Assignment 2.b for **STATGR6103**

submitted to Professor Andrew Gelman

Advait Rajagopal

18 September 2016

1 Question 1

1.1 Is there evidence from the data that the effect is not constant across frequencies? Explain, and justify your answer quantitatively.

A customary look at the data seems to suggest that there may be an effect of the electromagnetic field on the brains and that this may vary across frequencies. However this needs a lot more quantitative justification.

I start by understanding the data. Intuitively I expect that sham treatment should have no real effect and thus the effects should be centered around one because the recorded value is a ratio. Let us call the treatment group 'A1', the control group associated with it 'A2', the sham treatment group 'B1' and the control group associated with it 'B2'[1]. Thus in this terminology we expect 'B1/B2' to be one. If it is not then the 'control' itself has some effect on the chicken brains. However if the electromagnetic field has an impact on the brains and moreover this effect varies across frequencies, we expect to see a pattern in the ratio of 'A1/A2' and frequencies.

I verify in Figure 1 that the ratio 'B1/B2' is approximately one.

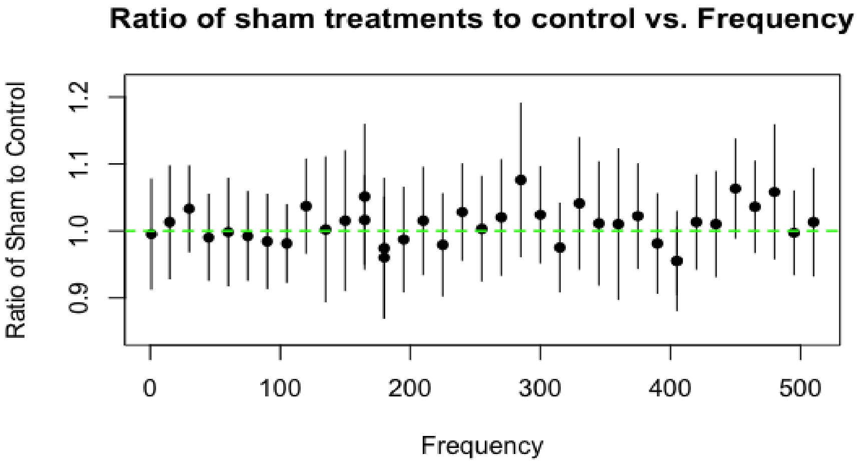

Figure 1: Ratio of sham treatments to control is centered around one

[1] http://www.stat.columbia.edu/ gelman/research/published/ChanceEthics1.pdf

Now I plot the relationship between the ratio of treatment over control and frequency to find out if there is any effect and most importantly an effect that is *variant across frequencies*. This is shown in Figure 2.

