Lab Report in Bayesian Learning

Laboration 2

TDDE07

David Grönberg, Erik Lindgren
Davgr686, Erili040

LINKÖPINGS UNIVERSITET

25-04-2020

Contents

List of Figures

List of Tables

1. Assignments

1.1 Assignment 1

We are given a dataset containing daily average temperatures in Linköping over the year 2018. The dataset consists of two variables: *time* and *temp*, where *time* is the number of days since the new year divided by 365 and *temp* is the recorded average temperature (in Celsius) for a given time. The task is to perform Bayesian analysis of the quadratic regression:

$$temp = \beta_0 + \beta_1 * time + \beta_2 * time^2 + \epsilon, \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

1.1.1 a)

The conjugate prior is used for the linear regression model. We then have the following joint prior for β and σ:

$$\beta|\sigma^2 \sim N(\mu_0, \sigma^2\Omega_0^{-1})$$

$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$$

The hyper-parameters are initiated with $\mu_0 = (-10, 100, -100)$, $\Omega_0 = 0.01 * I_3$, $\nu_0 = 4$ and $\sigma_0^2 = 1$. We draw from the prior and plot the regression curves and modify the hyper-parameters until we think the results reflect our prior beliefs.

Figure 1.1 shows 100 generated regression curves for the hyper-parameters: $\mu_0 = (-10, 130, -130)$, $\Omega_0 = 1 * I_3$, $\nu_0 = 4$ and $\sigma_0^2 = 1$. We argue that these values produce fairly reasonable regression curves, where in the winter time the temperature is around -5 and around 20 in the summer.

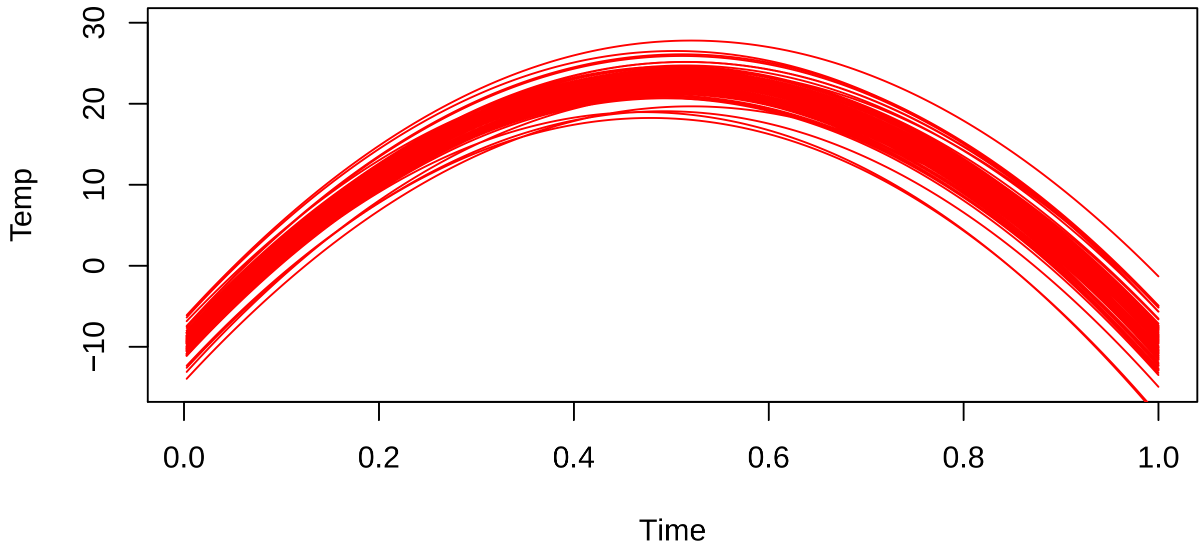

Figure 1.1: Regression curves from 100 prior draws

1.1.2 b)

Next we simulate from the joint posterior distribution of β and σ^2.

$$\beta|\sigma^2 \sim N(\mu_n, \sigma^2\Omega_n^{-1})$$

$$\sigma^2 \sim Inv - \chi^2(\nu_n, \sigma_n^2)$$

where

$$\mu_n = (X'X + \Omega_0)^{-1}(X'X\hat{\beta} + \Omega_0\mu_0)$$
$$\Omega_n = X'X + \Omega_0$$
$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \nu_0\sigma_0^2 + (y'y + \mu_0\Omega_0\mu_0 - \mu_n\Omega_n\mu_n)$$

Figure 1.2 shows a histogram of values for 10,000 draws from the posterior β and σ^2.

