Homework 1: STA465

Homework 1 is due on Wednesday, February 2nd by 23:59 EST. The homework assignment is worth 25 points in total.

Question 1: Posterior Distribution (5 pts)

Derive the closed-form expressions of the following two posterior distributions. These are common results so feel free to look through http://www.stat.columbia.edu/~gelman/book/BDA3.pdf for help.

Question 1.1

Suppose you have T observations, $\{y_i\}_{i=1}^T$, such that

$$y_i \sim N(\mu, \sigma = 3)$$

If we assign the following prior distribution for μ,

$$\mu \sim N(s, v^2)$$

what is the posterior distribution for μ? Show all steps and derivations.

Question 1.2

Suppose you have T observations, $\{y_i\}_{i=1}^T$, such that

$$y_i \sim Binomial(N, p)$$

for a fixed value of N. We would like to estimate the parameter p, and specify the following prior distribution,

$$p \sim Beta(\alpha, \beta)$$

where $Beta(\alpha, \beta)$ denotes the Beta distribution. Derive the posterior distribution for p. Show all steps and derivations.

Question 2: Simulation (5 pts)

Question 2.1

Simulate 20 data sets of size 50 from the following linear regression model,

$$y_i = \beta_0 + \beta_1 \cdot x + \epsilon_i$$

where $\epsilon_i \sim N(0, 5)$, $\beta_0 = 1$, $\beta_1 = -1$. Simulate values of x using the following R code:

```r
set.seed(465)
x <- rpois(n = 50, lambda = 5)
```

Make a plot for each simulation, with the true line overlaid.

Question 2.2

Simulate a data set from the following multilevel regression model,

$$y_{ij} = \beta_{0j} + \beta_{1j} \cdot x + \epsilon_{ij}$$
$$\beta_{0j} \sim N(-3, 1)$$
$$\beta_{1j} \sim N(1, 1)$$

for $\epsilon_i \sim N(0, 1)$, $j \in \{1, \ldots, 25\}$, and $i \in \{1, \ldots, 100\}$. Simulate values of x using the following R code:

```r
set.seed(465)
x <- rpois(n = 100, lambda = 5)
```

Make a plot for each group, with the true line overlaid.

Question 3: Prior Predictive Distributions (5 pts)

Given a prior distribution, the distribution of unknown but observable y is,

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

where $p(y)$ reflects the distribution of the observations and $p(\theta)$ denotes the prior distribution of the parameters in the model. This is called the *prior predictive distribution*.

To sample from this distribution, we can do so in two steps:

- first sample θ^j from the prior distribution
- using the value of θ^j, sample y^j from $p(y|\theta^j)$

Question 3.1: Linear Regression

Let's return to the simple linear regression model. In class, we simulated the following data set:

```r
#-----------------------------------------------
# Setting the values of the parameters
#-----------------------------------------------

beta0  <- 1
beta1 <- 0.5
sigma <- 1

#-----------------------------------------------
# Simulating covariate values + data
#-----------------------------------------------
```

```
set.seed(17)

x <- runif(n = 100, min = 1, max=5)
y.mean <- beta0 + beta1*x
y <- rnorm(n = 100,
           mean = y.mean,
           sd = sigma)

sim.data <- tibble(x,y, y.mean)
```

For the three following sets of prior distributions, generate 20 data sets from the respective *prior predictive distribution*. Prior distribution candidates:

- $\beta_0 \sim N(0,1), \beta_1 \sim N(0,1), \sigma \sim Gamma(shape = 1, scale = 1)$
- $\beta_0 \sim N(0,1000), \beta_1 \sim N(0,1000), \sigma \sim Gamma(shape = 1000, scale = 1000)$
- $\beta_0 \sim Unif(0,1), \beta_1 \sim Unif(-1,0), \sigma \sim Exp(1)$

For each sets of prior distributions, make two plots to relay the results:

- Graph the 20 prior predictive data sets and overlay the true line.
- Draw the curve of the prior distribution and add a vertical line at the true value of β_0, β_1, σ.

Write a short paragraph on the major implications of the choice of the prior distribution as seen via simulation from the prior predictive.

