Bayesian Learning lab2-revised

Joris van Doorn // Weng Hang Wong

5/18/2020

1. Linear and polynomial regression

The dataset TempLinkoping.txt contains daily average temperatures (in Celcius degrees) at Malmslätt, Linköping over the course of the year 2018. The response variable is temp and the covariate is

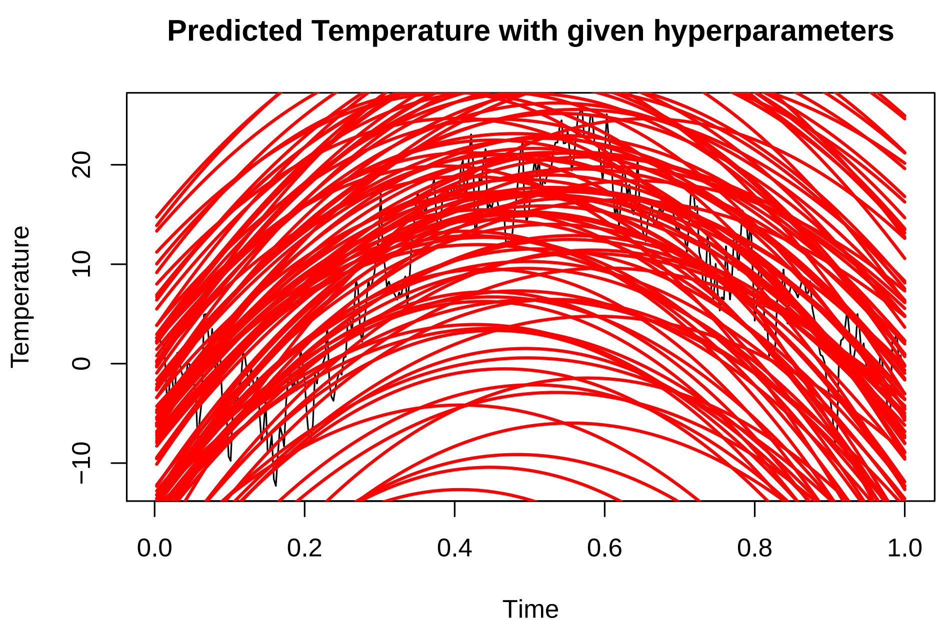
$$time = \frac{the\ number\ of\ days\ since\ beginning\ of\ year}{365}$$

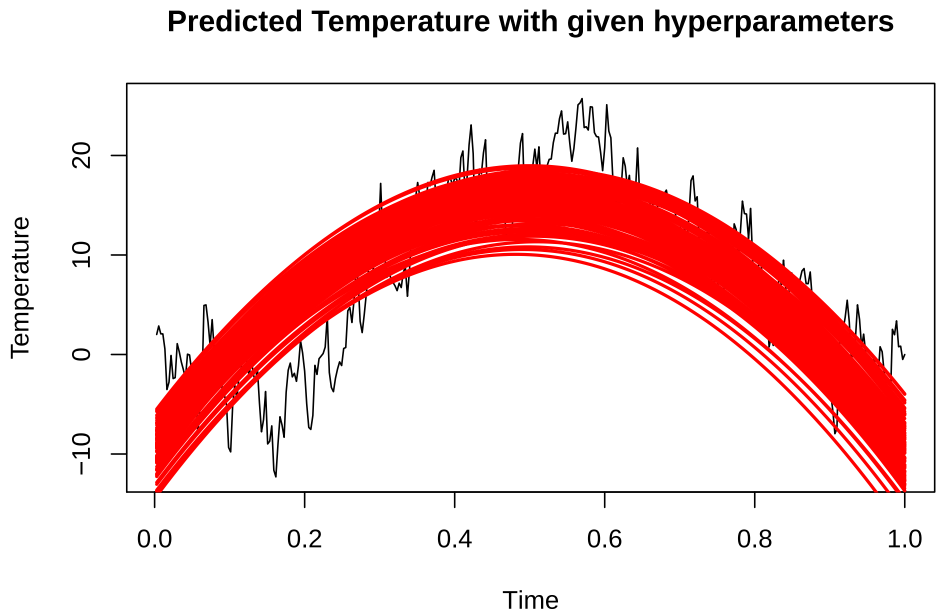
The task is to perform a Bayesian analysis of a quadratic regression