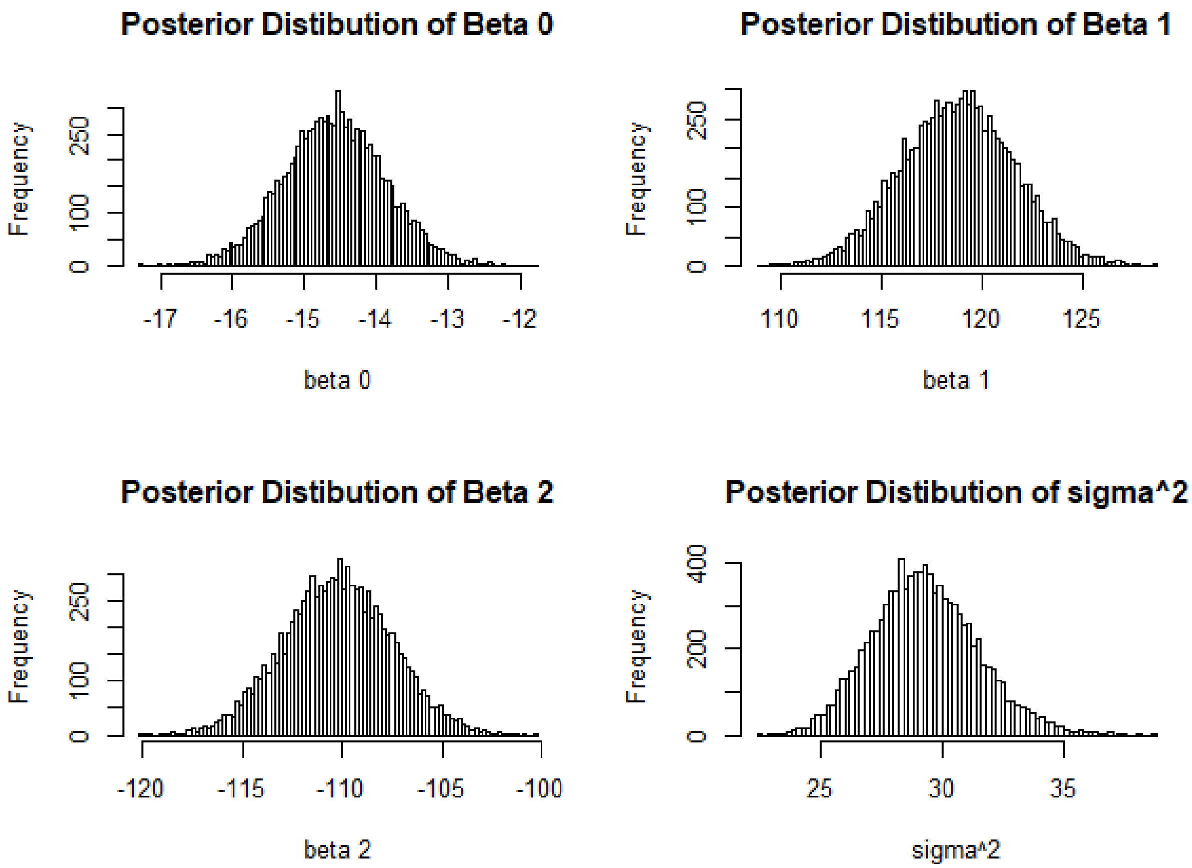

Figure 1.2: Posterior distribution of β and σ^2

Figure 1.3 shows a plot of the temperature data (blue scatter points) and a curve for the posterior median of the regression function $f(time) = \beta_0 + \beta_1 * time + \beta_2 * time^2$ (black line). The lower 2.5% and upper 97.5% credible interval for $f(time)$ is represented by the red and green line, respectively.

