Assignment 6.a for **STATGR6103**

submitted to Professor Andrew Gelman

Advait Rajagopal

12 October 2016

1 Question 1

Tesla is the world's leading manufacturer of long-range electric vehicles. In the process of creating these vehicles, Tesla has also built a vast network of charging stations, called "superchargers", that allow customers to charge their cars' batteries to 80% in just 40 minutes (as opposed to 4-8 hours with standard home/public outlets). If electric cars are going to be the future, this supercharging infrastructure is necessary for customers to be able to drive long distances without worrying about range anxiety or having to stop for inconvenient amounts of time. At the same time, however, building superchargers is costly, and the location of new superchargers needs to be optimized.
As such, Tesla needs your help to understand the relationship between supercharger remoteness and usage. The attached data set contains 100 superchargers in Tesla's network, and for each supercharger contains:

- remoteness: the distance of the supercharger from the nearest big city
- usage: the proportion of Tesla customers in the nearest big city who use this supercharger on a regular basis

1.1 Part A

Explore the data and create a model that describes the relationship between supercharger remoteness and usage. Explicitly write you prior, likelihood, and posterior.
I start by simply plotting the data. Observing an inverse trend in the relationship between usage and remoteness (i.e, usage decreases as remoteness increases), I have reason to believe that perhaps a logit transformation of the "usage" data might yield a more substantial linear trend. This is convenient also because the "usage" data is a proportion that lies between 0 and 1 and so doing a logit transform of this data will allow me to fit a normal regression eventually. I then plot the transformed "usage" data against the usage and confirm what I suspect is a downward sloping trend. Figure 1 shows this exploration process. Figure 1(a) shows the initial scatter plot of the raw usage data plotted against the remoteness of the superchargers. Figure 1(b) shows the scatterplot of the logit transformed data with a simple linear regression line that confirms the downward sloping trend.

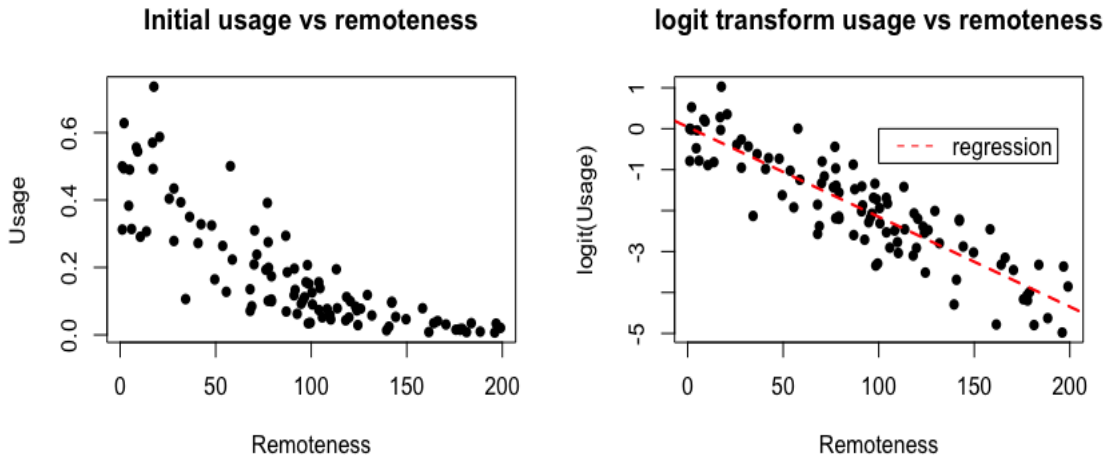

Figure 1: (a) Shows the raw usage data for the superchargers; (b) Shows the logit transformed usage data for the superchargers

Thus if we call the data for the i^{th} supercharger's usage, "u_i" then the logit transform is as follows;

$$z_i = logit(u_i) = log\Big(\frac{u_i}{1-u_i}\Big)$$

The linear regression line in Figure 1(b) is a simple linear regression and has a slope and intercept of -0.021 and 0.042 respectively.
Now I begin to create a model that fits the transformed data z_i. Consider the model;

$$z_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

I use a noninformative prior for the slope and intercept parameters called β_1 and β_0 respectively. I further assume errors are normally distributed and thus the error ϵ has a normal distribution as given below;

$$\epsilon \sim \mathrm{N}(0, \sigma^2)$$

The noninformative **prior distribution** on the parameters (assuming independence of the scale and location parameters) then becomes;

$$p(\beta_0, \beta_1, \sigma^2) \propto (\sigma^2)^{-1}$$

The likelihood function for z is as given below;

$$p(z_i|\beta_0, \beta_1, \sigma^2) \sim \mathrm{N}(\beta_0 + \beta_1 * x_i, \sigma^2)$$

