STAT 578 - Fall 2019 - Data Analyis Report

Frederick (Eric) Ellwanger - fre2

December 14, 2019

Introduction

Marijuana contains a mind-altering chemical known as THC, which can be found in the leaves and buds of the plant. Marijuana can be ingested in several different ways, most notably by smoking, eating, or drinking teas made with the leaves and buds.[1] Marijuana use is especially a concern among adolescents because it has been shown to have negative consequences on educational outcomes of frequent marijuana users. Another concern about adolescent marijuana use is the unkown impact it may have on the still developing brains of these young people.

Marijuana use among adolescents hit a peak around 1980 of close to 35% of surveyed repondants having used marijuana within the last 30 days. Those numbers have decreased, but have risen in recent years to close to 23% due to a belief that it is safe to use since many states have legalized the medical use of marijuana, and some states have outright legalized the use of recreational marijuana use.[2]

Studies have shown a difference in marijuana use by gender. There are probably a myriad of reasons for this, but it is important to understand if an organization would want to target information to males and females to try to reduce the usage rates among adolescents overall.[3]

Data

The data for this analysis can be found in the file `marijuanause.csv` and contains 236 rows of data representing indivdual adolescents. Each row contains 6 columns of data:

column_name	description
female	A value 0 or 1 representing an indicator variable for whether the adolescent is female (1) or male (0).
use1976	A value 0 or 1 representing an indicator variable for whether the adolescent used marijuana (1) or not (0) in the year 1976.
use1977	A value 0 or 1 representing an indicator variable for whether the adolescent used marijuana (1) or not (0) in the year 1977.
use1978	A value 0 or 1 representing an indicator variable for whether the adolescent used marijuana (1) or not (0) in the year 1978.
use1979	A value 0 or 1 representing an indicator variable for whether the adolescent used marijuana (1) or not (0) in the year 1979.
use1980	A value 0 or 1 representing an indicator variable for whether the adolescent used marijuana (1) or not (0) in the year 1980.

Table 1 shows the total number of adolescents that used marijuana in a particular year as well as the fraction of the sample population that represents. The table clearly shows a rising use of marijuana among adolescents in this survey.

Table 2 shows a break down by gender. From this table it is clear that Male adolescents in this survey have a higher proportion of marijuana use than the Female adolescents in every year of this study.

[1] National Institute on Drug Abuse, https://www.drugabuse.gov/publications/research-reports/marijuana/]

[2] Child Trends, https://www.childtrends.org/indicators/marijuana-use

[3] Gender Differences in Adolescent Marijuana Use and Associated Psychosocial Characteristics, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3359836/