As the plot shows, the 95% credible interval for $f(time)$ does not contain most of the data points because the error term: $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ is not included in $f(time)$.

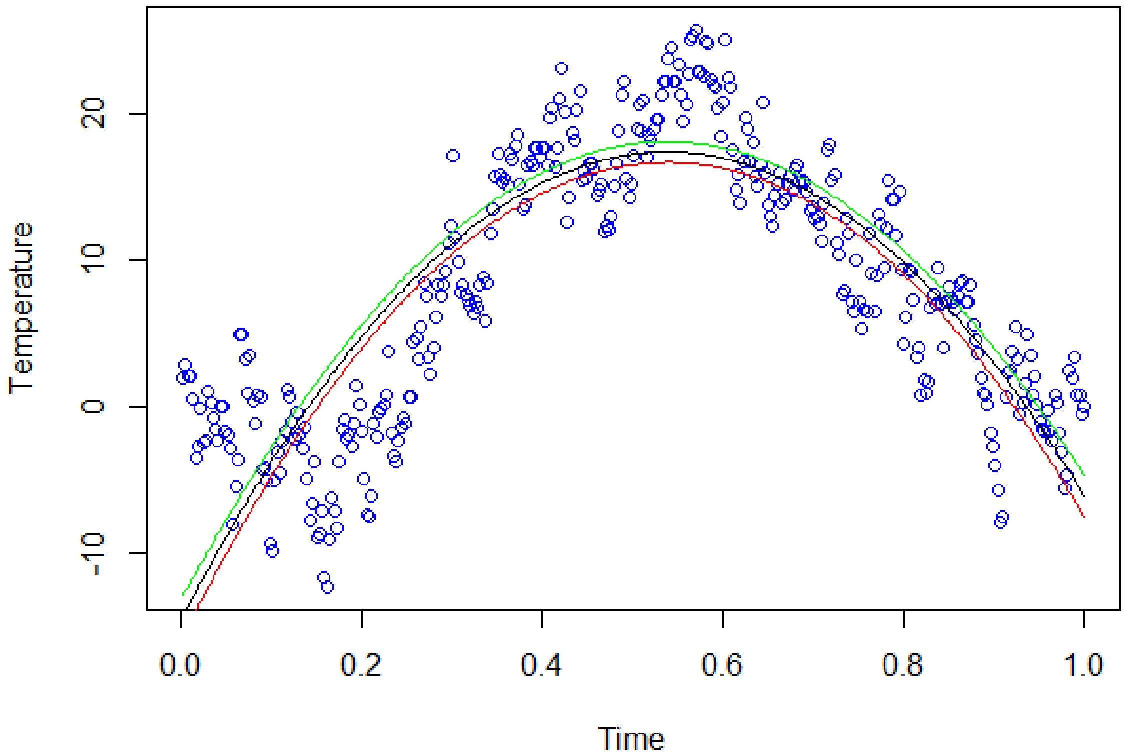

Figure 1.3: Posterior median of $f(time)$

1.1.3 c)

Next, the time with the highest expected temperature, denoted $\hat{x}$ is simulated from the posterior distribution of β.

The highest expected temperature is found where the derivative of the regression function is zero. This yields the following formula for $\hat{x}$