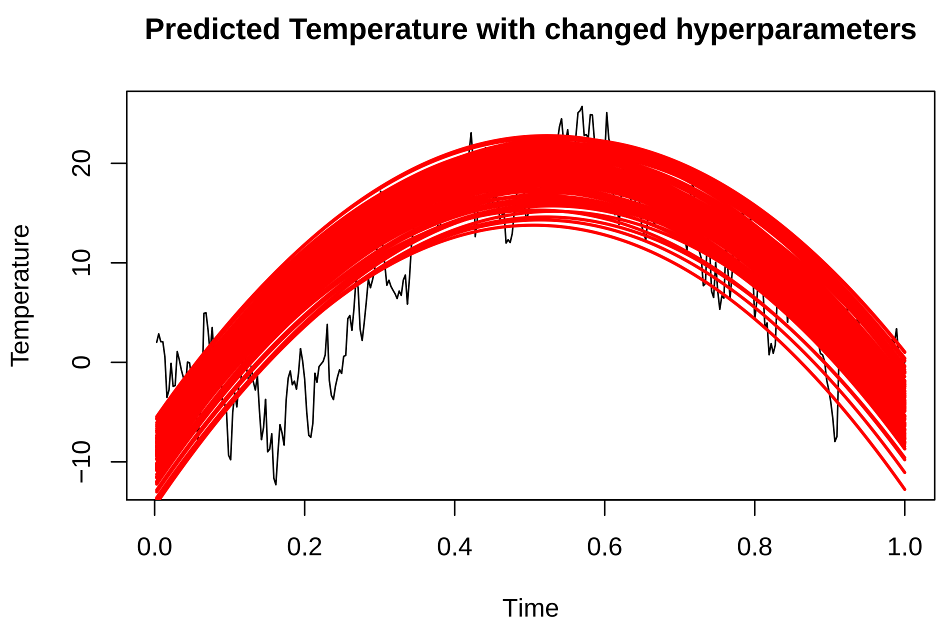
$$temp = \beta_0 + \beta_1 * time + \beta_2 * time^2 + \epsilon, \epsilon \sim^{iid} N(0, \sigma^2)$$

a.

Determining the prior distribution of the model parameters. Use the conjugate prior for the linear regression model. Your task is to set the prior hyperparameters μ_0, Ω_0, ν_0 and σ_0^2 to sensible values. Start with $\mu_0 = (-10, 100, -100)^T, \Omega_0 = 0.01 \cdot I_3, \nu_0 = 4$ and σ_0^2. Check if this prior agrees with your prior opinions by simulating draws from the joint prior of all parameters and for every draw compute the regression curve. This gives a collection of regression curves, one for each draw from the prior. Do the collection of curves look reasonable? If not, change the prior hyperparameters until the collection of prior regression curves agrees with your prior beliefs about the regression curve.