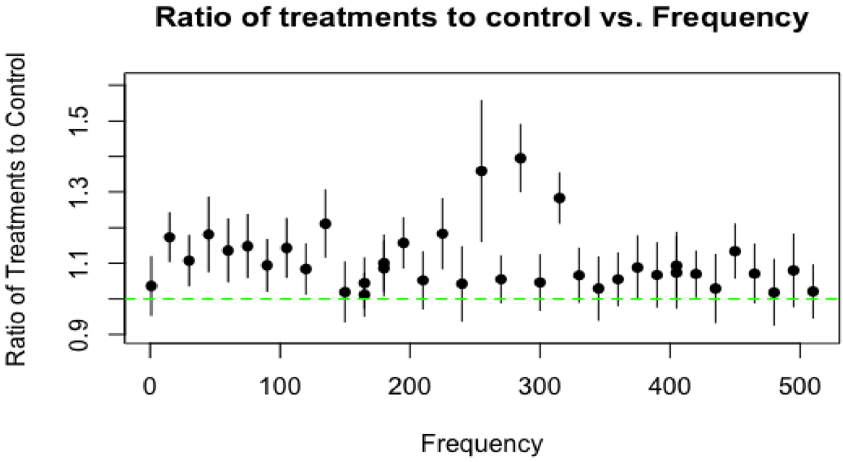

Figure 2: Effect of exposure to the field is positive

Upon looking at Figure 2 it becomes clear that there is some positive effect of exposing the brains to electromagnetic waves. There seem to be some spikes in effect particularly between 200 and 300 Hz and this might suggest that there is some *significant* effect there. However this could be attributed to pure noise or some sampling variability and it is our responsibility to check this using multilevel or hierarchical Bayesian methods. I thus perform a multilevel regression of treatments effects across frequencies and the recorded data points are called y_j and these are unbiased estimates of the true treatment effect θ_j. Thus we assume $y_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$. We further believe that $\theta_j \sim \mathrm{N}(\mu_\theta, \tau_\theta^2)$. Since we have data for the σ_j^2 we believe it is equal to the standard error of y_j and we attempt to make inferences about the hyperparameters that govern the distribution of θ_j, namely μ_θ and τ_θ^2 [2].

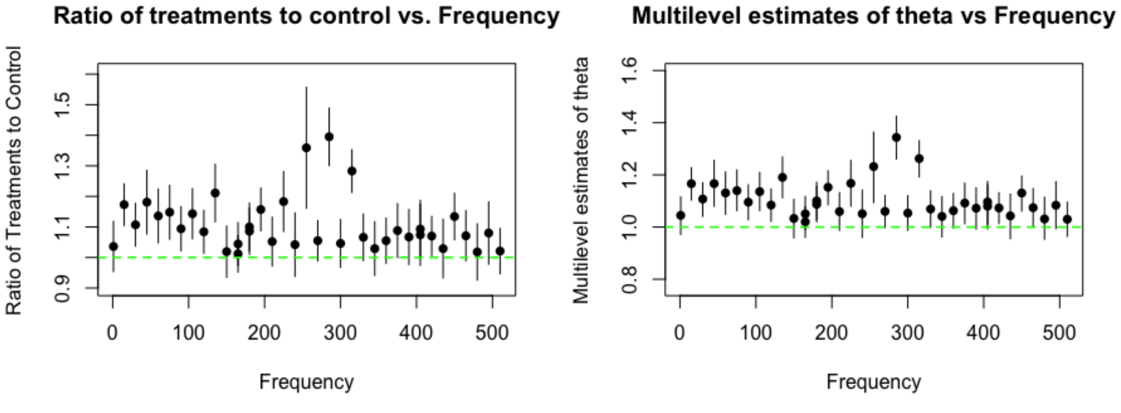

Figure 3: (a)Figure 3a shows the original raw estimates of treatment effects as a ratio of treatments to control. (b)Figure 3b shows the estimates of the same ratio based on a multilevel Bayesian estimation.

[2] For a clear exposition of the model and justification of the assumption of normality, see Gelman et al.(2008) at [http://www.stat.columbia.edu/ gelman/research/unpublished/multiple2.pdf]

We must note that the estimates of θ in Figure 3b are obtained by partial pooling. After estimating $\tau = 0.07$ from the full hierarchical model, I use a value of $\tau = 0.05$ for the partial pooled model. Upon examining Figure 3 it becomes clear that the seeming variant effect that was present across frequencies from the just the raw estimate A1/A2 seems to diminish significantly across the same frequencies. The estimates bunch closer together and there is no real *significant* region of frequency as thought earlier[3]. Figure 3b reveals that there exists some variation across frequencies but to suggest there is a clear distinction between some significant and nonsignificant frequencies is not a valid inference. We must consider sampling variation or just pure noise as well while examining the treatment effects.

1.2 What is the role of the "sham" treatment? Why is it performed at all?

As explored in section 1.1 and Figure 1 the ratio B1/B2 is very close to one in most cases. Thus after the data collection process we find out that a sham treatment is almost analogous to control yielding the ratio B1/B2 = 1. However before we start the experiment we don't really know that 'control' is actually a control. There may be an effect of just placing the chicken brain in water. Thus we have groups A1, A2, B1 and B2 corresponding to exposure, control, sham treatment and sham control. If sham treatment is analogous to control we expect groups A2, B1 and B2 to be the same. But we cannot guarantee this a priori. So it is useful to perform the sham treatment to get a differential measure of the actual effects of exposure to electromagnetic waves.