Figure 1.4 shows a histogram of the posterior distribution of $\hat{x}$

$$\hat{x} = -\frac{\beta_1}{2\beta_2}$$

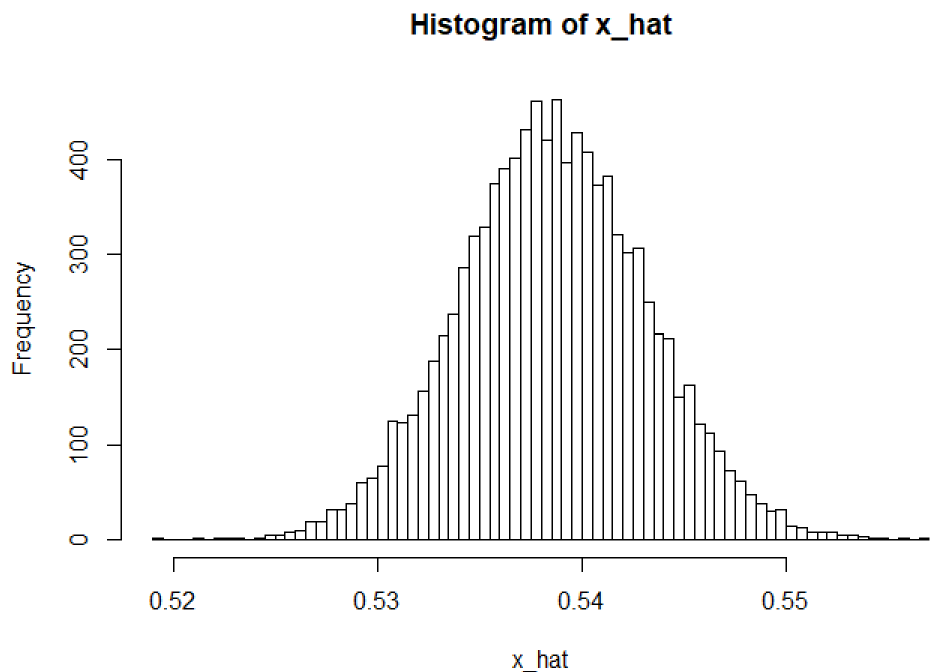

Figure 1.4: Posterior distribution of $\hat{x}$

1.1.4 d)

If we want to estimate a polynomial model of order 7 and suspect that higher order terms may not be needed, we can set μ_0 to low values for higher order terms, centering the β distribution around lower values. We can also set λ (in $\Omega_0 = \lambda I$)to larger values, giving a smoother fit.

1.2 Assignment 2

We are given a dataset containing 200 observations of women with nine different variables. The variables describes the amount of children, years of education, years of working, age, husbands income, and whether or not the women is currently working. The dataset also contains a constant for the intercept.

1.2.1 a)

We are tasked to approximate the posterior distribution of a 8-dim parameter vector β with a multivariate normal distribution:

$$\beta|y, X \sim N(\beta, J_y^{-1}(\beta)$$

where β is the posterior mode and J(β) is the observed Hessian evaluated at the posterior mode. The prior used to find the approximate posterior distribution was:

$$\beta \sim N(0, \tau^2 I)$$

where $\tau = 10$.
The resulting numerical values for β and J(β) can be seen in table 1.1 and figure 1.5.

Table 1.1: Numerical values for the posterior mode (β)

0.6267	-0.0197	0.1802	0.1675	-0.1446	-0.0820	-1.3591	-0.0246

	V1	V2	V3	V4	V5	V6	V7	V8
1	-2.266040090	-3.338901e-03	6.545166e-02	1.179135e-02	-0.0457803996	3.029370e-02	0.1887501416	0.0980249308
2	-0.003338901	-2.528075e-04	5.610290e-04	3.125693e-05	-0.0001414992	3.588591e-05	-0.0005067051	0.0001444227
3	0.065451664	5.610290e-04	-6.218251e-03	3.558101e-04	-0.0018962834	3.243051e-06	0.0061347158	-0.0017527385
4	0.011791353	3.125693e-05	3.558101e-04	-4.351746e-03	0.0142491568	1.340957e-04	0.0014690636	-0.0005437045
5	-0.045780400	-1.414992e-04	-1.896283e-03	1.424916e-02	-0.0555788806	3.299256e-04	-0.0032084847	-0.0005120396
6	0.030293701	3.588591e-05	3.243051e-06	1.340957e-04	0.0003299256	-7.184679e-04	-0.0051842393	-0.0010953068
7	0.188750142	-5.067051e-04	6.134716e-03	1.469064e-03	-0.0032084847	-5.184239e-03	-0.1512642569	-0.0067690578
8	0.098024931	1.444227e-04	-1.752739e-03	-5.437045e-04	-0.0005120396	-1.095307e-03	-0.0067690578	-0.0199724321

Figure 1.5: Numerical values for the observed Hessian(J(β)) evaluated at the posterior mode.