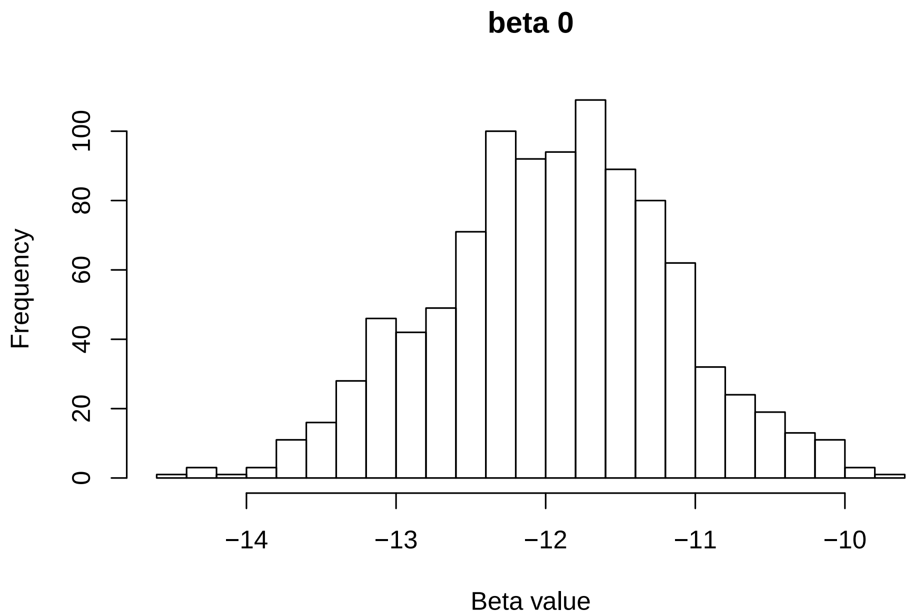


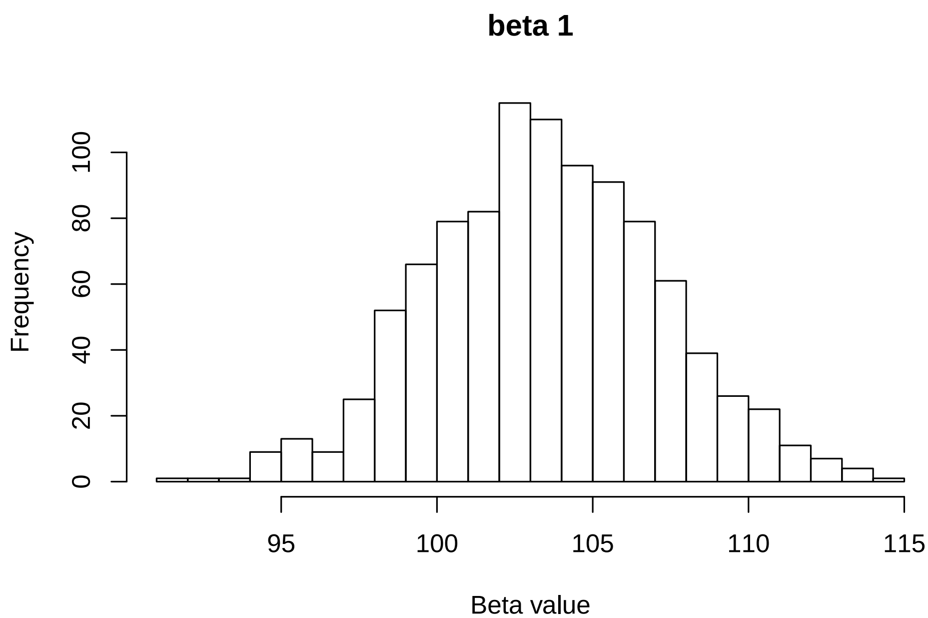


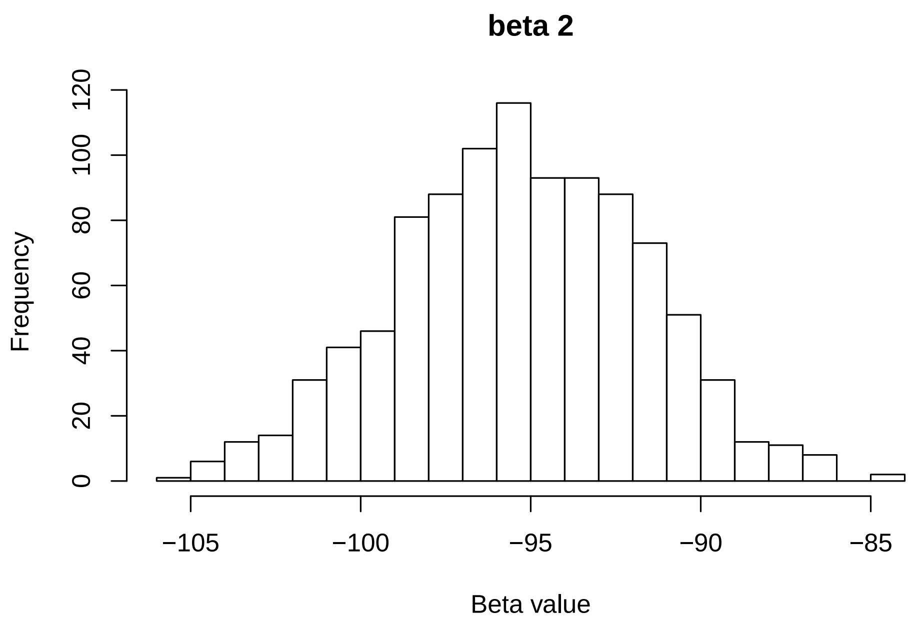
b.

