STAT 578: Advanced Bayesian Modeling

Week 4 – Lesson 2

Normal Hierarchical Model in R/JAGS

Fall 2019

Prediction for 2016 Polls

Consider a hypothetical new national poll conducted just before the 2016 US presidential election.

- What would we expect its estimate for the Clinton lead to be?
- With what probability would it *clearly* indicate a Clinton lead (beyond its margin of error)?

Model for New Poll

Let

$$\tilde{y} \;=\; \text{Clinton lead (percentage points) in new poll}$$

To make $\tilde{y}$ comparable to the observed poll results $y_1, \ldots, y_7$, let

$$\tilde{y} \mid \tilde{\theta} \;\sim\; \mathrm{N}(\tilde{\theta}, \tilde{\sigma}^2)$$

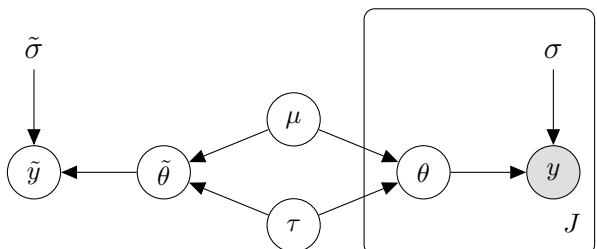
$$\tilde{\theta} \mid \mu, \tau \;\sim\; \mathrm{N}(\mu, \tau^2)$$

where $2\tilde{\sigma}$ is the new poll's margin of error, assumed known.

Note:

- $\tilde{\theta}$ is conditionally independent of $\theta_1, \ldots, \theta_7$ (given μ, τ) and has the same distribution as they do. Hence, it is exchangeable with them.
- The new poll is as if sampled from the same "population" of polls as the others.

DAG Model



We can extend the JAGS code (with the approximately flat hyperprior) to simulate $\tilde{y}$ (and $\tilde{\theta}$). In `polls20163.bug`:

```
model {

  for (j in 1:length(y)) {
    y[j] ~ dnorm(theta[j], 1/sigma[j]^2)
    theta[j] ~ dnorm(mu, 1/tau^2)
  }

  mu ~ dunif(-1000,1000)
  tau ~ dunif(0,1000)

  y.tilde ~ dnorm(theta.tilde, 1/sigma.tilde^2)
  theta.tilde ~ dnorm(mu, 1/tau^2)

  lead.ind <- y.tilde > 2*sigma.tilde

}
```

Note the line (deterministic relation)

```
lead.ind <- y.tilde > 2*sigma.tilde
```

which creates an **indicator variable**: It equals 1 when $\tilde{y} > 2\tilde{\sigma}$, and 0 otherwise.

The condition $\tilde{y} > 2\tilde{\sigma}$ means that the estimated Clinton lead exceeds zero by more than its margin of error.

For illustration, suppose the new poll has a margin of error of 2:

$$\tilde{\sigma} \;=\; 1$$

Now perform the analysis with R (`rjags`):

```
> m3 <- jags.model("polls20163.bug", c(as.list(d), sigma.tilde=1))
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 7
   Unobserved stochastic nodes: 11
   Total graph size: 49

Initializing model

  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
Warning messages:
1: In jags.model("polls20163.bug", c(as.list(d), sigma.tilde = 1)) :
  Unused variable "poll" in data
2: In jags.model("polls20163.bug", c(as.list(d), sigma.tilde = 1)) :
  Unused variable "ME" in data
```

```
> update(m3, 2500)  # burn-in
  |**************************************************| 100%

> x3 <- coda.samples(m3, c("y.tilde","lead.ind"), n.iter=10000)
  |**************************************************| 100%
```

```
> summary(x3)

Iterations = 3501:13500
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean     SD Naive SE Time-series SE
lead.ind 0.865 0.3417 0.003417       0.004308
y.tilde  3.721 1.8510 0.018510       0.024372

2. Quantiles for each variable:

            2.5%   25%   50%   75% 97.5%
lead.ind 0.00000 1.000 1.000 1.000 1.000
y.tilde  0.06936 2.717 3.735 4.779 7.368
```

Approximate 95% posterior predictive interval for $\tilde{y}$:

$$(0.07,\, 7.37)$$

Approximate posterior predictive probability that Clinton is clearly leading (by more than the margin of error) in the new poll:

$$\Pr(\tilde{y} > 2\tilde{\sigma} \mid y) \;\approx\; 0.87$$

Note: Mean of an indicator variable is probability it equals 1. (Why?)