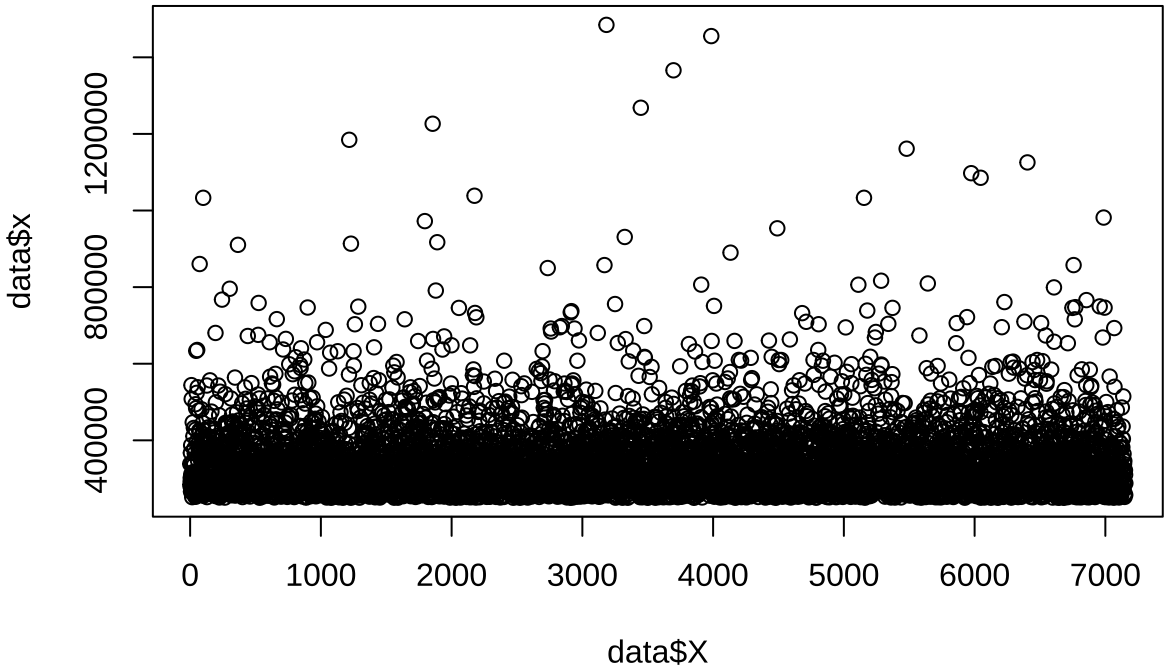
takehome

2024-11-02

Question 1

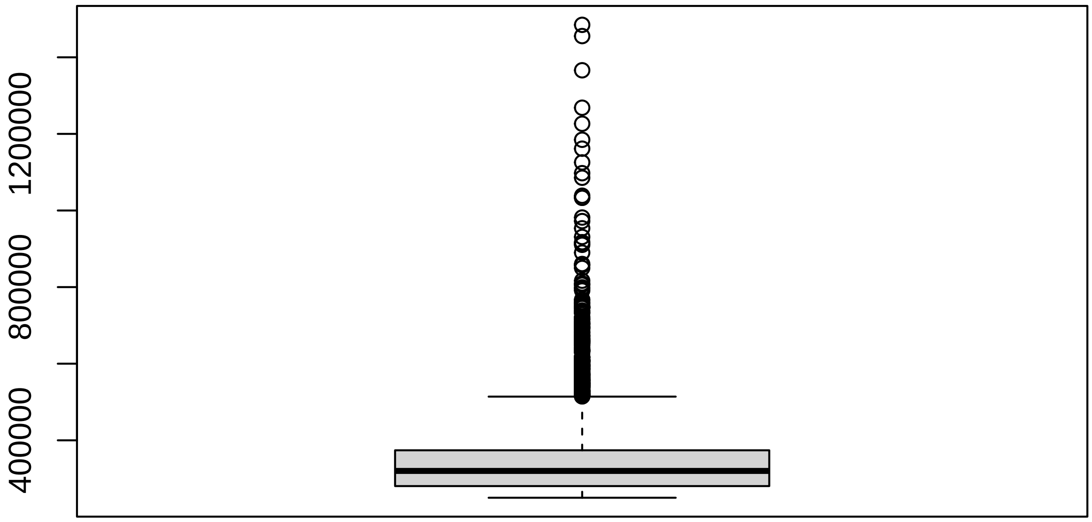
```
data <- read.csv("income20train-2.csv", header = T)
data[,2] -> y
```

```
plot(data$X,data$x)
```