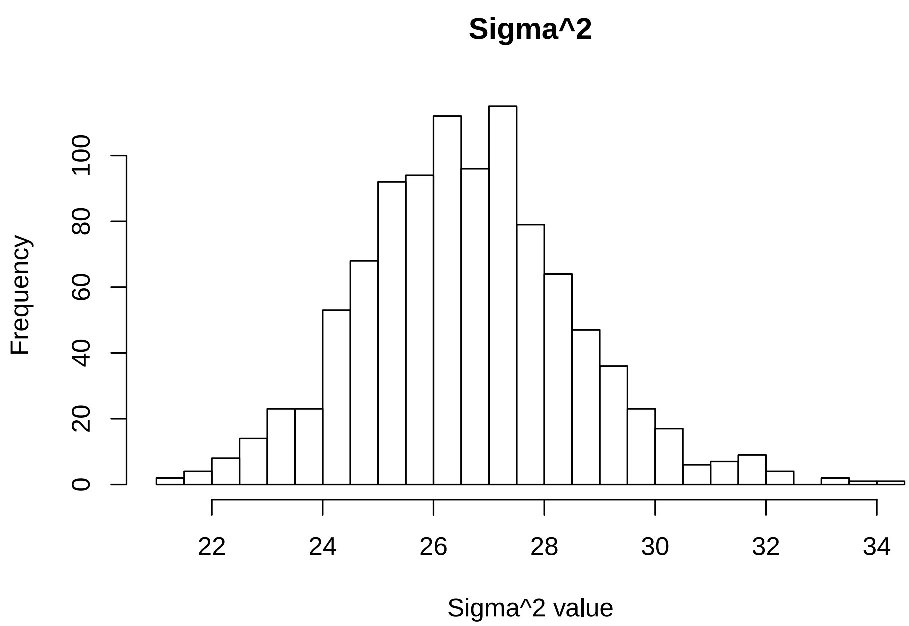
Write a program that simulates from the joint posterior distribution of β_0,β_1,β_2,*and* σ^2. *Plot the marginal posteriors for each parameter as a histogram. Also produce another figure with a scatter plot of the temperature data and overlay a curve for the posterior median of the regression function* $f(time) = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2$, *computed for every value of time. Also overlay curves for the lower 2.5% and upper 97.5% posterior credible interval for f (time). That is, compute the 95% equal tail posterior probability intervals for every value of* *time and then connect the lower and upper limits of the interval by curves. Does the interval bands contain most of the data points? Should they?*

From the graph below, the parameters are simulated from the joint posterior distribution. The marginal posteriors for each parameter β_0,β_1,β_2,and σ^2 are shown below.

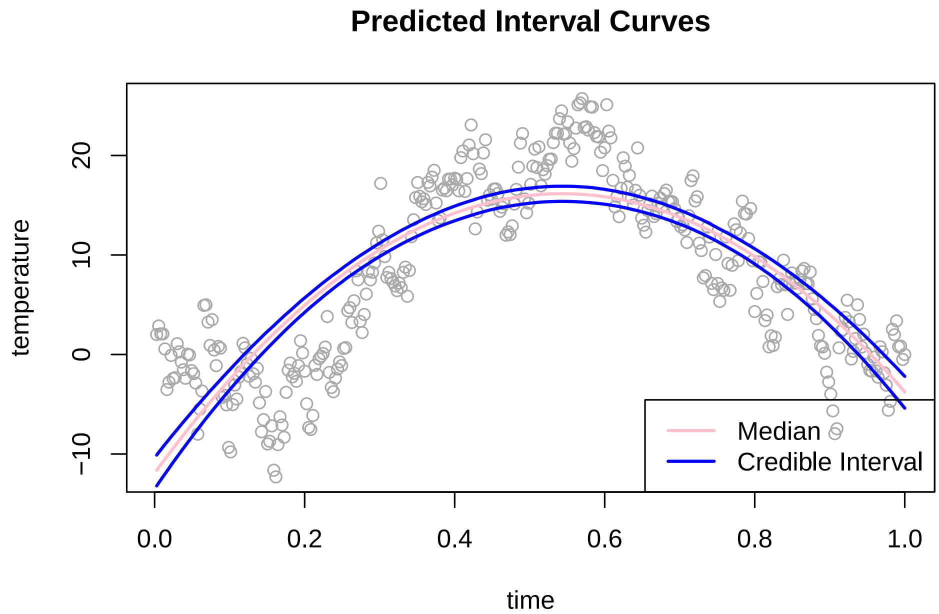








Here is a scatter plot of the temperature data with the median and credible interval curves. However, most of the data points are not contained in the 95% posterior credible interval, they should not contained most of the data points, since it didn't include the ε in the regression function and the uncentainty parameter here has particular probability.



c.

