Assignment 6.b for **STATGR6103**

submitted to Professor Andrew Gelman

Advait Rajagopal

17 October 2016

1 Question 1

In the example of binary outcomes on page 147, it is assumed that the number of measurements, n, is fixed in advance, and so the hypothetical replications under the binomial model are performed with $n = 20$. Suppose instead that the protocol for measurement is to stop once 13 zeros have appeared.

1.1 Part A

Explain why the posterior distribution of the parameter θ under the assumed model does not change.

In the initial setting the binary outcomes have a binomial likelihood, and the probability parameter θ has a beta posterior distribution. So for n trials with y successes occurring with probability θ, the likelihood is as follows;

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

With a uniform prior on θ the posterior distribution is as follows,

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}$$

Under the new measurement protocol I stop when 13 zeros have appeared. This is a negative binomial distribution, where I stop generating new numbers as soon as 13 zeros have appeared. The functional form of the negative binomial is;

$$p(y|\theta) = \binom{n-1}{y-1}\theta^y(1-\theta)^{n-y}$$

With this likelihood the posterior distribution of θ is;

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}$$

which is the same as it was under the previous measurement protocol. The only change is in the constant which drops out under proportionality. This is why the posterior distribution of the parameter θ under the new model does not change.

1.2 Part B

Perform a posterior predictive check, using the same test quantity, T = number of switches, but simulating the replications y^{rep} under the new measurement protocol. Display the predictive simulations, $T(y^{rep})$, and discuss how they differ from Figure 6.5.

I draw θ from its Beta(8,14) posterior distribution and use it to simulate replications under the new measurement protocol. I also create a "switch" function in R to count the number of switches in a specific replication. I plot the histogram of switches in the replicated data alongside Figure 6.5.

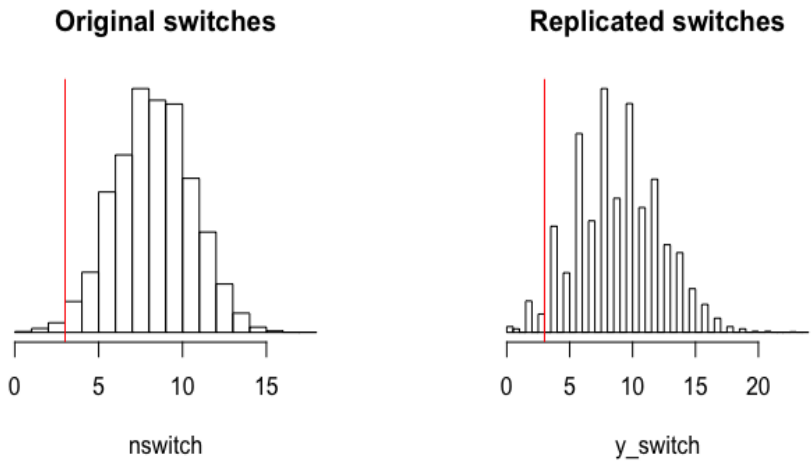

Figure 1: (a)Figure 6.5 from Gelman's BDA 3(vertical line at T(y) = 3); (b) The switches in the replicated data (vertical line at T(y^{rep}) = 3).

It is immediately obvious that there are irregular peaks in the replicated data. Also because n is not fixed, the number of switches can be more than 20. The distribution for replicated switches has much longer tails. It is also visible from Figure 1(b) that the peaks are alternating.

2 Question 2

In `http://www.stat.columbia.edu/ gelman/research/published/jbes01m045r3.pdf` we estimate the effects of survey incentives under different conditions. We had difficulty because the coefficient estimates were so noisy: See table 2 of that table, which shows estimates and half-widths of 50% intervals. The estimates were so noisy that we could not include all the three-way interactions, even though we wanted to do so. And our estimates of main effects and two-way interactions are also uncomfortably noisy. For this assignment, you should use informative priors on regression coefficients to do a better job. These are not raw data on individual survey responses. Rather, each line of the file represents a different experimental condition. The file has 101 lines of data, which correspond to between 2 and 5 experimental conditions in each of 39 survey experiments.
This assignment has two parts.