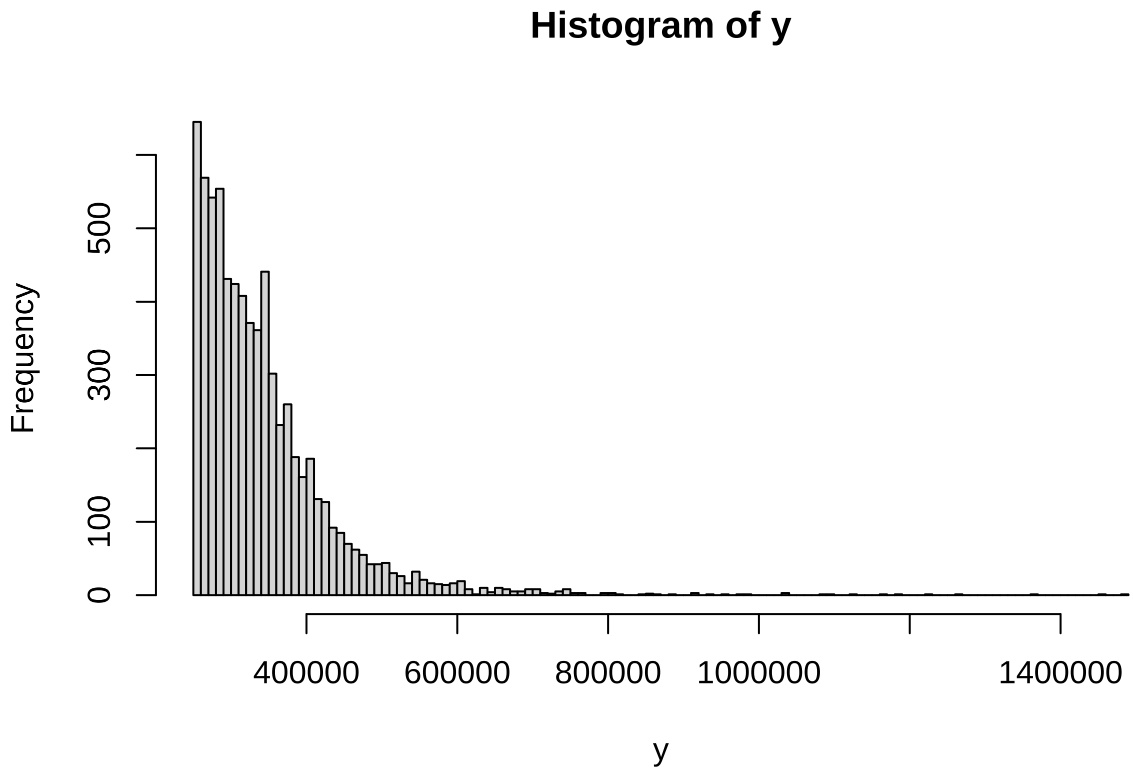


(a)

```
boxplot(y)
```



```
hist(y, breaks = 100)
```



```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  250034  280433  320084  341938  373952 1484705
```

```
mean(y)
```

```
## [1] 341938.1
```

Important to note that the incomes tend to be dense towards the minimum rather than being in the middle; these numbers skew heavily to the right. We can try to fit a Pareto distribution to this later to see how is models the data and run our test set against it. We have a median of 320084 and a mean of 341938.1 while he have a max of 1484705. Also note the amount of outliers in the boxplot and where these outliers are. Also note how these outliers range wider than that of the not unusual data. This shows support to the phenomenon of American income inequality (even among the 20% richest in the nation).