It is of interest to locate the time with the highest expected temperature (that is, the time where f (time) is maximal). Let's call this value $\widetilde{x}$ Use the simulations in b) to simulate from the posterior distribution of $\widetilde{x}$

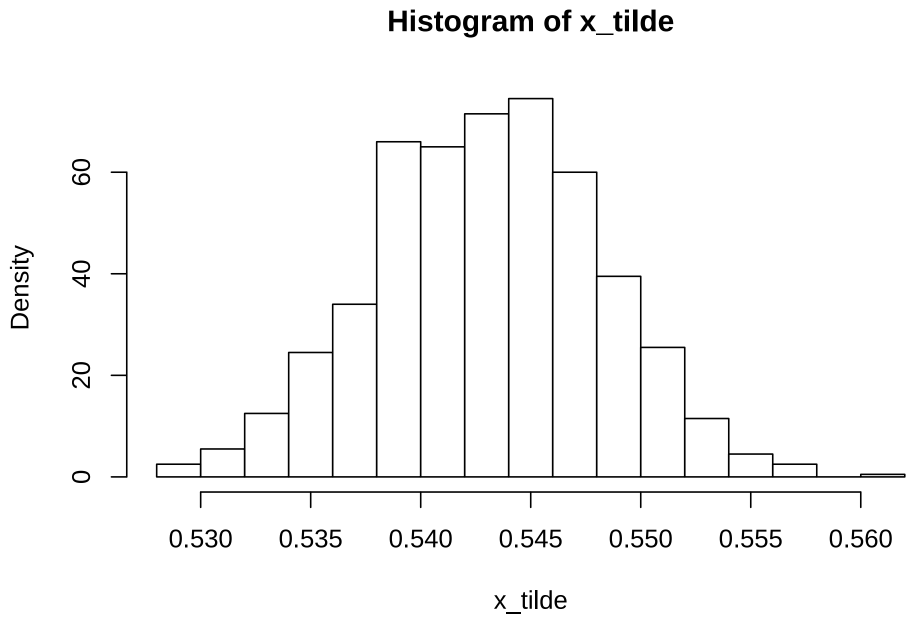
The first derivative of f(time) will be maximal when it equal to zero.

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$$

$$0 = \beta_1 + 2\beta_2 x$$

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

```
## The expected highest expected temperature is 0.5430181
```



d.

Say now that you want to estimate a polynomial model of order 7, but you suspect that higher order terms may not be needed, and you worry about overfitting. Suggest a suitable prior that mitigates this potential problem. You do not need to compute the posterior, just write down your prior. [Hint: the task is to specify μ_0 and Ω_0 in a smart way.]

To prevent overfitting we suggest adding a regularization term. The proposed prior would like as follows:

$$\beta_i|\sigma^2 \sim^{iid} N(0, \frac{\sigma^2}{\lambda})$$

where λ will be the smoothness/shrinkage/regularization term. Ω_0 and λ are relate as $\Omega_0 = \lambda I$. A change in λ does not directly affect μ_0. However, it does influence μ_n through Ω_0. The larger λ is, the more shrinkage.

2. Posterior approximation for cassification with logistic regression

The dataset WomenWork.dat contains n = 200 observations (i.e. women) on the following nine variables:

a.

Consider the logistic regression

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

where y is the binary variable with y = 1 if the woman works and y = 0 if she does not. x is a 8-dimensional vector containing the eight features (including a one for the constant term that models the intercept). The

Variable	Data type	Meaning	Role
Work	Binary	Whether or not the woman works	Response
Constant	1	Constant to the intercept	Feature
HusbandInc	Numeric	Husband's income	Feature
EducYears	Counts	Years of education	Feature
ExpYears	Counts	Years of experience	Feature
ExpYears2	Numeric	(Years of experience)/10)^2	Feature
Age	Counts	Age	Feature
NSmallChild	Counts	Number of child <7 years in household	Feature
NBigChild	Counts	Number of child >6 years in household	Feature

goal is to approximate the posterior distribution of the 8-dim parameter vector β with a multivariate normal distribution

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

where $\hat{\beta}$ is the posterior mode and $J(\hat{\beta}) = -\frac{\delta^2 lnp(\beta|y)}{\delta\beta\delta\beta^T}|_{\beta=\hat{\beta}}$ is the observed Hesian evaluated at the posterior mode. Note that $J(\hat{\beta})$ is an 8x8 matrix with second derivatives on the diagonal and cross-derivatives $\frac{\delta^2 lnp(\beta|y)}{\delta\beta_i\delta\beta_j}$ on the offdiagonal. It is actually not hard to compute this derivative by hand, but don't worry, we will let the computer do it numerically for you. Now, both $\hat{\beta}$ and $J(\hat{\beta})$ are computed by the optim function in R. I want you to implement you own version of this. You can use my code as a template, but I want you to write your own file so that you understand every line of your code. Don't just copy my code. Use the prior $\beta \sim N(0, \tau^2 I)$, with $\tau = 10$. Your report should include your code as well as numerical values for $\hat{\beta}$ and $J(\hat{\beta})$ for the WomenWork data.