Question 3.2: Multilevel/Hierarchical Model

Let's extend the simple linear model to a multilevel model. We will use the following simulated data set for this assignment:

```
#-----------------------------------------------
# Setting the values of the parameters
#-----------------------------------------------

set.seed(17)

nu.mu <- 2
tau.mu <- 0.5
nu.beta <- -1
tau.beta <- 0.5
mu.hm <- rnorm(n=20, mean = nu.mu, sd=tau.mu)
beta.hm <- rnorm(n=20, mean = nu.beta, sd= tau.beta)
sigma <- 1

#-----------------------------------------------
# Simulating covariate values + data
#-----------------------------------------------

x.hm <- runif(n = 100, min = 1, max=5)
y.mean.hier <- c(rep(mu.hm, each = 100) +
```

```
                      rep(beta.hm, each = 100)*
                      rep(x.hm, 20))
y.hier <- rnorm(n = 20*100, mean = y.mean.hier, sigma)

sim.data.hier <- tibble(x = rep(x.hm, 20), y.hier,
                        y.mean.hier,
                        group = paste("Group",
                                      rep(1:20, each = 100)))
```

For the three following sets of prior distributions, generate a data set of the same size from the respective *prior predictive distribution*. Prior distribution candidates:

- $\nu_\mu \sim N(0,1), \nu_\beta \sim N(0,1), \tau_\mu \sim Gamma(shape = 1, scale = 1), \tau_\beta \sim Gamma(shape = 1, scale = 1), \sigma \sim Gamma(shape = 1, scale = 1)$
- $\nu_\mu \sim N(0,1000), \nu_\beta \sim N(0,1000), \tau_\mu \sim Gamma(shape = 1000, scale = 1000), \tau_\beta \sim Gamma(shape = 1, scale = 1000), \sigma \sim Gamma(shape = 1000, scale = 1000)$
- $\nu_\mu \sim Unif(0,1), \nu_\beta \sim Unif(0,1), \tau_\mu \sim Exp(1), \tau_\beta \sim Exp(1), \sigma \sim Exp(1)$

For each sets of prior distributions, make three plots to relay the results:

- Graph the prior predictive data sets and overlay the line from which the data was generated.
- Overlay the 20 original lines (individually, in separate panels) onto the 20 simulated lines
- Draw the curve of the prior distributions and add a vertical line at the true value of $\nu_\mu, \nu_\beta, \tau_\mu, \tau_\beta, \sigma$.

Write a short paragraph on the major implications of the choice of the prior distribution as seen via simulation from the prior predictive.

Question 4: Fitting Linear Models (5 pts)

Question 4.1: Fitting a Linear Regression Model

```
load("bayes-vis.RData")
latcab <- GM[GM$super_region == 5,]
```

Let's fit a linear regression model in INLA to the PM2.5 data set, using only the data from Latin America/Caribbean using two different sets of priors (your choice). Plot histograms of the posterior draws. Report the 95% credible intervals for each parameter estimate. Organize the results in a table.

Question 4.2: Fitting a Multilevel Regression Model

Let's fit a linear regression model in INLA to the PM2.5 data set using two different sets of priors (your choice). Plot histograms of the posterior draws. Report the 95% credible intervals for each parameter estimate.

Question 5: Posterior Predictive Distributions (5 pts)

The posterior predictive distribution for a new observation $\tilde{y}$, can be written as,

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

where $p(\theta|y)$ reflects the joint posterior distribution of the parameters.

To sample from the posterior predictive distribution, we can do so in two steps:

- first sample θ^j from the posterior distribution $p(\theta|y)$ in INLA
- using the value of θ^j, sample $\tilde{y}^j$ from $p(\tilde{y}|\theta^j)$

Question 5.1

Using the results from the two models in Question 4.1, generate 100 data sets from the posterior predictive distribution. Draw density curves comparing the simulated data sets to the original data set. Write 1-2 sentences per plot about how the simulated data sets compare to the original data set.

Question 5.2

Using the results from the two models in Question 4.2, generate 100 data sets from the posterior predictive distribution. Draw density curves comparing the simulated data sets to the original data set. Recall the hierarchical/multilevel structure of the model. Write 1-2 sentences per plot about how the simulated data sets compare to the original data set.