P8133 Week 8 Session 2: Binomial outcome and hierarchical model

0. Binomial outcome
   Suppose for each of N items $n \in 1:N$, we observe y_n successes out of S_n trials. Then our model will be

$$y_n \sim \text{Binomial}(S_n, \theta_n)$$

   where y_n is the number of response, S_n is the number of trials and θ_n is the response rate.
   To begin with, we will only focus on outcome for biomarker "EGFR".

   Create dataset

```
y = c(6, 1, 10, 0, 11, 1, 0, 1, 11, 0, 6, 0, 9, 11, 25, 1)
S = c(17, 7, 25, 1, 20, 3, 3, 1, 27, 3, 16, 0, 23, 14, 39, 4)
Drug = rep(c("Erlotinib", "Erlotinib+bexaorotane", "Vandetanib", "Sorafenib"), each = 4)
Marker = rep(c("EGFR", "KRAS/BRAF", "VEGF", "RXR/Cyclin D1"), times = 4)
mydata = data.frame(y, S, Drug, Marker)

mydata1 = mydata[which(mydata$Marker == "EGFR"),]
```

   Can you use uniform prior distribution and derive posterior distribution?

1. Bayesian model with binomial outcome

   (a). We will first run the stan model as above

```
stan_data = list(N = 4, y = mydata1$y, S = mydata1$S)
stan_fit1a = stan(file = "1aBinomial.stan", data = stan_data,
warmup = 500, iter = 1000, chains = 4, seed = 1)
print(stan_fit1a)
```

   (b). We can also use a logistic regression model with predictors and an intercept as follows.

$$logit(\theta_n) = X\beta$$

   where θ_n is the response rate. (Note: X is our design matrix and here is an intercept + one categorical predictor)

```
stan_data = list(N = 4, X = model.matrix(~Drug, data = mydata1), y = mydata1$y,
S = mydata1$S, D = 4)
stan_fit1b = stan(file = "1bLogisticRegression.stan", data = stan_data,
warmup = 500, iter = 1000, chains = 4, seed = 1)
print(stan_fit1b)
```

   Please extract posterior samples from logistic models and transform them to θ_n, and compare with part (a).

2. Binomial outcome with different prior
Next, let's assume we have some prior knowledge on the drug combination Erlotinib+bexaorotane and we are quite sure that this combination is much better than singlely using Erlotinib in terms of response on biomarker EGFR. Our model can be formulated as:

$$y_n \sim \text{Binomial}(S_n, \theta_n)$$
$$logit(\theta_n) = X\beta$$
$$\beta_2 \sim N(1, 1)$$

Run stan model

```
stan_data = list(N = 4, X = model.matrix(~Drug, data = mydata1),
y = mydata1$y, S = mydata1$S, D = 4)
stan_fit2 = stan(file = "2LogisticRegression.stan", data = stan_data,
warmup = 500, iter = 1000, chains = 4, seed = 1)
print(stan_fit2)
```

What is the posterior distribution of β_2 compared that in Q1 part (b)?

3. Hierarchical Bayesian model

(a). Hierarchical binomial model
First, we will use data from 1 drugs and 4 biomarkers and try to build a simple bayes hierarchical model. Assume now, we will only focus on drug "Erlotinib+bexaorotane".

The model will be

$$y_n \sim \text{Binomial}(S_n, \theta_n)$$
$$\theta_n \sim \text{beta}(a, b)$$
$$a \sim \exp(1)\text{I}_{\{a>1\}}$$
$$b \sim \exp(1)\text{I}_{\{b>1\}}$$

where y_n is the number of response, S_n is the number of trials and θ_n is the response rate.

Run stan model

```
mydata2 = mydata[which(mydata$Drug == "Erlotinib+bexaorotane"),]
stan_data = list(N = 4, y = mydata2$y, S = mydata2$S)
stan_fit3a = stan(file = "3aBinomialHierarchical.stan", data = stan_data,
warmup = 500, iter = 1000, chains = 4, seed = 1)
print(stan_fit3a)
```

What's the posterior distribution of θ_n,a and b?

(b). Hierarchical logistic regression
Now, we will use all data from 4 drugs and 4 biomarkers. Suppose each binomial outcome y_n has an associated level, $m_n \in \{1, ..., M\}$. Each outcome will also have an associated predictor vector $x_n \in R^D$. Each level m gets its own coefficient vector $\beta_m \in R^D$.

Mathematically, the hierarchical model places a prior on the coefficients $\beta_{m,d} \in R$, which is also estimated with the data. In this case, we will assume a normal distribution as the prior for β. Our model can be formulated as:

$$y_n \sim \text{Binomial}(S_n, \theta_n)$$
$$logit(\theta_n) = X\beta_m$$
$$\beta_{1,d}...\beta_{m,d} \sim N(\mu_d, \sigma_d^2)$$
$$\mu_d \sim N(0, 10^4)$$

Run stan model

```
stan_data = list(N = 16, X = model.matrix(~Drug, data = mydata), y = mydata$y,
S = mydata$S, D = 4, m = rep(seq(1:4), times = 4), M = 4)
stan_fit3b = stan(file = "3bLogisticHierarchical.stan", data = stan_data,
warmup = 500, iter = 1000, chains = 4, seed = 1)
print(stan_fit3b)
```

What is the posterior distribution of β_2? What is the advantage or disadvantage of hierarchical logistic model?