We are then tasked to compute the 95% credible interval for the β of the NSmallChild variable. In 1.6 we can see the posterior distribution of the β variable. The 95% credible interval for the β variable is between -2.1343751 and -0.5752881.

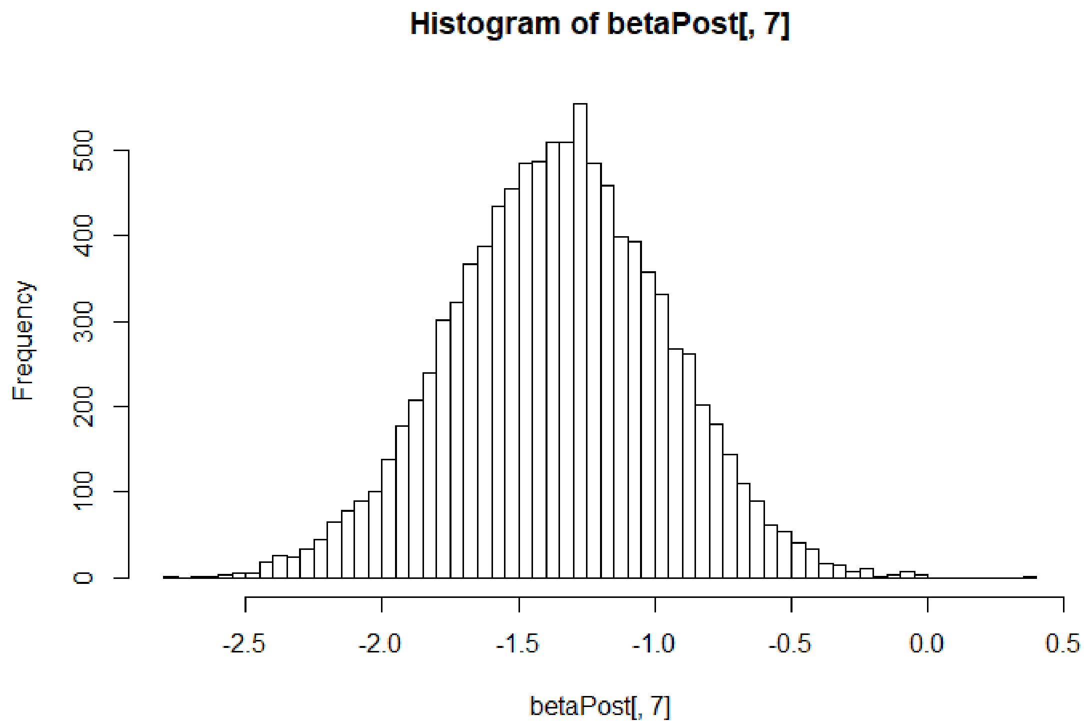

Figure 1.6: The posterior distribution of the beta for the NSmallChild variable

The β for NSmallChild variable has the largest posterior mean of all the features. We can therefore say that the NsmallChild has an important role in predicting whether or not a woman work.

1.2.2 b)

The next task is to plot the predictive distribution of the work variable for a specific woman. The woman is 40 years old, has two children(3 and 9 years old), 8 years of education, 10 years of education and her husband has an income of 10. The predictive distribution follows the following logistical regression function:

$$Pr(y = 1|x) = \frac{exp(x^T\beta)}{1 + exp(x^T\beta)}$$

The resulting plot of the predictive distribution can be seen in 1.7.

