum((y-mean(y))^2)

invchi=((N-1)*s2)/(rchisq(1000,df=N-1))

bhat <- mean(rnorm(mean=mean(y),
      sd=sqrt(invchi/N),n=1000))
```

- We can sample independently since σ^2 is independent of β in the joint posterior
- (Also, note how the Inv-χ^2 is nicely sampled from the χ^2 distribution)

Prediction on $\tilde{y}$

- We can make predictions on future observations of y using posterior inference
- We make use of the posterior predictive distribution:

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)\mathrm{d}\theta \\
&= \int p(\tilde{y}|\theta)p(\theta|y)\mathrm{d}\theta
\end{aligned}
$$

- For our above example:

$$p(\tilde{y}|y) \propto \int \exp\left\{-\frac{1}{2\sigma^2}(\tilde{y}-\beta)^2\right\}\exp\left\{-\frac{1}{2\tau_*^2}(\beta-\mu_*)^2\right\}\mathrm{d}\theta$$

- Can show that $E[\tilde{y}|y] = \mu_*$ and $Var(\tilde{y}|y) = \sigma^2 + \tau_*^2$
- Useful for model checking with out-of-sample y