2.1 Part A

For the first part, download the data, graph them, and fit some simple regression models in Stan. Display the data and the fitted model. You do not need to make the same sorts of graphs that are in the published paper, but you should make some graphs that show the data and fitted model together.
I downloaded the data and made some graphs. These are shown in Figure 2. The data that we are given contains information from 39 surveys. It contains the following information.

1. The response rate for the survey
2. The dollar value of the incentive (value)
3. The mode of the survey (telephone or face to face)
4. The type of incentive (gift or cash)
5. The timing of the incentive (before or after survey)
6. The burden of the survey (low or high)

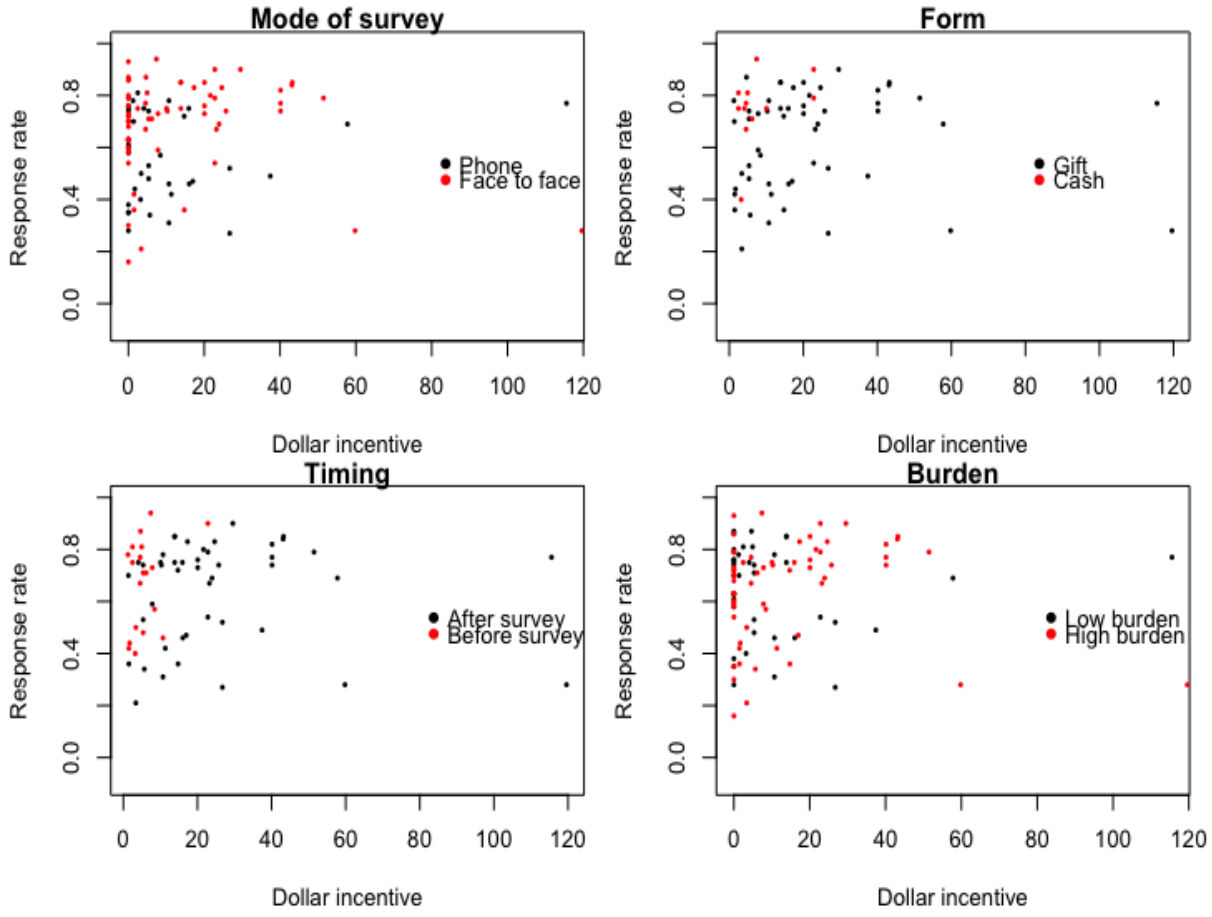

Figure 2: The data are quite noisy and no clear relationships can be observed.