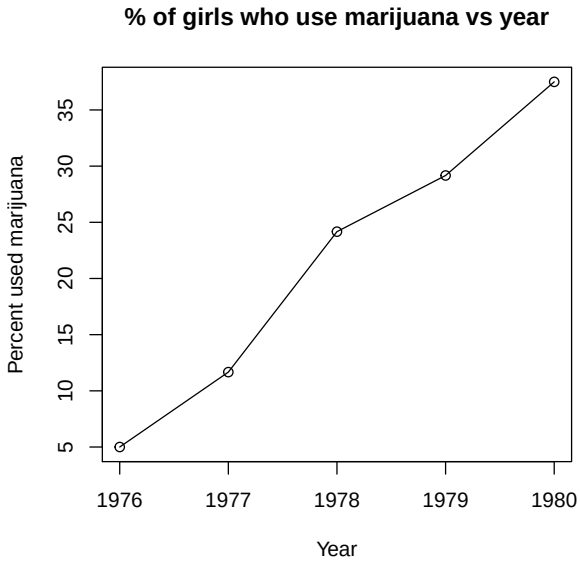
Table 1: Data Summary

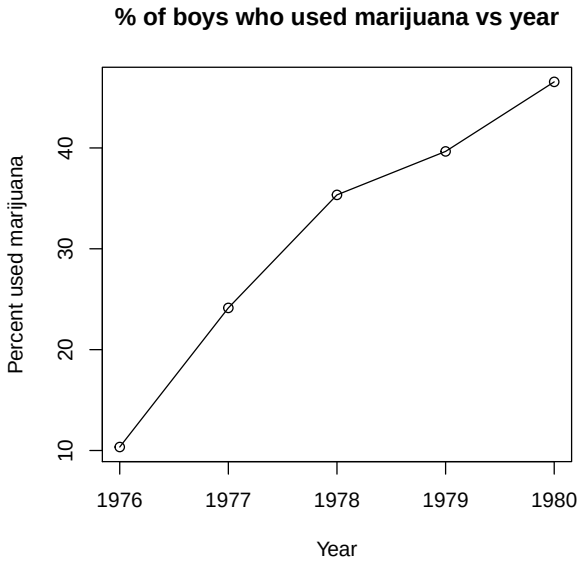
	female	use1976	use1977	use1978	use1979	use1980
Count	120.000	18.000	42.000	70.000	81.000	99.000
Percentage	0.508	0.076	0.178	0.297	0.343	0.419

Table 2: Usage rates by gender and year

	1976	1977	1978	1979	1980
Female					
# Female Use	6.000	14.000	29.000	35.000	45.000
Proportion Female Use	0.050	0.117	0.242	0.292	0.375
Male					
# Male Use	12.000	28.000	41.000	46.000	54.000
Proportion Male Use	0.103	0.241	0.353	0.397	0.466

The below graphs show the rates of marijuana use among boys and girls in the years 1976-1980. These graphs clearly show an increasing use of marijuana over the study years. It can also be seen that the proportion of male adolescents using marijuana in a given year is greater than the proportion of female adolescent using marijuana.





First Model

First Model - Section A

Here is the code for the First model using logistic regression with an ordinary intercept term, a coefficient for the centered and rescaled year number, and a coefficient multiplying the centered indicator of being female with no error term.

```
model {
   for (i in 1:236) {
      for (j in 1:5) {
           uses[i, j] ~ dbern(prob[i, j])
         logit(prob[i, j]) <- beta0 + beta.yr*year.scaled[j] + beta.sex*sex.cent[i]

           usesrep[i, j] ~ dbern(prob[i, j])
       }
   }
   beta0 ~ dt(0, 1/(10^2), 1)
   beta.yr ~ dt(0, 1/(2.5^2), 1)
   beta.sex ~ dt(0, 1/(2.5^2), 1)
}
```

Model 1 - Section B

The First model was run using 4 chains. 2,000 iterations per chain of burn in was used with overdispersed starting points for the top level parameters. 10,000 iterations per chain were then used to check for convergence for 'beta0' (the intercept), 'beta.yr' (the coefficient for the centered and scaled year number), and 'beta.sex' (the coefficient for the centered indicator for being female). Gelman Rubin diagnostics and trace plots showed no issues with convergence. After showing convergence diagnostics were good, an additional 2,000 iterations per chain were used to get information for the above parameters as well as probabilities for marijuana use in a given year. Thinning was not used in this analysis.

Table 3: MCMC Summary for Model 1

Chains	4
BurnIn	2000
ConvergenceTesting	10000
Iterations	2000
Thinning	1

The effective number of parameters was over 4,000 for each parameter.

