# Advanced Data Analysis in R

## Bayesian Modeling in R

Michael DeWitt

2018-03-17 (Updated 2019-04-15)

# Bayesian Modeling in R

# A Thought Exercise

You are already Bayesian!

You just didn't know it!

# A Coin

What is the probability a given coin is fair?

# Frequentist

If you didn't answer 100% or 0% you're Bayesian!

# Bayes Theorem

Just derrived from identities of probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Prior, Likelihood, Posterior Distribution

# Prior- What do we know (or not)?

Prior is *subjective* and represents a range of potential values

Specified by a distribution

**Informative**

Restricts range of likely values to small range

More informative adds more “weight” to the prior vs the data

**Noninformative/ Uninformative**

Wider range of possibilities

More closely approximates the Maximum likelihood estimates

# Likelihood - The Data Generating Process

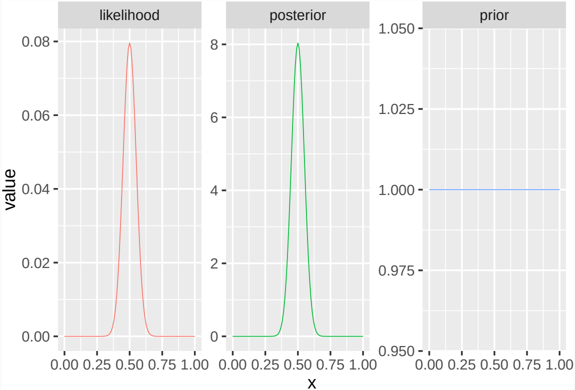
Likelihood is the distribution for the data generating process

Examples

- Poisson process -> poisson likelihood function
- Binomial process -> binomial likelihood function
- Normal distibution -> normal likelihood function
- Ordered categorical -> ordered categorical likelihood function

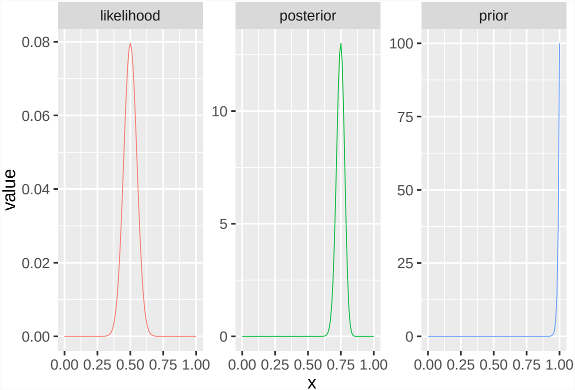
# Posterior - What we make inferences on!

$$P(A|B) \sim Prior * Likelihood$$



# Strong Priors

Say, I had a stronger prior. . .



# Bayesian Workflow

# Doing Bayesian Inferences

Bayesian inference and modeling techniques can be applied across the board

In MLE approachs you often make assumptions without even realising it!

Bayes requires you to be more deliberate

# Write Your Model

What is your data generating process?

*We are estimating the support for a given referendum. Thus our population has a choice, either 1 (support) or 0 (do not support). We do not have any good estimates from previous literature for overall support.*

# Write The DGP in Math

**Likelihood Function**

Series of bernouilli trials -> binomial likelihood function

$$P(y_i = 1|income, gender_i) \sim logit^{-1}(\mu_i)$$

Where,

$$\mu = normal(\beta_1 * income + \beta_2 * gender + intercept, \sigma)$$

**Prior**

Additionally say that we believe that the impact of these two metrics aren't too strong

$$\beta_1 \sim N(0, 0.5)$$

$$\beta_1 \sim N(0, 0.5)$$

# Simulate Your Data Generating Process

```r
n <- 1000
gender <-rep(x = 0:1, length.out = n)
income <- rnorm(n, 0, 1)
mu <- gender * 1.5 + income * 2
y <- rbinom(n,1, prob = plogis(mu))

dat <- data.frame(gender =gender,
                  income = income,
                  y = y)
```

# Domain Specific Languages

Implementation of Bayesian Data Analysis can be done manually...

But there exist domain specific languages to handle most cases:

- BUGS
- JAGS
- Stan
- Hand coded samplers

# Enter `brms`

`brms` makes Bayesian Modeling easy

Compiles traditional R and `lme4` syntax to Stan

Utilises Hamiltonian Monte Carlo with a No U-Turn Sampler

# Specifying a Model in `brms`

Models can specified in-line or separately using the `bf` function

```
library(brms)
(model <- bf(y ~ gender + income))
## y ~ gender + income
```

# Inspect Priors

The `get_prior` function allows the user to see what priors can be specified

```
get_prior(model, dat, bernoulli())
##                  prior     class   coef group resp dpar nlpar
## 1                              b
## 2                              b gender
## 3                              b income
## 4 student_t(3, 0, 10) Intercept
```

# Specify Priors

Priors can then be specified using the existing disribution families

```
my_priors <- c(
  prior(normal(0, 0.5), class = "b", coef = "gender"),
  prior(normal(0, 0.5), class = "b", coef = "income"))
```

# Model Family

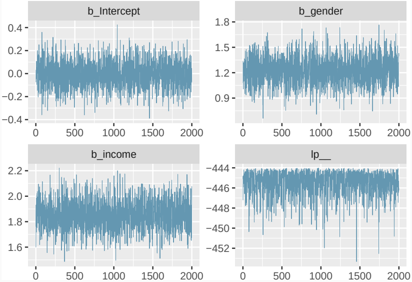
```
fit <- brm(model, my_priors,
           data =  dat,
           family = bernoulli(),
           inits = 1000, cores = 2,
           chains = 2, seed = 1234, refresh = 0)
```

# Posterior Checks

- Convergence
  - Trace Plots `tracplot`
  - Rhat metrics
  - Effective Sample Size
- Posterior Predictive Intervals
  - Was there a good fit between the model and the data

# So Let's Check

```
library(tidybayes)
library(bayesplot)
mcmc_trace(as.matrix(fit))
```



# Inferences

```
summary(fit)
##  Family: bernoulli
##   Links: mu = logit
## Formula: y ~ gender + income
##    Data: dat (Number of observations: 1000)
## Samples: 2 chains, each with iter = 2000; warmup = 1000; thin
##          total post-warmup samples = 2000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rha
## Intercept    -0.01      0.11    -0.22     0.23       2373 1.0
## gender        1.23      0.16     0.92     1.54       1787 1.0
## income        1.85      0.12     1.63     2.08       1595 1.0
##
## Samples were drawn using sampling(NUTS). For each parameter,
## is a crude measure of effective sample size, and Rhat is the
## scale reduction factor on split chains (at convergence, Rhat
```

# Advantages of Bayesian Analysis

- Takes advantage of expert opinion
  - Especially helpful with small samples size studies
  - Reduces possibility of wildly odd results
- Easier communications (more intuitive to discuss probabilities)
- Studies can build on one another
  - Results from one study can be supplied directly as a prior into a replciation or another study

# Drawbacks of Bayesian Inference

- Not as widely utilised in major publications
- Computationally intensive
- Picking a prior
- Heuristics exist

# References