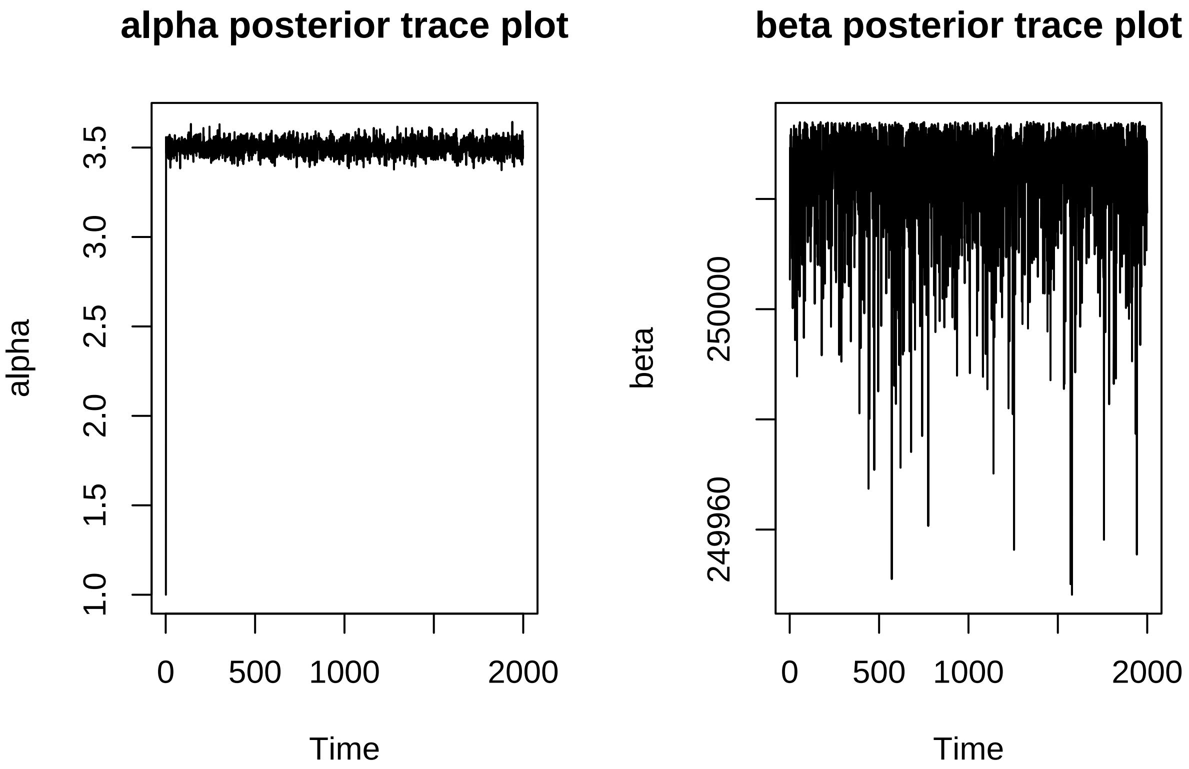
(b)

```
- consider an initial $\alpha_{b-1}$
- sample $\beta_{b-1}$ from $Mono(n\alpha_{b-1}+1, min(y_1, ..., y_n))$
- enter loop for $B$ amount of times
    – sample $\alpha_b$ from $\Gamma(n+1, \sum_{i=1}^{n}[ln(y_i)] - nln(\beta_{b-1}))$
    – sample $\beta_b$ from $Mono(n\alpha_b+1, min(y_1, ..., y_n))$
    – $\beta_b$ is set to $\beta_{b-1}$
- these $\alpha$'s and $\beta$'s are stored in a vector.
```

```
# prep the mono sample function
```