Table 4: Effective Sample Sizes for parameters of Model 1

	Effective Sample Sizes
beta0	4149
beta.yr	4401
beta.sex	4913

Model 1 - Section C

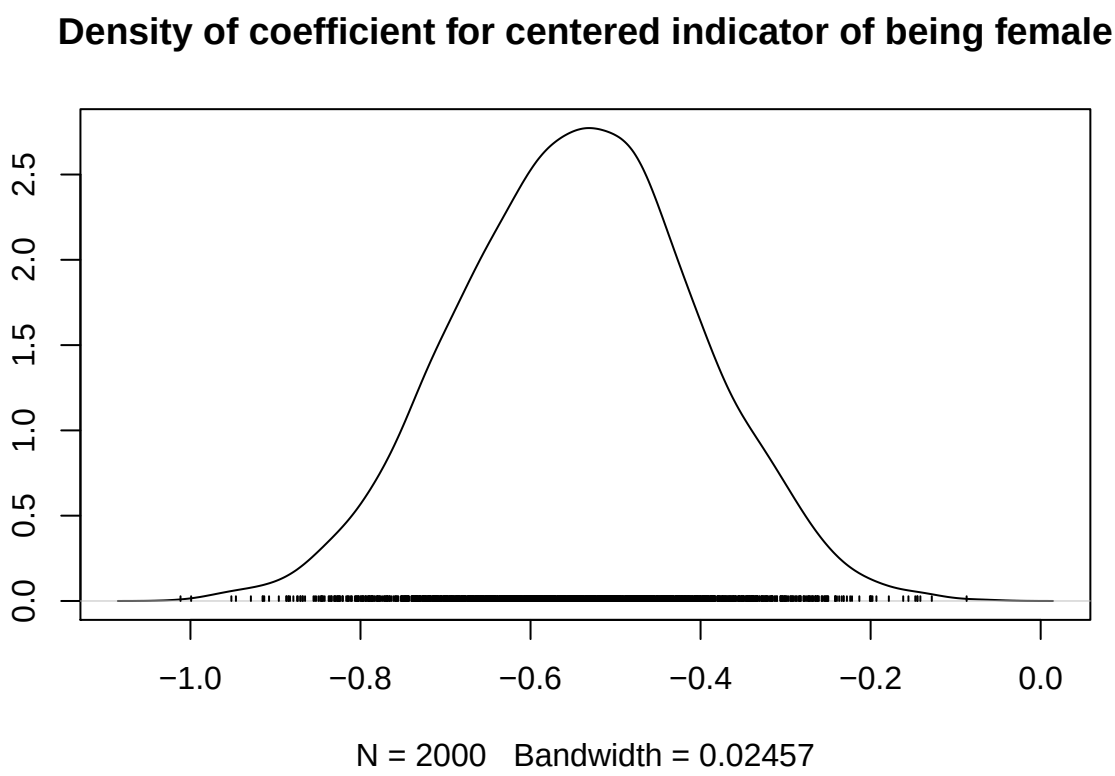
The approximate means, standard deviations, and 95% posterior central interval - represented by the values between the 2.5% and 97.5% approximate posterior quantiles - for Model 1 are shown in the table below.

Table 5: Approximate Posterior data for parameters of Model 1

	Mean	Std Dev	2.5% Post Qtl	97.5% Post. Qtl
beta0	-1.153	0.075	-1.303	-1.004
beta.yr	1.499	0.166	1.175	1.833
beta.sex	-0.545	0.140	-0.821	-0.278