This means that the **joint likelihood** is given by the following expression;

$$\begin{aligned} p(z|\beta_0, \beta_1, \sigma^2) &\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\Big(\frac{-(z_i - \beta_0 - \beta_1 * x - \epsilon_i)^2}{2\sigma^2}\Big) \\ &= \sigma^{-n} \exp\Big(\frac{\sum_{i=1}^{n} -(z_i - \beta_0 - \beta_1 * x - \epsilon_i)^2}{2\sigma^2}\Big) \\ &= \sigma^{-n} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - \beta_0 - \beta_1 * x - \epsilon_i)^2\Big) \end{aligned}$$

Therefore the **posterior distribution** is derived as below;

$$p(\beta_0, \beta_1, \sigma^2|z, x) \propto p(\beta_0, \beta_1, \sigma^2)p(z|\beta_0, \beta_1, \sigma^2)$$
$$= \sigma^{-n-2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - \beta_0 - \beta_1 * x - \epsilon_i)^2\Big)$$

1.2 Part B

Fit your model in Stan and use the posterior draws to simulate the usage of the 100 superchargers in the data set. Compare your simulations to the actual data.
I fit the model described in section 1.1 in Stan and I provide a summary of the estimates of the parameters of the posterior distribution in Table 1;

Table 1: Posterior parameter estimates

	mean	2.5%	25%	50%	75%	97.5%	R-hat
β_0	0.04	-0.21	-0.04	0.04	0.12	0.29	1
β_1	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	1
σ	0.44	0.39	0.42	0.44	0.46	0.51	1

We observe the desired negative slope denoted by β_1 and a very good convergence with R-hat = 1 for all parameters. Figure 2 shows an overview of the model with the original data and inverse logit transformed "predicted usage" values.[1]

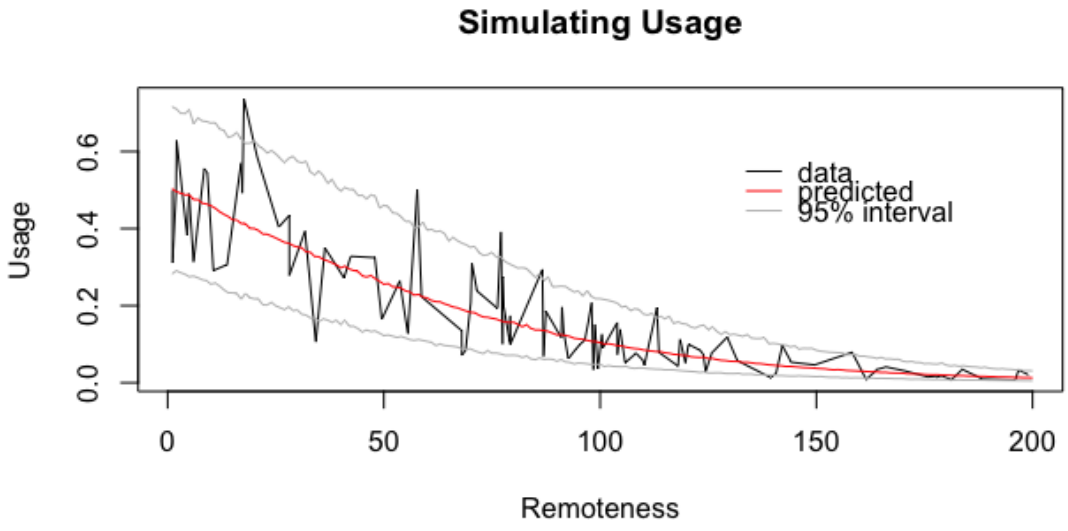

Figure 2: The 95% interval leave out roughly 5 points each time the model is run and the model fits the data quite well

Figure 3 shows that the raw usage values and the predicted usage values follow a roughly similar trend. The simulations are relatively good estimates of the actual data.