Having looked at Singer et al.(1999) and Gelman et. al(2003) it is clear to me that the real quantities of interest are the value of the incentive and the response rate and I believe the goal is to model the relationship between these two variables with additional information provided by the other factors (time, mode etc.). I start by running a simple linear regression on all the variables with the value of the incentive and the other variables as predictors and the response rate as the dependent variable. However I do not use the raw response rate and value, but the difference between the response rate in the treatment and control condition with no incentive. The benefit of using the normal linear model is that we have easily interpretable coefficients.

The linear models I run however regard all surveys as interchangeable which is not entirely true and correspond to a case of complete pooling. A summary of the model and the posterior estimates of the parameters is given in Table 1. Figure 3 has data overlaid with the fitted model.

Table 1: Posterior Means (standard deviation) of Coefficients

	Model 1	Model 2
Intercept	0.03 (0.01)	0.03 (0.01)
Value	0.001 (0.0004)	0.0008 (0.0015)
Mode	-0.02 (0.02)	-0.05 (0.03)
Timing	0.04 (0.02)	0.01 (0.03)
Form	-0.05 (0.02)	-0.03 (0.03)
Burden	0.04 (0.02)	-0.01 (0.03)
Value × Mode		0.01 (0.002)
Value × Timing		0.01 (0.002)
Value × Form		-0.003(0.003)
Value × Burden		0.003(0.003)