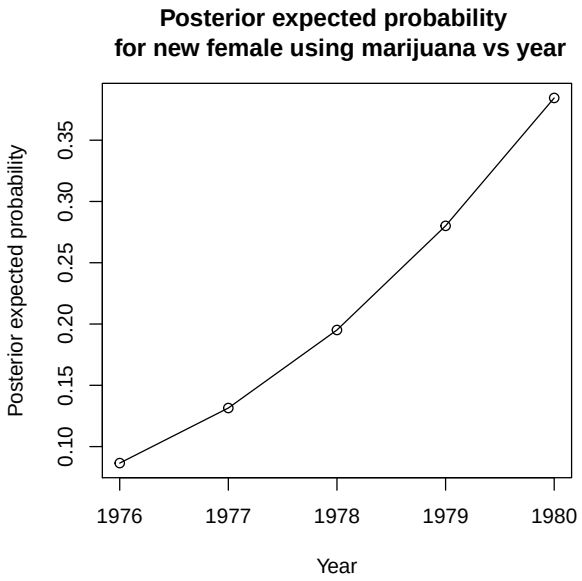
Model 1 - Section D

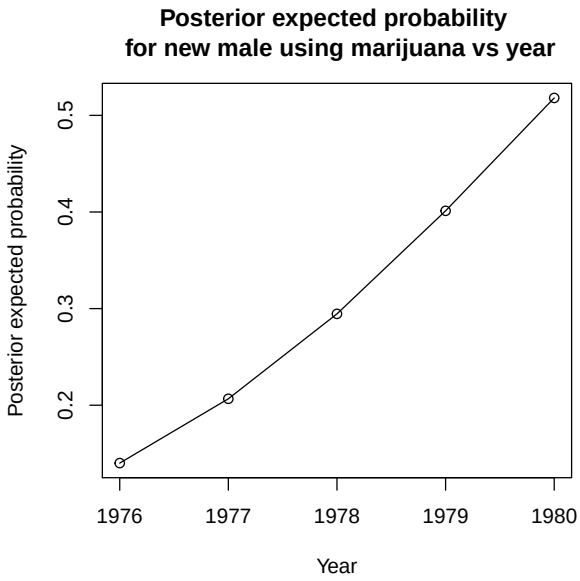
The approximate posterior probability that the coeffiecient for the centered indicator of being female is greater than 0 ($Pr(\beta.female > 0)$ is equal to 0. This indicates that for a given year that females appear to be less likely to use marijuana than males This can clearly be seen in the density plot below.



Model 1 - Section E

The below graphs show the posterior expected probability that a newly sampled female and male adolescent would use marijuana for each year. A clear trend of increasing use for both males and females over the years can be seen. The charts also show that males have a higher rate of using marijuana than females in any given year.





Model 1 - Section F

The value of Plummer's DIC is approximated using 100,000 iterations:

The approximate effective number of parameters for Model 1 is 3.002.
The approximate penalized deviance - Plummers DIC - for Model 1 is 1255.192

Second Model

Model 2 - Section A

Here is the code for the Second model using logistic regression with an ordinary intercept term, a coefficient for the centered and rescaled year number, and a coefficient multiplying the centered indicator of being female with a separate random effect for each adolescent.

```
model {
   for (i in 1:236) {
      for (j in 1:5) {
           uses[i, j] ~ dbern(prob[i, j])
         logit(prob[i, j]) <- beta0 + beta.yr*year.scaled[j] + beta.sex*sex.cent[i] + epsilon[i]

           usesrep[i, j] ~ dbern(prob[i, j])
      }
      epsilon[i] ~ dnorm(0, 1/(sigmaperson^2))
   }

   beta0 ~ dt(0, 1/(10^2), 1)
   beta.yr ~ dt(0, 1/(2.5^2), 1)
   beta.sex ~ dt(0, 1/(2.5^2), 1)
   sigmaperson ~ dunif(0,100)
}
```

Model 2 - Section B

The second model was run using 4 chains. 8,000 iterations per chain of burn in was used with overdispersed starting points for the top level parameters. 20,000 iterations per chain were then used to check for convergence for 'beta0' (the intercept), 'beta.yr' (the coefficient for the centered and scaled year number), 'beta.sex' (the coefficient for the centered indicator for being female), and 'sigmaperson' (the random effect for each person). Gelman Rubin diagnostics and trace plots showed no issues with convergence. After showing convergence diagnostics were good, an additional 40,000 iterations per chain were used to get information for the above parameters as well as probabilities for marijuana use in a given year. Thinning was not used in this analysis.

Table 6: MCMC Summary for parameters of Model 2

Chains	4
BurnIn	8000
ConvergenceTesting	20000
Iterations	40000
Thinning	1

The effective number of parameters was over 4,000 for each parameter.

Table 7: Effective sample sizes for parameters of Model 2

	Effective Sample Size
beta0	5556
beta.yr	16619
beta.sex	9875
sigmaperson	6891

Model 2 - Section C

The approximate means, standard deviations, and 95% posterior central interval - represented by the values between the 2.5% and 97.5% approximate posterior quantiles - for Model 2 are shown in the table below.