Compute an approximate 95% credible interval for the variable NSmallChild. Would you say that this feature is an important determinant of the probability that a women works?

	Constant	HusbandInc	EducYears	ExpYears	ExpYears2	Age	NSmallChild	NBigChild
Constant	2.2660245	0.0033388	-0.0654508	-0.0117911	0.0457795	-0.0302936	-0.1887509	-0.0980243
HusbandInc	0.0033388	0.0002528	-0.0005610	-0.0000313	0.0001415	-0.0000359	0.0005067	-0.0001444
EducYears	-0.0654508	-0.0005610	0.0062182	-0.0003558	0.0018963	-0.0000032	-0.0061347	0.0017527
ExpYears	-0.0117911	-0.0000313	-0.0003558	0.0043516	-0.0142487	-0.0001341	-0.0014689	0.0005437
ExpYears2	0.0457795	0.0001415	0.0018963	-0.0142487	0.0555768	-0.0003299	0.0032081	0.0005120
Age	-0.0302936	-0.0000359	-0.0000032	-0.0001341	-0.0003299	0.0007185	0.0051842	0.0010953
NSmallChild	-0.1887509	0.0005067	-0.0061347	-0.0014689	0.0032081	0.0051842	0.1512632	0.0067691
NBigChild	-0.0980243	-0.0001444	0.0017527	0.0005437	0.0005120	0.0010953	0.0067691	0.0199723

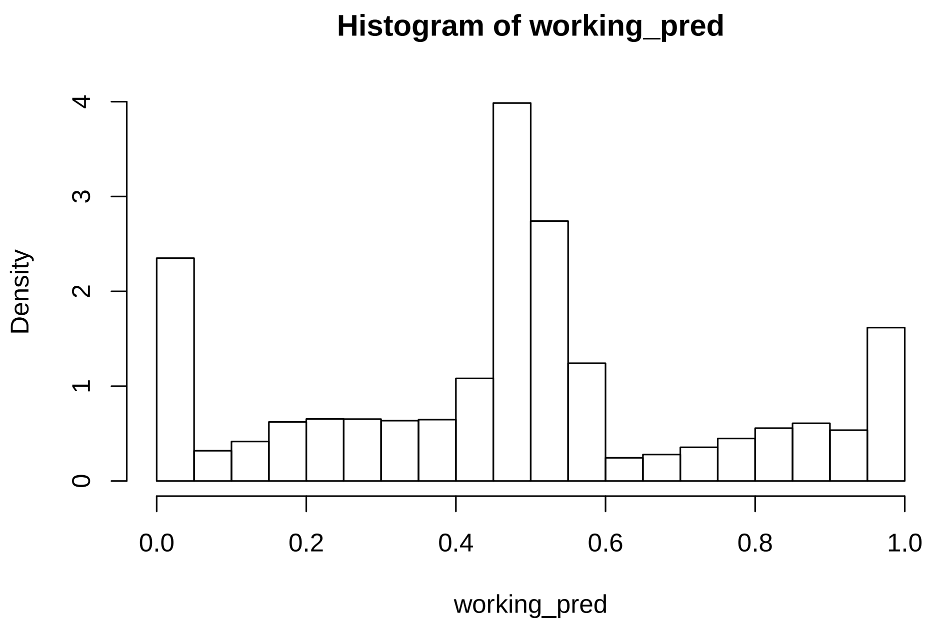
	Verification	Beta_hat	Beta_std
Constant	0.6443036	0.6267765	1.5053320
HusbandInc	-0.0197746	-0.0197917	0.0158998
EducYears	0.1798806	0.1802230	0.0788555
ExpYears	0.1675127	0.1675676	0.0659669
ExpYears2	-0.1443595	-0.1446093	0.2357472
Age	-0.0823403	-0.0820669	0.0268042
NSmallChild	-1.3625024	-1.3591545	0.3889257
NBigChild	-0.0254299	-0.0246916	0.1413234

NSmallChild
-2.1141903
-0.5896295

The number of small children seem to matter. The coefficient is by far the largest, especial compared to the number of larger children a woman has. This would also intuitivaly make sense, because the earlier years of a childs life it demands more attention and it would therefore be more likely for one of the parents to remain at home and not have a job.

b.