[1] see R Code and Stan Code

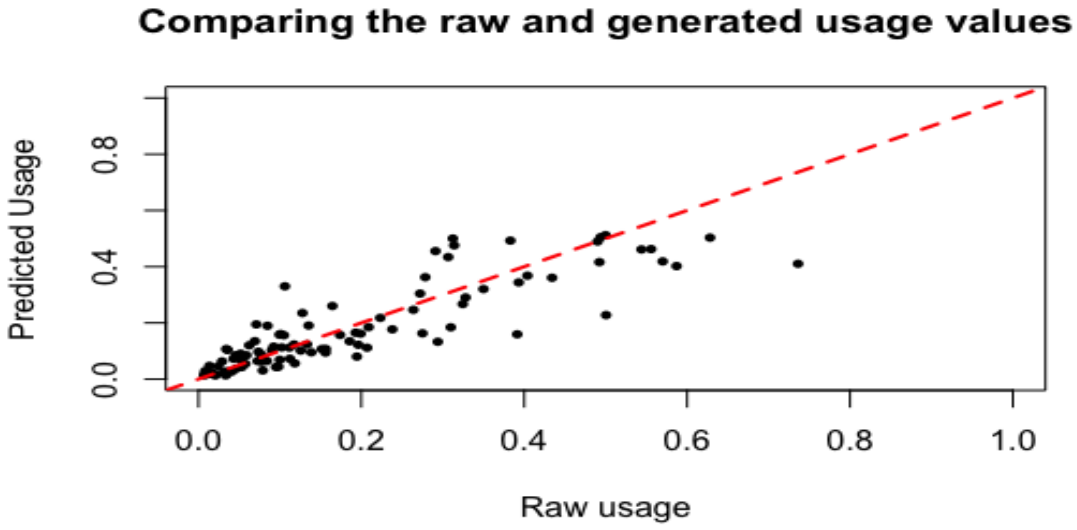

Figure 3: There is a strong correlation between the raw and predicted usage

1.3 Part C

Plot the scatterplot of the raw data, and on top of this, plot the posterior mean, 2.5th percentile, and 97.5th percentile of usage as a function of remoteness. Discuss your model's performance.

Figure 4 shows the desired scatterplot of the raw data with the posterior mean estimates of usage, along with the 95% interval of usage as a function of the remoteness.

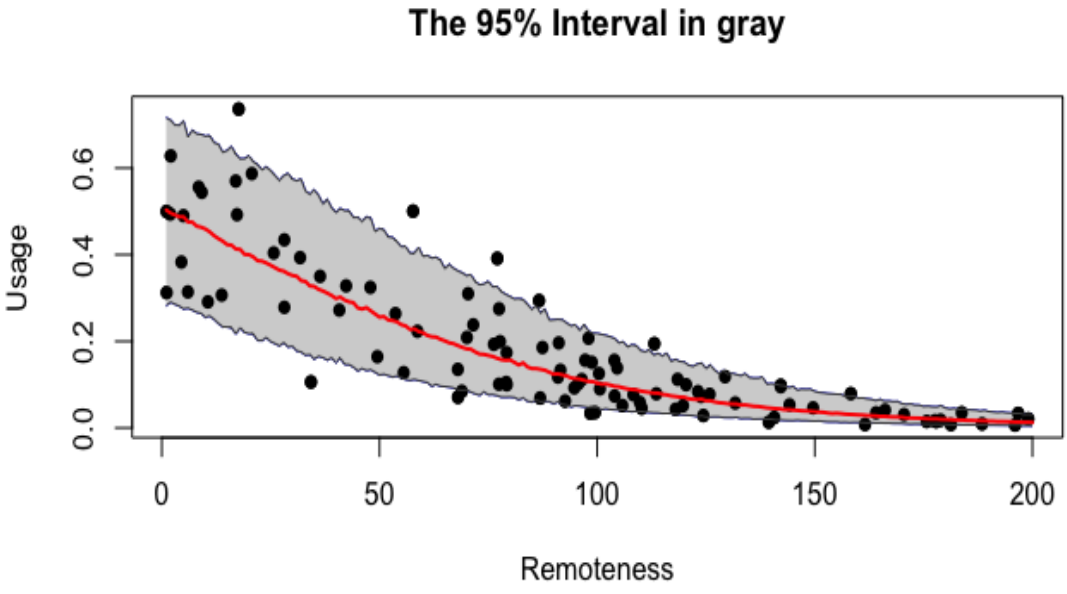

Figure 4: Raw usage, predicted usage, 95% intervals

The model predicts the data fairly well and leaves out some very obvious outliers in the 95% interval. Moreover, the variance parameter for usage converges as remoteness increases which is an accurate depiction of the true usage behavior which starts to cluster as remoteness crosses a 100.