Table 8: Approximate posterior data for parameters of Model 2

	Mean	Std Dev	2.5% Post Qtl	97.5% Post. Qtl
beta0	-2.40	0.279	-2.98	-1.891
beta.yr	2.97	0.288	2.42	3.550
beta.sex	-1.01	0.458	-1.93	-0.131
sigmaperson	2.95	0.307	2.40	3.599

Model 2 - Section D

The value of Plummer's DIC is approximated using 100,000 iterations:

The approximate effective number of parameters for Model 2 is 167.779.
The approximate penalized deviance - Plummers DIC - for Model 2 is 674.192

This value is considerably smaller than the Plummers DIC value for Model 1. Model 2 seems to be the preferred model to use.

Conclusions

The second model that includes an indivisual error term for each adolescent does a better job at fitting the data. This is evident in the Plummers DIC values, where Model 2 is significantly lower than Model 1.

This model shows an overall increase in marijuana use in both males and females as the year number increases. The model further shows that males are more likely to use marijuana in a given year than females.

Appendix

This portion of the report contains all of the r code used in this analysis.

```
#Code for Part 2 - Data Description

library(knitr) # loads the kable package required for kableExtra
library(kableExtra)

#df <- data.frame(column_name = character(), description = character())

df = data.frame(column_name = 'female', description = 'A value 0 or 1 representing an indicator variable

df = rbind(df, data.frame(column_name = 'use1976', description = 'A value 0 or 1 representing an indica

df = rbind(df, data.frame(column_name = 'use1977', description = 'A value 0 or 1 representing an indica

df = rbind(df, data.frame(column_name = 'use1978', description = 'A value 0 or 1 representing an indica

df = rbind(df, data.frame(column_name = 'use1979', description = 'A value 0 or 1 representing an indica

df = rbind(df, data.frame(column_name = 'use1980', description = 'A value 0 or 1 representing an indica

kable(df, booktabs = T)  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, bold = T)
```

```
#Code for Part 2 - Data Summary

#Read in data
marijuana <- read.csv("marijuanause.csv")

#Get summary statistics
Total = sapply(marijuana, sum)
Percentage = Total/nrow(marijuana)
stats = rbind(Total, Percentage)

#table(as.data.frame(stats))
kable(stats, caption = "Data Summary")  %>%
  kable_styling(full_width = T, latex_options = "hold_position")
```

```
#Code for Part 2 - Data Plots

#subset girls
girls = subset(marijuana[2:6], marijuana$female == 1)
girlsuse = apply(girls, 2,mean)
girlstotal = apply(girls, 2, sum)

#subset boys
boys = subset(marijuana[2:6], marijuana$female == 0)
boysuse = apply(boys, 2, mean)
boystotal = apply(boys, 2, sum)

both = as.data.frame(rbind(girlstotal, girlsuse, boystotal, boysuse))
```

```r
row.names(both) = c("# Female Use", "Proportion Female Use", "# Male Use", "Proportion Male Use")
colnames(both) = c("1976", "1977", "1978", "1979", "1980")

kable(both, caption = "Usage rates by gender and year")  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  pack_rows("Female", 1, 2) %>%
  pack_rows("Male", 3, 4)

#setup for 2 plots side by side
par(mfrow=c(1, 2))
plot(girlsuse, type = 'o', main="Proportion of girls who use marijuana vs year",
     ylab="Proprtion used marijuana", xlab="Year", xaxt='n')