1.3 Consider two different summaries of treatment effect: *(i)*Mean ratios for exposed treatments vs. controls (1.036 at 1 Hz, 1.173 at 15 Hz, etc.), or *(ii)*Mean ratio for exposed treatments vs. controls, divided by mean ratio for sham treatments vs. controls (1.036/0.995 at 1 Hz, 1.173/1.013 at 15 Hz, etc.). Which of these two is a better estimate of the treatment effects? Use the data to address this question.

I think that using the mean ratios for treatment versus control or A1/A2 is better than using $\frac{A1/A2}{B1/B2}$ and will provide a better estimate of treatment effects. I have looked at the data long and hard and feel that A1/A2 is a more honest measure of the actual treatment. There are several reasons for this.

- First of all a sham treatment corresponding to a certain 'frequency' seems meaningless as the machine that produces electromagnetic waves is turned off at this time. This means that dividing the first value A1/A2 by the corresponding B1/B2 value at 1 Hz seems unproductive and moreover an undesirable transformation of the data which actually captures treatment effects. Each A1/A2 observation appears paired with a B1/B2 observation and corresponds to a certain frequency. I don't think this taking a ratio of ratios will contribute to our goal of estimating treatment effects

- If somehow B1/B2 and A1/A2 were correlated in some way, then it would still make sense for us to take into account this correlation across frequencies and factor this into analysis. In that case the transformation of each data point (by division) would reflect only the true treatment effect.

[3] Refer footnote 1 and Blackman et.al (1988)

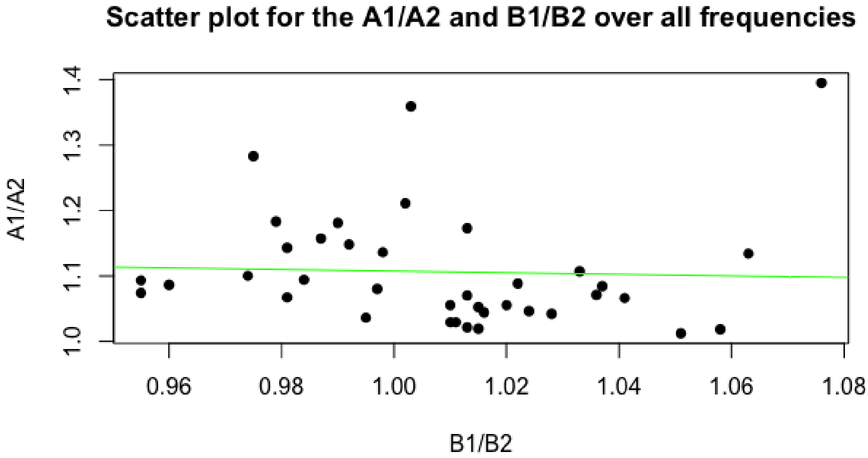

Figure 4: No correlation between exposure and sham effects

Figure 4 makes it very clear that there is no correlation between the ratio of treatment to control and the ratio of sham to control effects. This means that the two ratios are not related and thus dividing them by each other will not be a better measure of treatment effects. The better summary of treatment effects will be given by the measure A1/A2 alone.