1.4 Part D

Tesla has proposed building its first Russian supercharger near Moscow. It wants to build the supercharger as far away from Moscow as possible but wants the usage to be at least 5% (with high confidence). Use your posterior simulations to suggest the optimal remoteness for this new supercharger.

If Tesla wants to build the supercharger as far away from Moscow as possible but wants to ensure that the usage is at least 5% it should build the supercharger around 95 units of distance away from Moscow. Figure 5 shows the lower bound of the 95% interval.

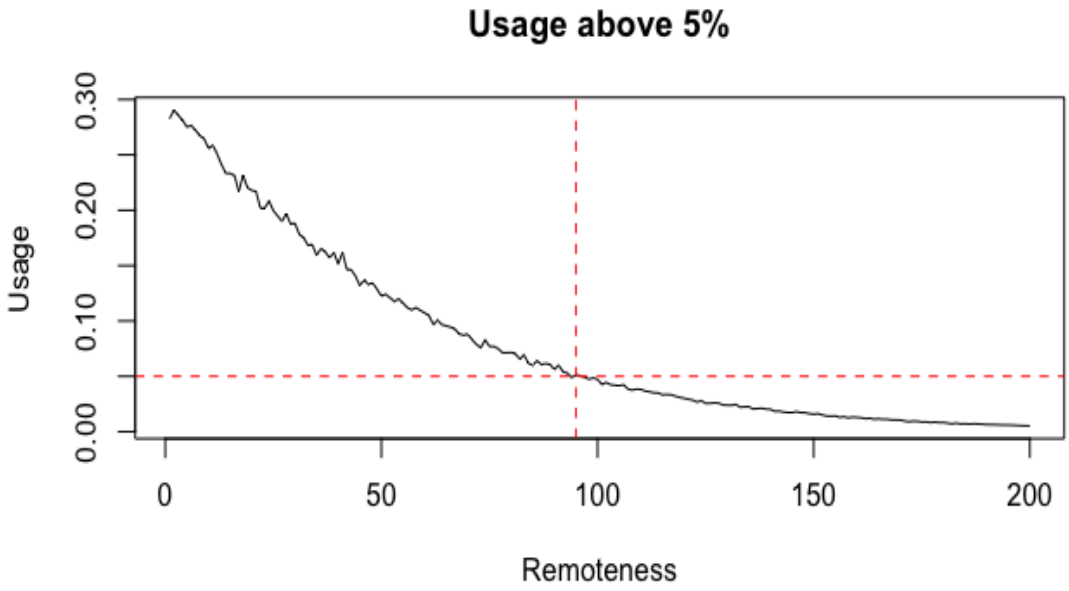

Figure 5: Minimum usage above 5% at 95 units

2 Question 2

Write a few sentences about what you might want to do for your final project. This does not commit you to anything. If you have more than one possible idea, write a sentence on each.

1. Bayesian estimation of the relation between infant mortality rate and health expenditure in India using multilevel modeling to capture time and state-wise variation. I would like to carry out a deeper analysis of the same problem with more sophisticated techniques.

2. To analyze the relationship between Public debt/GDP ratio and its impact on growth. Would like to show that the "distribution" of effects is centered at 0 and has a positive or negative impact on growth based on associated financial stress in the market.

3. Use household survey data to carry out a meta-analysis of voting behavior. This is a large dataset and the possibilities are varied.