axis(side=1, at=c(1:5), labels=c("1976","1977","1978","1979","1980"))

plot(boysuse, type = 'o', main = "Proportion of boys who used marijuana vs year",
     ylab="Proportion used marijuana", xlab="Year", xaxt='n')
axis(side=1, at=c(1:5), labels=c("1976","1977","1978","1979","1980"))
```

```r
#Code for Model 1 - Part A

years = c(1976, 1977, 1978, 1979, 1980)

d1 <- list(uses = as.matrix(marijuana[,2:6]),
           sex.cent = as.vector(scale(marijuana$female, scale = FALSE)),
           year.scaled = as.vector(scale(years, scale=2*sd(years))))

inits1 <- list(list(beta0=10, beta.yr=10, beta.sex=10),
               list(beta0=-10, beta.yr=-10, beta.sex=-10),
               list(beta0=-10, beta.yr=10, beta.sex=-10),
               list(beta0 = 10, beta.yr=-10, beta.sex=10))


library(rjags)
m1 <- jags.model("model1.bug", d1, inits1, n.chains=4, n.adapt=1000, quiet = TRUE)

#Burn-in
update(m1, 2000)

#Get samples
x1 <- coda.samples(m1, c("beta0", "beta.yr","beta.sex"), n.iter=10000)

#Check convergence
gelman.diag(x1, autoburnin=FALSE)

#Another check for convergence
gelman.plot(x1)

#Another check for convergence
plot(x1)

#Add additional variable - prob and yrep
x1 <- coda.samples(m1, c("beta0","beta.yr","beta.sex","prob","usesrep"), n.iter=2000)
```

```r
#Check for effective sample size > 4000
effsize = data.frame(effectiveSize(x1[,c("beta0", "beta.yr", "beta.sex")]))
```

```r
#Code for Model 1 - Part C

#Create summary of MCMC x1
x1.sum = summary(x1[, c("beta0", "beta.yr", "beta.sex")])
x1sumdf1 = data.frame(Chains = x1.sum$nchain)
x1sumdf2 = data.frame(BurnIn = 2000)
x1sumdf3 = data.frame(Iterations = (x1.sum$end - x1.sum$start + 1))
x1sumdf4 = data.frame(Thinning = x1.sum$thin)
x1sumall = cbind(x1sumdf1, x1sumdf2, x1sumdf3, x1sumdf4)
x1sumall = t(x1sumall)
kable(x1sumall, caption = "Table 3 - MCMC Summary")  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, bold = T)

colnames(effsize) = "Effective Sample Size"
kable(effsize, caption = "Effective Sample Sizes")  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, bold = T)
```

```r
#Code for Model 1 - Part D

beta_sex = as.matrix(x1)[, "beta.sex"]
bg1 = mean(beta_sex > 0)

#Plot density of beta.sex
densplot(x1[,"beta.sex"], main = "Density of coefficient for centered indicator of being female")
```

```r
#Code for Model 1 - Part E

inv.logit = function(x){
  1/(1+exp(-x))
}

year.scaled = as.vector(scale(years, scale=2*sd(years)))

beta_0 = as.matrix(x1)[, "beta0"]
beta_yr = as.matrix(x1)[, "beta.yr"]
beta_sex = as.matrix(x1)[, "beta.sex"]

#pFUse.post = matrix(nrow = 8000, ncol = 5)
#pMUse.post = matrix(nrow = 8000, ncol = 5)
pFUse.post = rep(NA, 5)
pMUse.post = rep(NA, 5)

for(i in 1:5) {
  #Posterior probability of a female using marijuana for a given year
  pFUse.post[i] = mean(inv.logit(beta_0 + beta_yr*year.scaled[i] +
                                 beta_sex*(1 - mean(marijuana$female))))

  #Posterior probability of a male using marijuana for a given year
```

```
  pMUse.post[i] = mean(inv.logit(beta_0 + beta_yr*year.scaled[i] +
                                 beta_sex*(0 - mean(marijuana$female))))
}

#Plot
par(mfrow=c(1, 2))
plot(pFUse.post, type = 'o', main="Posterior expected probability \n for new female using marijuana vs y