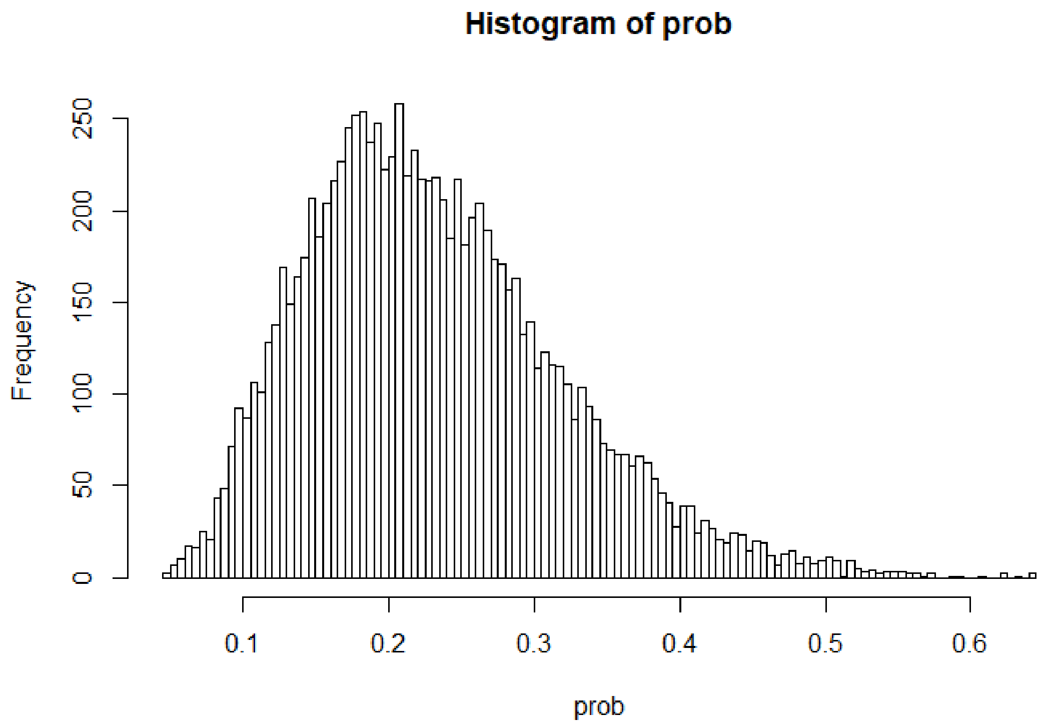

Figure 1.7: The predictive distribution for the work variable.

1.2.3 c)

Finally we are tasked to consider 10 women with the same features as the woman in 2(b). The task is to to plot the predictive distribution for the number of women that are working. As we have a sum of Bernoulli random variables, the distribution can be described as a binomial distribution. The resulting plot can be seen in 1.8.

Figure 1.8: Predictive distribution for the number of women that are working.