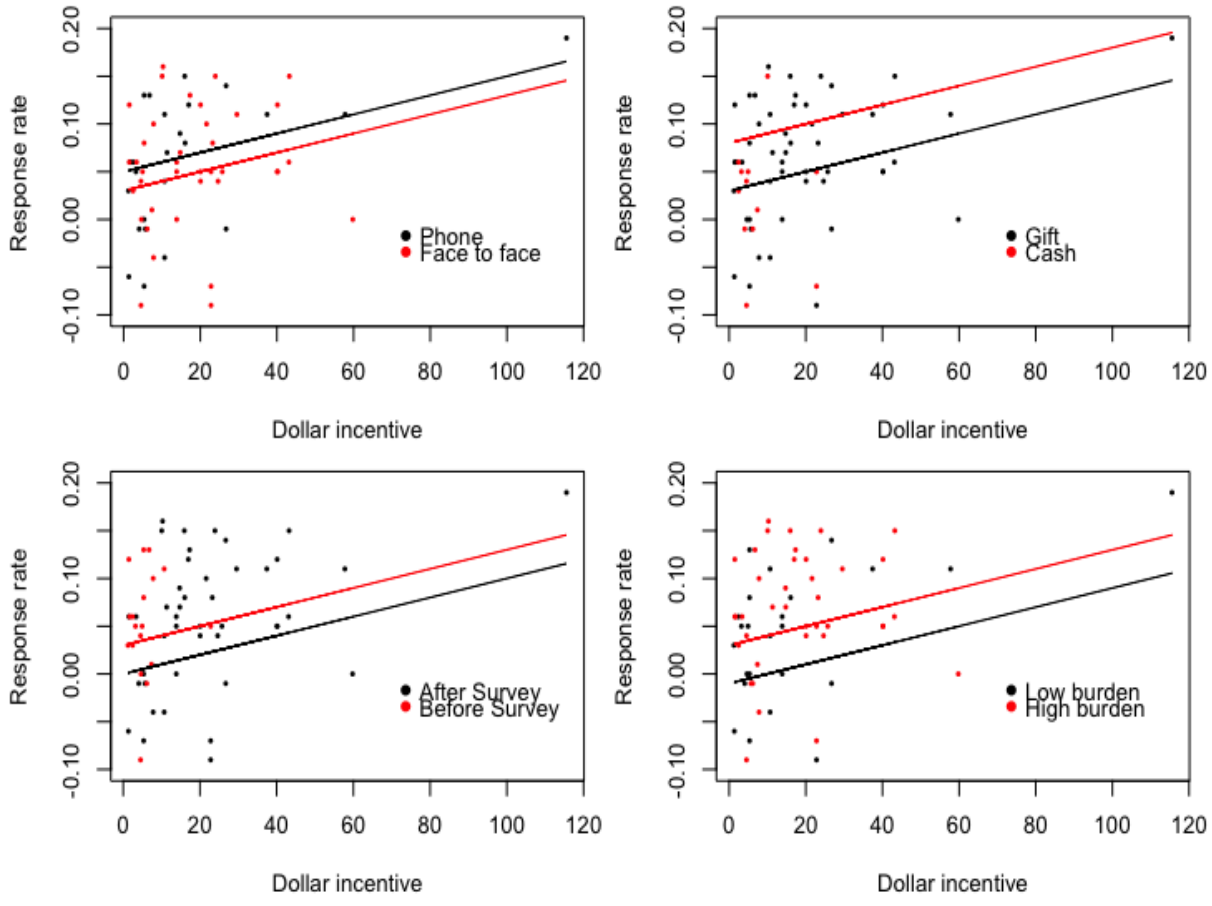

Figure 3: Shows the linear regression lines from the fitted model for each case.

It is clear that the data is quite noisy and the coefficients of regression do not have real meaning if they are calculated for the pooled data as survey level variation is to be accounted for. A Bayesian multilevel model is in order and we need informative priors on the coefficients to get meaningful results and understand how incentive affects response rates.

3 Code

3.1 R Code

Listing 1: R Code

```
#####
# Question 1
#####
data <- c(1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0)
data <- table(data)
class(data)
data[names(data==0)]

##Switches function
switches <- function(x)
{ switches <- sign(x)
sum(switches[-1] != switches[-length(x)])
}
switches(data)
#Replicating with original protocol
nswitch <- NULL
for ( i in 1:10000){
  theta <- rbeta(10000, 8,14)
  ygen <- rbinom(20, 1, mean(theta))
  nswitch[i] <- switches(ygen)
}
#Simulating y_rep with new protocol
y_rep_gen<-function(){
  theta <- rbeta (10000 ,8 ,14)
  y_rep_each <- rbinom (1,1,mean(theta))
  while (sum(y_rep_each==0) < 13)
    y_rep_each<- c(y_rep_each, rbinom(1,1,mean(theta)))
  return(y_rep_each)
}
y_rep_all <- NULL
for(i in 1:10000){
  y_rep_all[[i]]<-y_rep_gen() }
y_switch<-c()
for(i in 1:10000){
  y_switch[i]<-length(rle(y_rep_all[[i]])$values)-1
}
```