        ylab="Posterior expected probability", xlab="Year", xaxt='n')
axis(side=1, at=c(1:5), labels=c("1976","1977","1978","1979","1980"))
plot(pMUse.post, type = 'o', main="Posterior expected probability \n for new male using marijuana vs yea
        ylab="Posterior expected probability", xlab="Year", xaxt='n')
axis(side=1, at=c(1:5), labels=c("1976","1977","1978","1979","1980"))
```

```
#Code for Model 1 - Part F

load.module("dic")

mod1dic = dic.samples(m1, 100000)
```

```
#Code for Model 2 - Part A
years = c(1976, 1977, 1978, 1979, 1980)

d2 <- list(uses = as.matrix(marijuana[,2:6]),
           sex.cent = as.vector(scale(marijuana$female, scale = FALSE)),
           year.scaled = as.vector(scale(years, scale=2*sd(years))))

inits2 <- list(list(beta0=5, beta.yr=5, beta.sex=5, sigmaperson = 0.5),
               list(beta0=-5, beta.yr=-5, beta.sex=-5, sigmaperson = 5),
               list(beta0=5, beta.yr=5, beta.sex=-5, sigmaperson = 0.5),
               list(beta0=-5, beta.yr=-5, beta.sex=5, sigmaperson = 5))


library(rjags)
m2 <- jags.model("model2.bug", d2, inits2, n.chains=4, n.adapt=1000, quiet = TRUE)

#Burn-in
update(m2, 8000)

#Get samples
x2 <- coda.samples(m2, c("beta0", "beta.yr","beta.sex", "sigmaperson"), n.iter=20000)

#Check convergence
#gelman.diag(x2, autoburnin=FALSE)

#Another check for convergence
#gelman.plot(x2)

#Another check for convergence
#plot(x2)

#Add additional variable - prob and yrep
x2 <- coda.samples(m2, c("beta0","beta.yr","beta.sex","prob","sigmaperson","usesrep"),
                   n.iter=40000)
```

```r
#Check for effective sample size > 2000
eff2size = data.frame(effectiveSize(x2[,c("beta0", "beta.yr", "beta.sex", "sigmaperson")]))
```

```r
#Code for Model 2 - Part B

#Create summary of MCMC x2
x2.sum = summary(x1[, c("beta0", "beta.yr", "beta.sex")])
x2sumdf1 = data.frame(Chains = x2.sum$nchain)
x2sumdf2 = data.frame(BurnIn = 8000)
x2sumdf3 = data.frame(Iterations = (x2.sum$end - x2.sum$start + 1))
x2sumdf4 = data.frame(Thinning = x2.sum$thin)
x2sumall = cbind(x2sumdf1, x2sumdf2, x2sumdf3, x2sumdf4)
x2sumall = t(x2sumall)
kable(x2sumall, caption = "Table 6 - Model 2 MCMC Summary")  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, bold = T)

colnames(eff2size) = "Effective Sample Size"
kable(eff2size, caption = "Table 7 - Model 2 Effective sample size")  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, bold = T)
```

```r
#Code for Model 2 - Part C

#Create summary of MCMC x2
x2.sum = summary(x2[,c("beta0", "beta.yr", "beta.sex", "sigmaperson")])

#Get posterior means of all parameters
x2.post.mean = x2.sum$statistics[,"Mean"]

#Get posterior standard deviations of all parameters
x2.post.sd = x2.sum$statistics[,"SD"]

#Get 95% posterior central interval for all parameters
x2post.95interval = x2.sum$quantiles[, c("2.5%", "97.5%")]

all = as.data.frame(cbind(x2.post.mean, x2.post.sd, x2post.95interval))
colnames(all) = c("Mean", "Std Dev", "2.5%", "97.5%")
row.names(all) = c("beta0", "beta.yr", "beta.sex", "sigmaperson")

kable(all, caption = "Table 8 - Model 2 Approximate posterior data")  %>%
  kable_styling(full_width = T, latex_options = "hold_position") %>%
  column_spec(1, bold = T)
```

```r
#Code for Model 2 - Part D
mod2dic = dic.samples(m2, 100000)
```