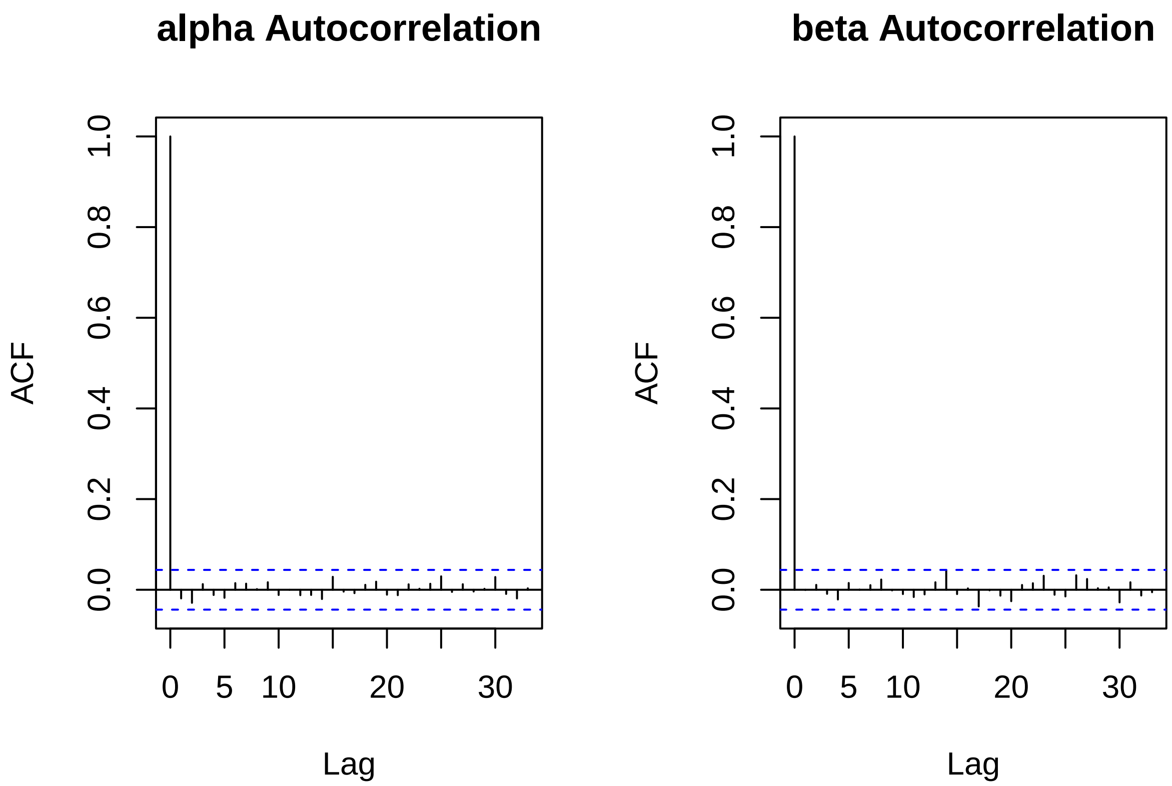
```r
rmono = function(n,alpha,beta){
  u = runif(n)
  x = exp(log(beta) + (log(u)/alpha))
  return(x)
}
# make this the last this
GIBBS1 <- function(B,data,alpha.init){
  set.seed(90909)
  # define n
  n = length(data)
  # make vector spaces
  alpha <- beta <- rep(0,B)
  # apply initials
  alpha[1] = alpha.init
  beta[1] = rmono(1,n*alpha.init + 1, min(data))

  #loop
  for(b in 2:B){
    rgamma(1,n+1,sum(log(data)) - n*log(beta[b-1])) -> alpha[b]
    rmono(1,n*alpha[b] + 1, min(data)) -> beta[b]
  }
  # output
  theta = cbind(alpha,beta)
  return(theta)
}
GIBBS1(2000,y,1) -> theta
```

```r
par(mfrow = c(1,2))
plot.ts(theta[,1], main = "alpha posterior trace plot", ylab = "alpha")
plot.ts(theta[,2], main = "beta posterior trace plot", ylab = "beta")
```



(c)

```
acf(theta[,1], main = "alpha Autocorrelation")
acf(theta[,2], main = "beta Autocorrelation")
```