3 Code

3.1 R Code

Listing 1: R Code

```
rm(list = ls())
library(foreign)
library(plyr)
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
getwd()
setwd("/Users/Advait/Desktop/New␣School/Fall16/BDA/Class10")
list.files()
df <- read.csv("supercharger.csv")
df <- arrange(df, by = Remoteness)

x <- df$Remoteness
usage <- df$Usage
range(usage)
logit_usage <- NULL
for(i in 1:100){
logit_usage[i] <- log(usage[i]/(1-usage[i]))
}
##Plot1
par(mfcol = c(1,2))
plot(x,usage,xlab = "Remoteness",ylab = "Usage",
     pch = 16, main = "Initial␣usage␣vs␣remoteness")
plot(x, logit_usage, xlab = "Remoteness", ylab = "logit(Usage)",
     pch = 16, main = "logit␣transform␣usage␣vs␣remoteness")
abline(lm(logit_usage ~ x), col = "red", lty = 2, lwd = 2)
legend(100,0, legend = "regression",
       col = "red", lty = 2)
##Stan Fit
stanc("final.stan")
finalfit <- stan("final.stan", data = list("logit_usage", "x"),
                 iter = 1000,
                 chains = 3)
print(finalfit)
ext <- extract(finalfit)
inter_usage <- colMeans(ext$logit_usage_rep)
```

Listing 2: R Code contd.

```
usage_pred <- NULL
for (i in 1:200){
usage_pred[i] <- exp(inter_usage[i])/(1+exp(inter_usage[i]))
}
plot(x,usage, pch = 16, main = "The␣model␣fits␣the␣data␣well",
     xlab = "Remoteness", ylab = "Usage")
lines(c(1:200),usage_pred, col = "red", pch = 16, cex = .7, lwd = 2)
legend(100,0.5, legend = "Logistic␣normal",
       col = "red", lwd = 2)

bar1 <- NULL
bar2 <- NULL
for (i in 1:200){
  bar1[i] <- quantile(ext$logit_usage_rep[,i],probs = 0.025 )
  bar2[i] <- quantile(ext$logit_usage_rep[,i],probs = 0.975 )
}
for (i in 1:200){
  bar1[i] <- exp(bar1[i])/(1+exp(bar1[i]))
  bar2[i] <- exp(bar2[i])/(1+exp(bar2[i]))
}
##Figure 2
plot(x, df$Usage, type = "l", xlab = "Remoteness",
     ylab = "Usage", main = "Simulating␣Usage")
legend(130,0.6, legend = c("data","predicted","95\%␣interval"),
       col = c("black", "red","gray"),lwd = 1, bty = "n")
lines(x, usage_pred, type = "l", col = "red")
lines(x,bar1, col = "gray", pch = 16, cex = .7)
lines(x,bar2, col = "gray", pch = 16, cex = .7)

##Figure 3
par(mfcol = c(1,1))
plot(df$Usage,usage_pred, pch = 16, cex = 0.7,
     xlab = "Raw␣usage",
     ylab = "Predicted␣Usage",
     main = "Comparing␣the␣raw␣and␣generated␣usage␣values",
     xlim = c(0,1),
     ylim = c(0,1))
abline(0,1, col = "red", lty = 2, lwd = 2)
```

Listing 3: R Code contd.

```
##Figure 4
moop <- c(1:200)
plot(x,usage, pch = 16, main = "The␣95\%␣Interval␣in␣gray",
     xlab = "Remoteness", ylab = "Usage")
lines(moop,usage_pred, col = "red", pch = 16, cex = .7, lwd = 2)
lines(moop,bar1, col = "blue", pch = 16, cex = .7)
lines(moop,bar2, col = "blue", pch = 16, cex = .7)
polygon(c(moop, rev(moop)), c(bar2, rev(bar1)),
        col = "lightgray", border = NA)
points(x,usage,pch = 16)
lines(moop,usage_pred, col = "red", pch = 16, cex = .7, lwd = 2)
lines(moop,bar1, col = "gray30", pch = 16, cex = .7)
lines(moop,bar2, col = "gray30", pch = 16, cex = .7)

##Figure 5
plot(x,bar1, pch = 16, main = "Usage␣above␣5%",
     xlab = "Remoteness", ylab = "Usage", type ="l")
abline(h = 0.05, col = "red", lty = 2)
abline(v = 95, col = "red", lty = 2)
```

3.2 Stan Code

Listing 4: Stan Code

```
data{
  real logit_usage[100];
  real x[100];
}
parameters{
  real b0;
  real b1;
  real<lower =0> sigma;
  real error[100];
}
model{
  error ~ normal(0,sigma);
  for (i in 1:100)
  logit_usage[i] ~ normal(b0 + b1*x[i] + error[i], sigma);
}
generated quantities{
  real logit_usage_rep[200];
  for (i in 1:200){
    logit_usage_rep[i] = normal_rng(b0 + b1*i, sigma);
  }
}
```