Listing 2: R Code contd.

```
#Plotting histograms
par(mfcol = c(1,2))
hist(nswitch, main="Original␣switches",
     yaxt = "n", ylab=NULL, breaks = 20)
abline(v = 3, col = "red")
hist(y_switch,main="Replicated␣switches",
     yaxt = "n", ylab=NULL, breaks = 50)
abline(v = 3, col = "red")

#####
# Question 2
#####
# #Read in data
rm(list = ls())
setwd("/Users/Advait/Desktop/New␣School/Fall16/BDA/Class11")
incentive <- read.table("http://www.stat.columbia.edu/
~gelman/bda.course/incentives_data_clean.txt",
                                    skip = 12)
str(incentive)
incentive$id <- c(as.factor(incentive$sid))
id <- incentive$id
unique(sid)
length(unique(sid))
#Initial plots
par(mfcol = c(2,2),  mar = c(4,4,1,1))
plot(incentive$v[incentive$m==-0.5],
     incentive$r[incentive$m==-0.5], pch = 16, cex = 0.5,
     xlab = "Dollar␣incentive",
     ylab = "Response␣rate", main = "Mode␣of␣survey",
     ylim = c(-0.1,1))
points(incentive$v[incentive$m==0.5],
       incentive$r[incentive$m==0.5], pch = 16, cex = 0.5,
       col = "red")
legend(80,0.6, legend =  c("Phone", "Face␣to␣face"),
       col = c("black", "red"), pch = 16, bty = "n")
##
plot(incentive$v[incentive$t==-0.5],
     incentive$r[incentive$t==-0.5], pch = 16, cex = 0.5,
     xlab = "Dollar␣incentive",
     ylab = "Response␣rate", main = "Timing",
     ylim = c(-0.1,1))
```

Listing 3: R Code contd.

```
points(incentive$v[incentive$t==0.5],
       incentive$r[incentive$t==0.5], pch = 16, cex = 0.5,
       col = "red")
legend(80,0.6, legend =  c("After␣survey", "Before␣survey"),
       col = c("black", "red"), pch = 16, bty = "n")
##
plot(incentive$v[incentive$f==-0.5],
     incentive$r[incentive$f==-0.5], pch = 16, cex = 0.5,
     xlab = "Dollar␣incentive",
     ylab = "Response␣rate", main = "Form",
     ylim = c(-0.1,1))
points(incentive$v[incentive$f==0.5],
       incentive$r[incentive$f==0.5], pch = 16, cex = 0.5,
       col = "red")
legend(80,0.6, legend =  c("Gift", "Cash"),
       col = c("black", "red"), pch = 16, bty = "n")
##
plot(incentive$v[incentive$b==-0.5],
     incentive$r[incentive$b==-0.5], pch = 16, cex = 0.5,
     xlab = "Dollar␣incentive",
     ylab = "Response␣rate", main = "Burden",
     ylim = c(-0.1,1))
points(incentive$v[incentive$b==0.5],
       incentive$r[incentive$b==0.5], pch = 16, cex = 0.5,
       col = "red")
legend(80,0.6, legend =  c("Low␣burden", "High␣burden"),
       col = c("black", "red"), pch = 16, bty = "n")
#####################################################################
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
incentiveprime <- incentive[complete.cases(incentive),]
dim(incentiveprime)

rdif <- incentiveprime$r.dif
vdif <- incentiveprime$v.dif
time <- incentiveprime$t
f <- incentiveprime$f
b <- incentiveprime$b
m <- incentiveprime$m
```

Listing 4: R Code contd.

```
stanc("6b.stan")
fit <- stan("6b.stan", data = list("rdif", "vdif","m",
"time", "f", "b"),
            iter = 1000, chains = 3)
print(fit)
ext1 <- extract(fit)
sd(ext1$b1)
###
stanc("6b2.stan")
fit2 <- stan("6b2.stan", data = list("rdif", "vdif","m",
"time", "f", "b"),
             iter = 1000, chains = 3)
print(fit2)
ext2 <- extract(fit2)
mean(ext2$b1)
sd(ext2$b1)
mean(ext2$b9)
############################
par(mfcol = c(2,2), mar = c(4,4,1,1))
##Mode
plot(incentiveprime$v.dif[incentiveprime$m==-0.5],
     incentiveprime$r.dif[incentiveprime$m==-0.5], pch = 16,
     cex = 0.5,
     xlab = "Dollar incentive",
     ylab = "Response rate",
     ylim = c(-0.1,0.2))
points(incentiveprime$v.dif[incentiveprime$m==0.5],
       incentiveprime$r.dif[incentiveprime$m==0.5], pch = 16,
       cex = 0.5,
       col = "red")
lines(incentiveprime$v.dif, 0.05 + 0.001*incentiveprime$v.dif)
lines(incentiveprime$v.dif, 0.03 + 0.001*incentiveprime$v.dif,
col = "red")
legend(70,0, legend =  c("Phone", "Face to face"),
       col = c("black", "red"), pch = 16, bty = "n")
```