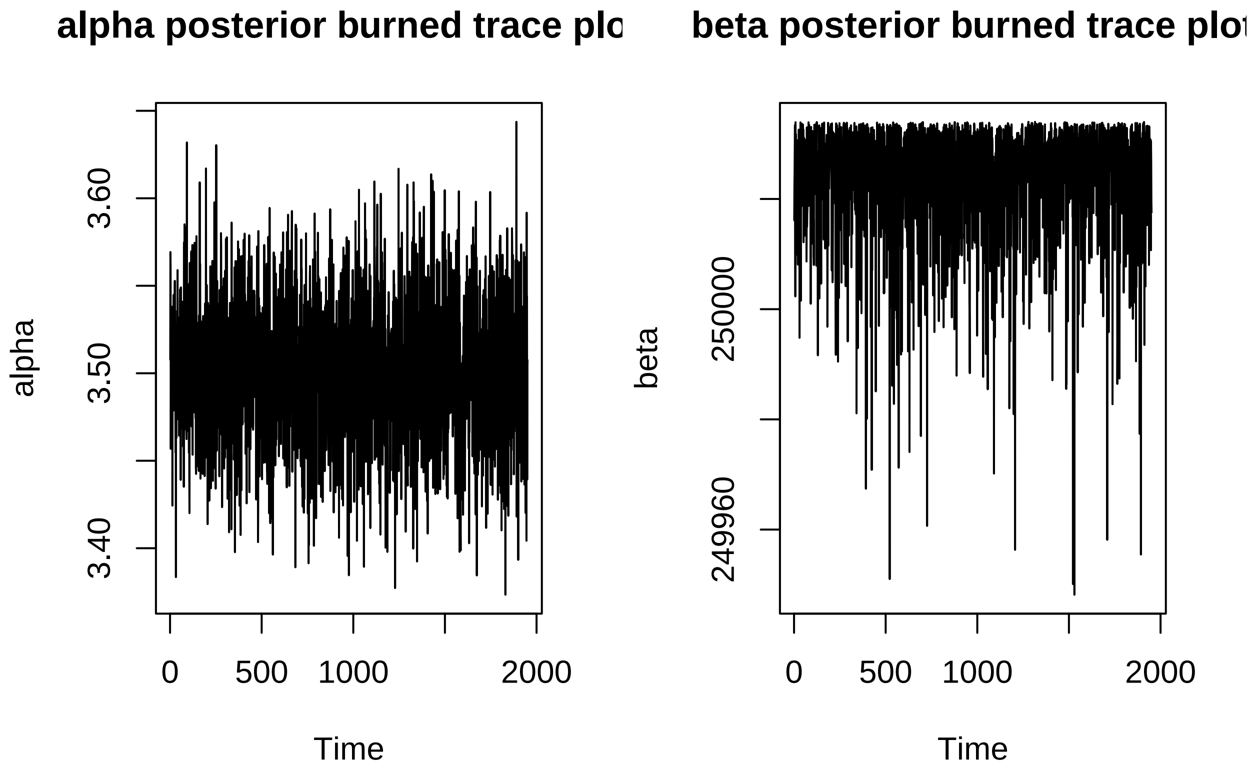


alpha Autocorrelation

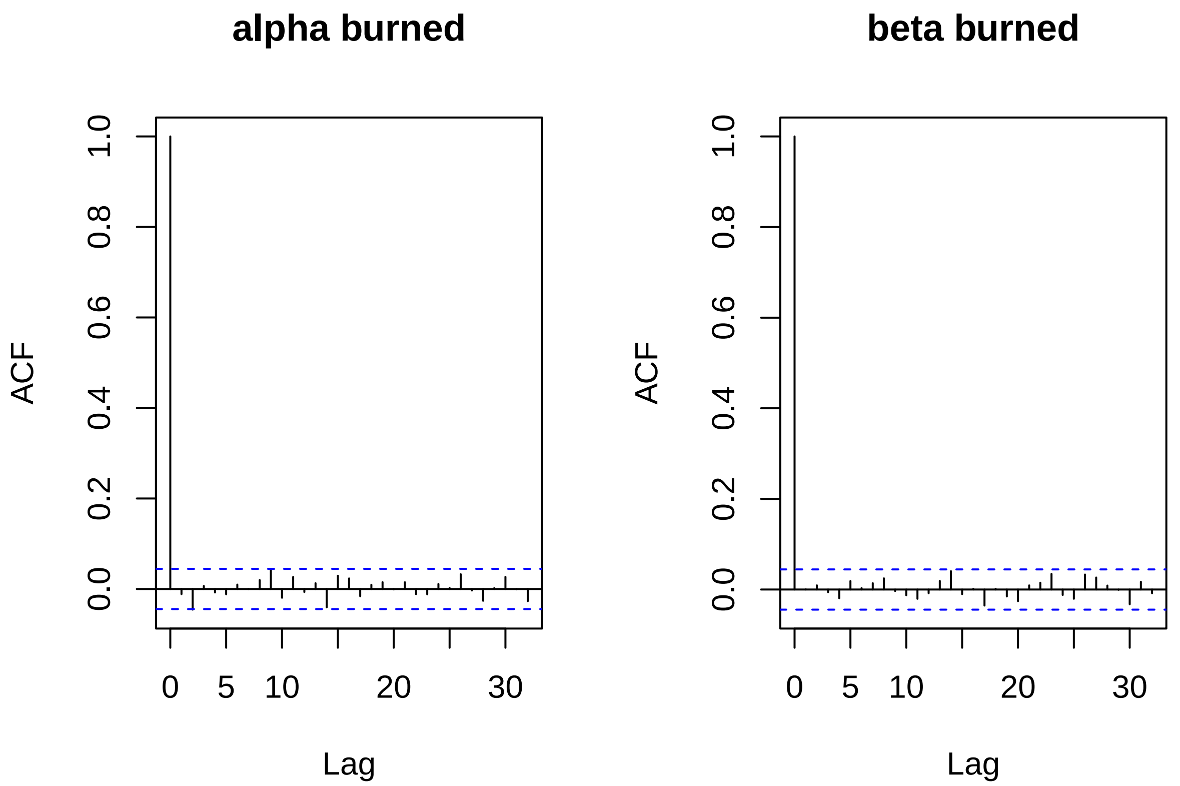
beta Autocorrelation

There seems to be clear stabilization with β and α doesn't need much to be burnt off. I will remove the the first 49.

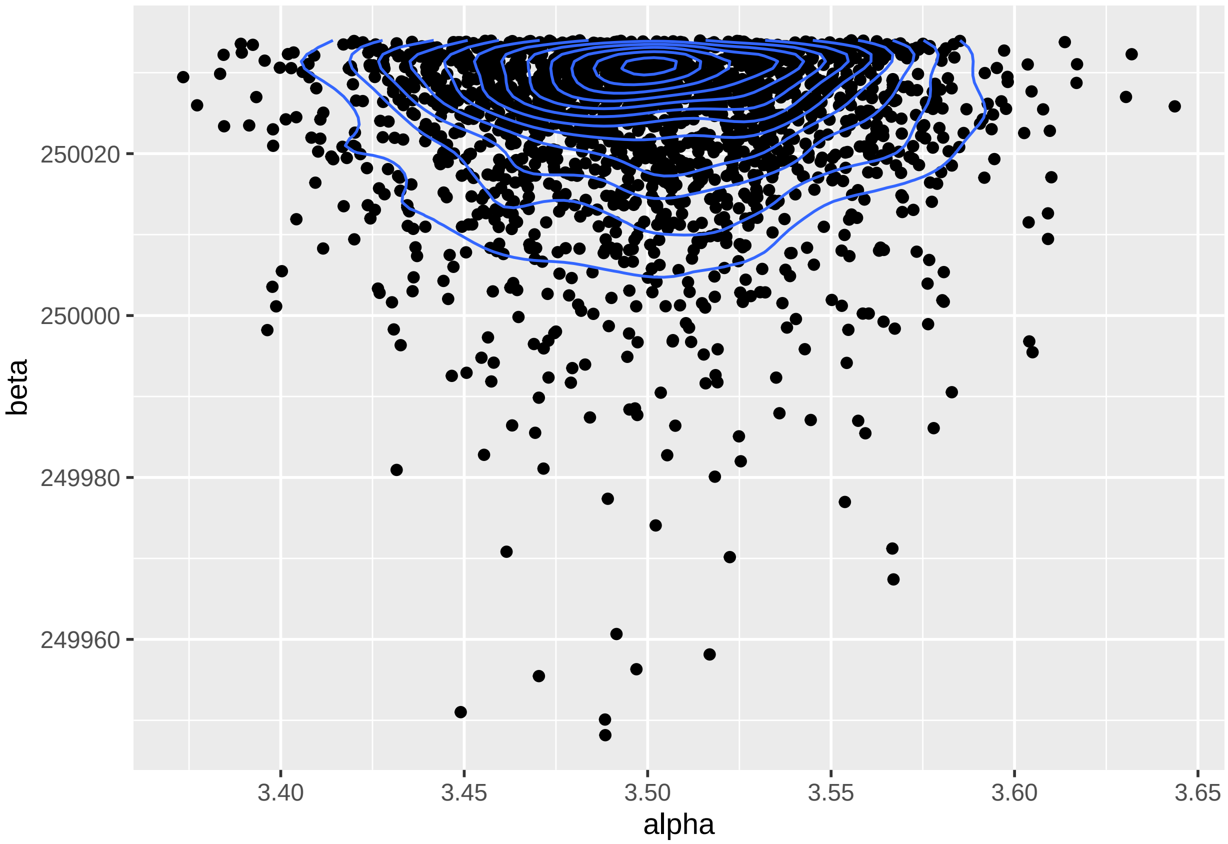
```r
# burning
B = length(theta[,1])
theta.burn = theta[50:B,]
# visuals
par(mfrow = c(1,2))
plot.ts(theta.burn[,1], main = "alpha posterior burned trace plot", ylab = "alpha")
plot.ts(theta.burn[,2], main = "beta posterior burned trace plot", ylab = "beta")
```