2 Code

2.1 R Code

```
getwd()
setwd("/Users/Advait/Desktop/New␣School/Fall16/BDA/Week3")

##Reading in the data
df <- read.table("data_chickens", header=FALSE)
str(df)
colnames(df) <- c("frequency","N1","ratio_sham","se_sham",
                                        "N2","ratio_exp","se_exp")
str(df)

#Plotting the sham values for frequencies
plot(df$frequency, df$ratio_sham,
     ylim = range(c(df$ratio_sham - 2.5*df$se_sham,
     df$ratio_sham + 2.5*df$se_sham)),
     pch = 16, col = "black",
     xlab = "Frequency", ylab = "Ratio␣of␣Sham␣to␣Control",
     main = "Ratio␣of␣sham␣treatments␣to␣control␣vs.␣Frequency")
arrows(df$frequency,
          df$ratio_sham - 2*df$se_sham,
          df$frequency,
            df$ratio_sham+ 2*df$se_sham,
            length=0, angle=90, code=2, col = "black")
abline(h=1, lty = 2, col = "green", lwd = 1.5)

#Plotting the treatment values for frequencies
plot(df$frequency, df$ratio_exp,
     ylim = range(c(df$ratio_exp - 2.5*df$se_exp,
     df$ratio_exp + 2.5*df$se_exp)),
     pch = 16, col = "black",
     xlab = "Frequency", ylab = "Ratio␣of␣Treatments␣to␣Control",
     main = "Ratio␣of␣treatments␣to␣control␣vs.␣Frequency")
arrows(df$frequency,
            df$ratio_exp - 2*df$se_exp,
            df$frequency,
            df$ratio_exp + 2*df$se_exp,
            length=0, angle=90, code=2, col = "black")
abline(h=1, lty = 2, col = "green", lwd = 1.5)

##Fit a multilevel model for ratio of treatments to control vs frequency
#Fit 1 is a full hierarchical model with all parameters estimated within the model
N <- length(df$frequency)
y <- df$ratio_exp
#tau <- 0.05
sigma <- df$se_exp
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
stanc("hier_chickens.stan")
fit1 <- stan("hier_chickens.stan",
                  data = list("N", "y", "sigma"),
                  iter = 1000, chains = 3)
print(fit1)
ext1 <- extract(fit1)
thetaval <- colMeans(ext1$theta)

#Fit 2 is a Partial pooling model with tau = 0.05
fit2 <- stan("hier_chickens.stan",
                  data = list("N", "y", "sigma","tau"),
                  iter = 1000, chains = 3)
print(fit2)
ext2 <- extract(fit2)
thetaval2 <- colMeans(ext2$theta)
results <- print(fit2)
str(print(fit2))
```

```
##Plotting raw estimates and multilevel estimates
lowerint <- sapply(as.data.frame(fit2), FUN = quantile, probs = 0.025)
upperint <- sapply(as.data.frame(fit2), FUN = quantile, probs = 0.975)

par(mfcol = c(1,2))
plot(df$frequency, df$ratio_exp,
     ylim = range(c(df$ratio_exp - 2.5*df$se_exp,
     df$ratio_exp + 2.5*df$se_exp)),
     pch = 16, col = "black",
     xlab = "Frequency", ylab = "Ratio of Treatments to Control",
     main = "Ratio of treatments to control vs. Frequency")
arrows(df$frequency,
     df$ratio_exp - 2*df$se_exp,
     df$frequency,
     df$ratio_exp + 2*df$se_exp,
     length=0, angle=90, code=2, col = "black")
abline(h=1, lty = 2, col = "green", lwd = 1.5)

plot(df$frequency, thetaval2,
     ylim = range(c(thetaval2 - 2.5*tau, thetaval2 + 2.5*tau)),
     pch = 16, col = "black", xlab = "Frequency",
     ylab = "Multilevel estimates of theta",
     main = "Multilevel estimates of theta vs Frequency")
for (i in 1:38){
arrows(df$frequency[i], lowerint[i], df$frequency[i], upperint[i],
length=0, angle=90, code=2, col = "black")
}
abline(h=1, lty = 2, col = "green", lwd = 1.5)

##Plot for correlation between A1/A2 and B1/B2
plot(df$ratio_sham,df$ratio_exp,
       xlab = "B1/B2",
       ylab = "A1/A2",
     main = "Scatter plot for the A1/A2 and B1/B2 over all frequencies",
     pch = 16)
abline (lm(df$ratio_exp ~ df$ratio_sham), col = "green")
```

2.2 Stan code

```
data{
int N;
real y[N];
real<lower=0> sigma[N];
real tau; //Here tau is 0.05 and pulled from R
}
parameters{
real theta [N] ;
real mu;
#real<lower=0> tau; //Here tau is an unknown parameter and Stan estimates it
}
model{
theta ~ normal(mu, tau);
y ~ normal(theta , sigma);
}
```