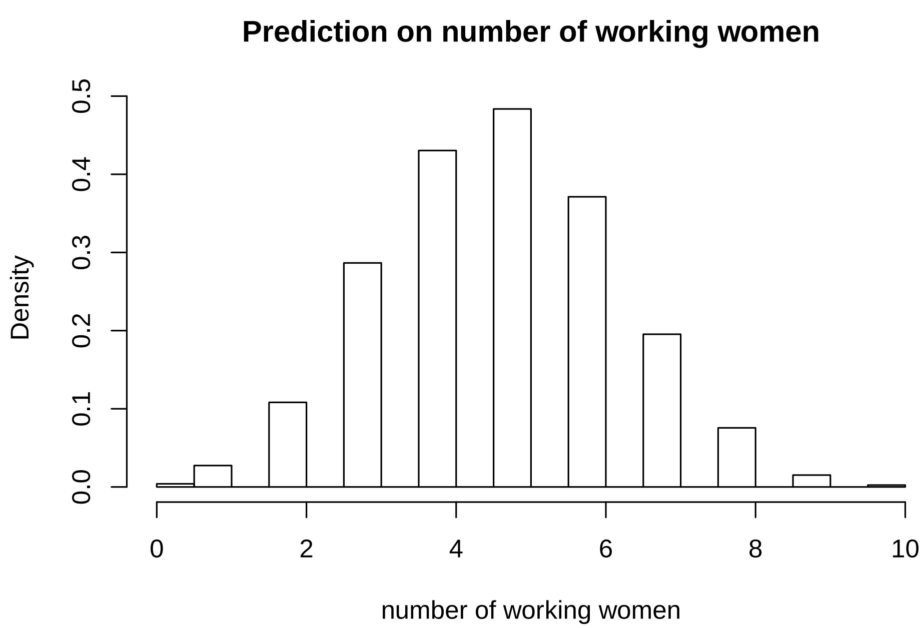
Write a function that simulates from the predictive distribution of the response variable in a logistic regression. Use your normal approximation from 2(a). Use that function to simulate and plot the predictive distribution for the Work variable for a 40 year old woman, with two children (3 and 9 years old), 8 years of education, 10 years of experience. and a husband with an income of 10. [Hints: The R package mvtnorm will again be handy. Remember my discussion on how Bayesian prediction can be done by simulation.]




```
## The expected value for this woman is:  0.4775278 , thus the model predicts that she is working.
```

2c.

Now, consider 10 women which all have the same features as the woman in 2(b). Rewrite your function and plot the predictive distribution for the number of women, out of these 10, that are working. [Hint: Which distribution can be described as a sum of Bernoulli random variables?]



Now, we have n=10 random variables and the probability of working(success) = 0.4775278, then we use binomial distribution to calcution the probability of working.