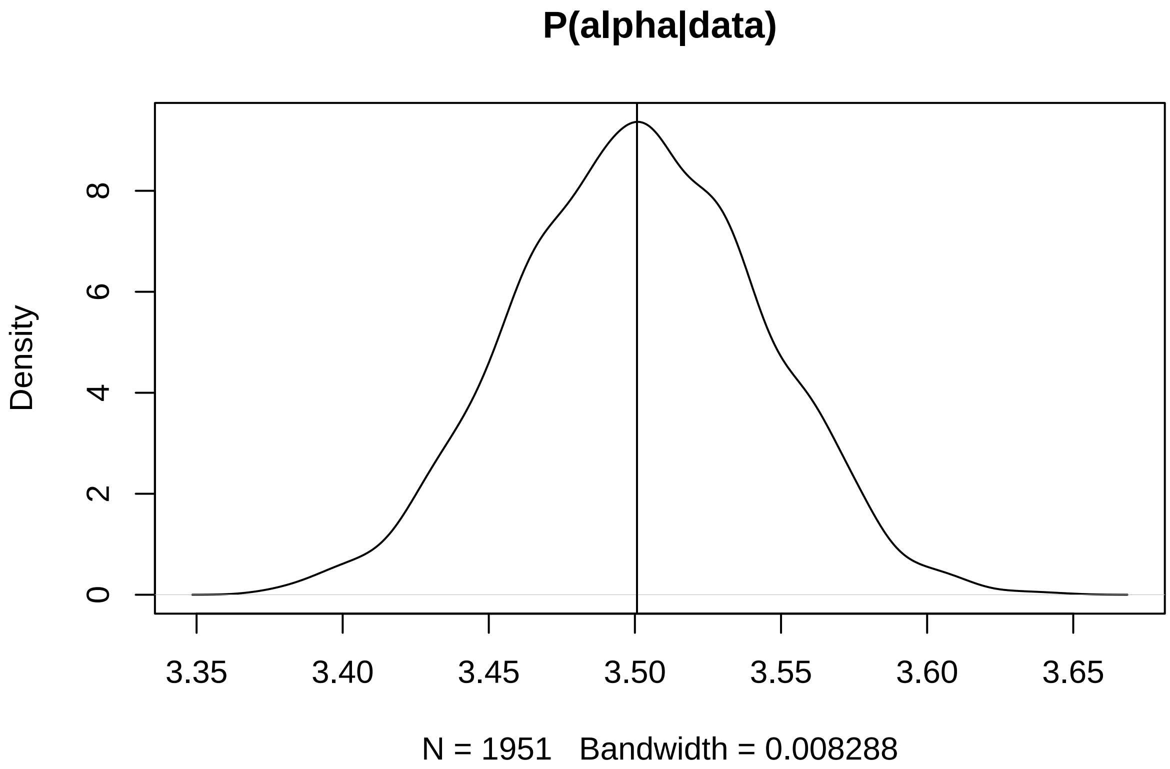
```
acf(theta.burn[,1], main = "alpha burned")
acf(theta.burn[,2], main = "beta burned")
```



```r
# generate bivariate scatterplot
post <- data.frame(theta.burn)
ggplot(post, aes(x = alpha, y = beta)) + geom_point() + geom_density2d()
```



```
# plot alpha/data
plot(density(theta.burn[,1]),
     main = "P(alpha|data)")
dens = density(theta.burn[,1])
point.est.a = dens$x[dens$y == max(dens$y)]
abline(v = point.est.a)
```