Listing 5: R Code contd.

```
##Time
plot(incentiveprime$v.dif[incentiveprime$t==-0.5],
     incentiveprime$r.dif[incentiveprime$t==-0.5], pch = 16,
     cex = 0.5,
     xlab = "Dollar incentive",
     ylab = "Response rate",
     ylim = c(-0.1,0.2))
points(incentiveprime$v.dif[incentiveprime$t==0.5],
       incentiveprime$r.dif[incentiveprime$t==0.5], pch = 16,
       cex = 0.5,
       col = "red")
lines(incentiveprime$v.dif, 0 + 0.001*incentiveprime$v.dif)
lines(incentiveprime$v.dif, 0.03 + 0.001*incentiveprime$v.dif,
col = "red")
legend(70,0, legend =  c("After Survey", "Before Survey"),
       col = c("black", "red"), pch = 16, bty = "n")
##Form
plot(incentiveprime$v.dif[incentiveprime$f==-0.5],
     incentiveprime$r.dif[incentiveprime$f==-0.5], pch = 16,
     cex = 0.5,
     xlab = "Dollar incentive",
     ylab = "Response rate",
     ylim = c(-0.1,0.2))
points(incentiveprime$v.dif[incentiveprime$f==0.5],
       incentiveprime$r.dif[incentiveprime$f==0.5], pch = 16,
       cex = 0.5, col = "red")
lines(incentiveprime$v.dif, 0.03 + 0.001*incentiveprime$v.dif)
lines(incentiveprime$v.dif, 0.08 + 0.001*incentiveprime$v.dif,
col = "red")
legend(70,0, legend =  c("Gift", "Cash"),
       col = c("black", "red"), pch = 16, bty = "n")
##Burden
plot(incentiveprime$v.dif[incentiveprime$b==-0.5],
     incentiveprime$r.dif[incentiveprime$b==-0.5], pch = 16,
     cex = 0.5,
     xlab = "Dollar incentive",
     ylab = "Response rate",
     ylim = c(-0.1,0.2))
points(incentiveprime$v.dif[incentiveprime$b==0.5],
       incentiveprime$r.dif[incentiveprime$b==0.5], pch = 16,
       cex = 0.5,
       col = "red")
```

3.2 Stan Code

Listing 6: Stan Code

```
data{
  #int id[101];
  real rdif[62];
  real vdif[62];
  real m[62];
  real time[62];
  real f[62];
  real b[62];
  }
parameters{
  real b0;
  real b1;
  real b2;
  real b3;
  real b4;
  real b5;
  real <lower = 0> sigma;
}
model{
for(i in 1:62){
  rdif[i] ~ normal(b0 + b1*vdif[i] + b2*m[i] + b3*time[i] +
  b4*f[i] + b5*b[i], sigma);
    }
}
```

Listing 7: Stan Code with interaction

```
data{
  #int id[101];
  real rdif[62];
  real vdif[62];
  real m[62];
  real time[62];
  real f[62];
  real b[62];
  }
parameters{
  real b0;
  real b1;
  real b2;
  real b3;
  real b4;
  real b5;
  real b6;
  real b7;
  real b8;
  real b9;
  real <lower = 0> sigma;
}
model{
for(i in 1:62){
  rdif[i] ~ normal(b0 + b1*vdif[i] + b2*m[i] + b3*time[i] + b4*f[i]
  + b5*b[i] + b6*(vdif[i]*m[i]) + b7*(vdif[i]*b[i]) +
  b8*(vdif[i]*f[i])
  + b9*(vdif[i]*time[i]), sigma);
    }
}
```