2. Code Appendix

2.1 Assignment 1

```r
# a

library(mvtnorm)
set.seed(1234)

data <- read.delim("TempLinkoping.txt", sep='\t')
data$temp <- ifelse(!is.na(data$X), data$X, data$temp)
data <- subset(data, select = -c(2))


mu_0 <- c(-10, 130, -130)
omega_0 <- 1 * diag(3)
var_0 <- 1
v_0 <- 4

predict_sim <- function(v_0, var_0, omega_0, x_vals) {

  var <- (v_0*var_0)/rchisq(1,v_0)
  betas <- rmvnorm(1, mean = mu_0, sigma = var * solve(omega_0))

  preds <- array(1:length(x_vals))
  for (i in 1:length(x_vals)) {
    preds[i] <- betas[,1] + betas[,2] * x_vals[i] + betas[,3] * x_vals[i]**2
  }
  return (preds)
}

plot(data$time, data$temp)
for (i in 1:100) {
  lines(data$time, predict_sim(v_0, var_0, omega_0, data$time), type="l", col="red")
}

# b
X <- as.matrix(cbind(1, data$time, data$time**2))
y <- as.matrix(data$temp)

beta_hat <- solve(t(X)%*%X)%*%(t(X)%*%y)
mu_n <- solve(t(X)%*%X + omega_0) %*% (t(X)%*%X%*%beta_hat + omega_0%*%mu_0)
omega_n <- t(X)%*%X + omega_0
v_n <- v_0 + length(data$time)
var_n <- (v_0 * var_0 + (t(y)%*%y + t(mu_0)%*%omega_0%*%mu_0 - t(mu_n)%*%omega_n%*%mu_n))/v_n

posterior_draw <- function(v_n, var_n, mu_n, omega_n) {

  var <- (v_n*var_n)/rchisq(1,v_n)
  betas <- rmvnorm(1, mean = mu_n, sigma = var[1] * solve(omega_n))
```

```r
  list(betas=betas, var = var)
}

n_sim <- 10000
posterior_betas <- matrix(0,n_sim,3)
posterior_vars <- c()
median_regression <- matrix(0,n_sim,length(data$temp))

for (i in 1:n_sim) {
  posterior_sim <- posterior_draw(v_n, var_n, mu_n, omega_n )
  posterior_betas[i,] <- posterior_sim$betas
  posterior_vars[i] <- posterior_sim$var

  for (x in 1:length(data$temp)) {
    median_regression[i,x] <- posterior_betas[i,1] + posterior_betas[i,2] * data$time[x] +
    posterior_betas[i,3] * data$time[x]**2
  }
}

medians <- apply(median_regression, 2, median)
cred_interval_lower <- c()
cred_interval_upper <- c()

for (u in 1:length(data$temp)) {
  cred_interval <- quantile(median_regression[,u],probs=c(0.025,0.975))
  cred_interval_lower[u] <- cred_interval[1]
  cred_interval_upper[u] <- cred_interval[2]
}

plot(data$time, data$temp, col="blue", xlab="Time", ylab="Temperature")
lines(data$time, medians, col="black")
lines(data$time, cred_interval_lower, col="red")
lines(data$time, cred_interval_upper, col="green")

par(mfrow=c(2,2))
hist(posterior_betas[,1], 100, main="Posterior Distibution of Beta 0", xlab="beta 0")
hist(posterior_betas[,2], 100, main="Posterior Distibution of Beta 1", xlab="beta 1")
hist(posterior_betas[,3], 100, main="Posterior Distibution of Beta 2", xlab="beta 2")
hist(posterior_vars, 100, main="Posterior Distibution of sigma^2", xlab="sigma^2")
par(mfrow=c(1,1))

# c
x_hat <- -posterior_betas[,2] / (2*posterior_betas[,3])
hist(x_hat, 100)
```

2.2 Assignment 2

```r
library("mvtnorm")
df = read.table("WomenWork.dat", header = TRUE)

#a)

tau = 10
mu = as.vector(rep(0,8))
sigma = tau^2*diag(8)
y = as.vector(df[,1])
X <- as.matrix(df[,2:9])

LogPostLogistic <- function(beta,y,X,mu,Sigma){
```

```
  n <- length(beta);
  linPred <- X%*%beta;

  logLik <- sum( linPred*y -log(1 + exp(linPred)));
  if (abs(logLik) == Inf) logLik = -20000
  logPrior <- dmvnorm(beta, matrix(0,n,1), Sigma, log=TRUE)
  return(logLik + logPrior)
}

initVal <- as.vector(solve(crossprod(X,X))%*%t(X)%*%y)
Results<-optim(initVal,LogPostLogistic,gr=NULL,y,X,mu,sigma,method=c("BFGS"),control=list(
    fnscale=-1),hessian=TRUE)

BTilde = Results$par
BHess = solve(Results$hessian)

betaPost <- rmvnorm(10000, BTilde, -BHess)
hist(betaPost[,7],100)
mean(betaPost[,7])

NrSmallChild = betaPost[,7]
quantils = quantile(NrSmallChild, c(0.025,0.975))

glmModel <- glm(Work~ 0 + ., data = df, family = binomial)

summary(glmModel)
quantils

#b)

xCase = as.matrix(t(c(1,10,8,10,1,40,1,1)))
xCase
prob = c()

betas = rmvnorm(10000, BTilde, -BHess)
prob = exp(xCase%*%t(betas))/(1+exp(xCase%*%t(betas)))
hist(prob,100)

#c)
workingWomen = rbinom(10000,10,prob)
hist(workingWomen,1000)
```