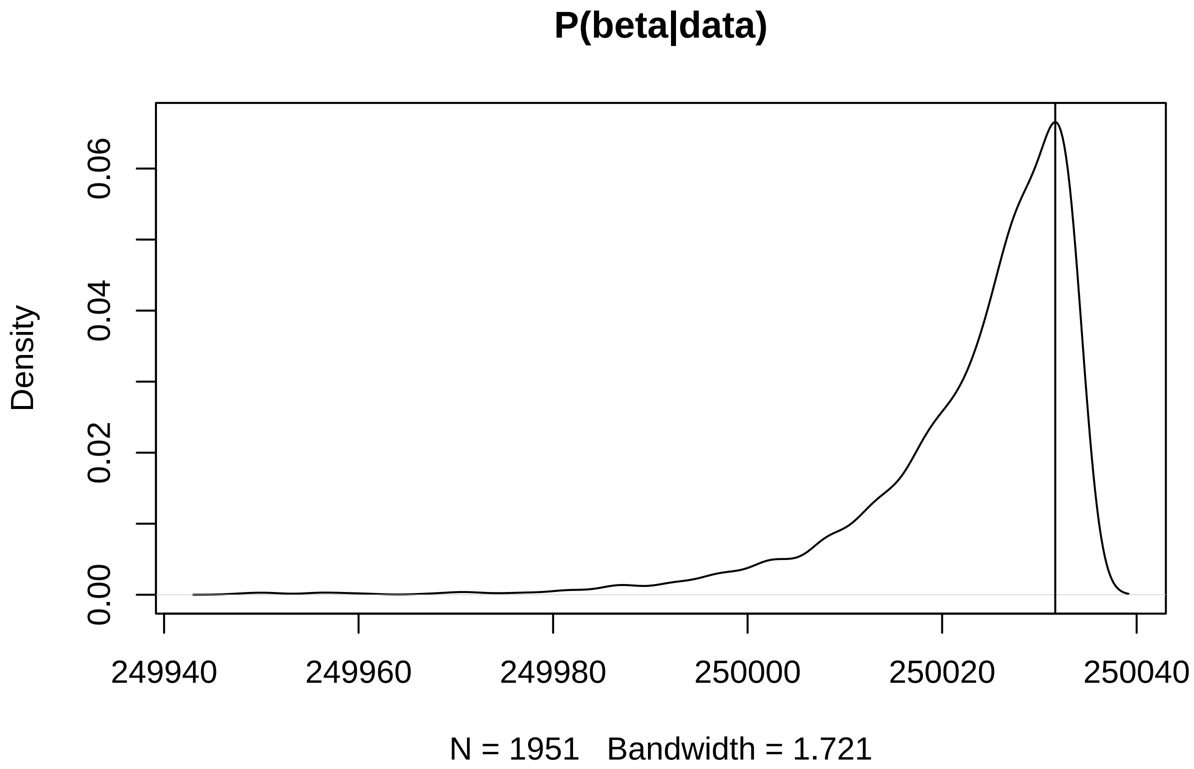


(d)

```
# confidence interval
c(quantile(theta.burn[,1],.025),quantile(theta.burn[,1],.975))
```

```
##     2.5%    97.5%
## 3.420024 3.580153
```

```
# plot beta/data
plot(density(theta.burn[,2]),
     main = "P(beta|data)")
dens = density(theta.burn[,2])
point.est.b = dens$x[dens$y == max(dens$y)]
abline(v = point.est.b)
```



```r
# confidence interval
c(quantile(theta.burn[,2],.025),quantile(theta.burn[,2],.975))
```

```
##     2.5%    97.5%
## 249994.6 250033.7
```

param.	lower	point est	upper
α	3.420024	3.50072	3.580153
β	249994.6	250031.6	250033.7

Note that the β parameter can be interpreted as an average lower end of the income data. Our most dense point for this parameter is 250031.6 and this is very close to our data's actual minimum of 250034. On the note about the