Appendix

```r
#----------------------------------------------------------------------------
library(mvtnorm)
# Q1a.
data0 <- read.table("TempLinkoping.txt", header = TRUE)
intercept <- rep(1,365)
data1 <- cbind(data0, "intercept"=intercept)
time2 <- data1$time^2
data1 <- cbind(data1, "time2"=time2)

#given hyperparameters
mu0=matrix(c(-10,100,-100))
omega0=diag(x=0.01, nrow=3, ncol=3)
v0=4
sigma20=1

#prior
PriorReg = function(mu0,omega0,v0,sigma20){
  set.seed(12345)
  for(i in 1:100){
    #using chi_sq to sample sigma^2
    chi_sample = rchisq(n=1, df=v0)
    sigma2 = v0*sigma20/chi_sample

    #using mvtnorm sample beta
    beta = rmvnorm(n=1, mean=mu0, sigma=sigma20*solve(omega0))

    #quadratic regression
    quad_regre= beta[1]+beta[2]*data0$time+beta[3]*(data0$time^2)+rnorm(1,mean=0, sd=sqrt(sigma2))
    lines(x=data0$time, y=quad_regre,col="red",lwd=2)
  }
}

### Check the given hyperpara
plot(data0, main="Predicted Temperature with given hyperparameters", ylab="Temperature", xlab="Time", t
PriorReg( mu0, omega0, v0, sigma20)

### change the hyperpara nu
plot(data0, main="Predicted Temperature with given hyperparameters", ylab="Temperature", xlab="Time", t
PriorReg( mu0, omega0, v0, sigma20=0.03)


# Change the hyperpara sigma
plot(data0, main="Predicted Temperature with changed hyperparameters", ylab="Temperature", xlab="Time",
PriorReg( mu0=matrix(c(-10,110,-105)), omega0, v0, sigma20=0.03)
#---------------------------------------------
#Q1b.

### find beta hat
n=dim(data0)[1]
X = data.frame(intercept=rep(1,n), x1=data0$time, x2=data0$time^2)
X = as.matrix(X)
```

```r
y = data0$temp
betaHat = solve(t(X)%*%X)%*%t(X)%*%y

### calculate mu, omega, nu sigma
mu_n = solve(t(X)%*%X+omega0) %*% (t(X)%*%X%*%betaHat+omega0%*%mu0)
omega_n = t(X)%*%X+omega0
v_n = v0 + n
sigma2_n = (v0*sigma20+(t(y)%*%y+t(mu0)%*%omega0%*%mu0-t(mu_n)%*%omega_n%*%mu_n))/v_n

### Marginal posterior
set.seed(12345)
paras = NULL
final = NULL
for(i in 1:1000){
  #using chi_sq to sample posterior sigma^2
  chi_sample = rchisq(n=1, df=v_n)
  post_sigma2 = v_n*sigma2_n/chi_sample

  #using mvtnorm sample posterior beta
  post_beta = rmvnorm(n=1, mean=mu_n, sigma=post_sigma2[1]*solve(omega_n))

  paras = cbind(post_beta,post_sigma2)
  final = rbind(paras, final)
}

colnames(final) = c("beta0","beta1","beta2","sigma2")

## histogram of each parameters
hist(final[,1], main="beta 0", xlab="Beta value", breaks=30)
hist(final[,2], main="beta 1", xlab="Beta value", breaks=30)
hist(final[,3], main="beta 2", xlab="Beta value", breaks=30)
hist(final[,4], main="Sigma^2",xlab="Sigma^2 value", breaks=30)
### median curve and intervals
post_beta = final[,1:3]

PredictedVal=matrix(0,nrow=n,ncol=nrow(post_beta))
for(i in 1:nrow(post_beta)){
  PredictedVal[,i] = X %*% post_beta[i,]
}

## find median and credible interval
medianInterval=c()
crediInterval = matrix(0,nrow=n,ncol=2)
for(i in 1:n){
  medianInterval[i] = median(PredictedVal[i,])
  crediInterval[i,] = quantile(PredictedVal[i,], c(0.025,0.975))

}

plot(data0, main="Predicted Interval Curves", col="darkgrey", ylab="temperature")
lines(data0$time,medianInterval, col="pink",lwd=2)
lines(data0$time,crediInterval[,1], col="blue",lwd=2)
lines(data0$time,crediInterval[,2], col="blue",lwd=2)
```

```
legend("bottomright",legend=c("Median", "Credible Interval"), col=c("pink","blue"),lwd=2 )

#--------------------------------------------------------------
#Q1c.

x_tilde = -post_beta[,2]/ (2*post_beta[,3])
cat("The expected highest expected temperature is",mean(x_tilde))

hist(x_tilde, freq=F, breaks=20)

# -------------------------
# Q2a.
library(knitr)
# loading data
data0 <- read.table("WomenWork.dat",header = T)

# setting initial values
tau <- 10
y <- data0[,1]
X <- as.matrix(data0[,2:9])
nCov <- dim(X)[2]
covNames <- names(data0)[2:9]

# Prior
mu <- as.vector(rep(0,nCov))
sigma <- tau^2*diag(nCov)

set.seed(12345)
# Logistic regression function that returns the regression coefficients
logiPost <- function(betas,y,X,sigma){
  pred <- as.vector(X%*%betas)
  loglike <- sum(y*pred-log(1+exp(pred)))
  logprior <- dmvnorm(betas, mean=rep(0,length(betas)), sigma, log=T)
  return(loglike+logprior)
}

# setting initial values
initVal <- as.vector(rnorm(dim(X)[2]))
# optimize over the betas
optRes <- optim(initVal,logiPost,gr=NULL,y,X,sigma,method="BFGS",control=list(fnscale=-1),hessian=T)

# retrieving betas
beta_hat <- optRes$par
beta_hes <- solve(-optRes$hessian)
beta_std <- as.matrix(sqrt(diag(beta_hes)))

# verifing results
model0 <- glm(Work~0+., data=data0, family=binomial)

# printing results
colnames(beta_hes) <- covNames
rownames(beta_hes) <- covNames
kable(beta_hes)
```

```
kable(data.frame(Verification=model0$coefficients,Beta_hat=beta_hat,Beta_std=beta_std))


#set.seed(12345)
#Small_beta_hat <- rmvnorm(n=1000, mean=beta_hat, sigma = beta_hes)
#CI_NSmallChild <- c(qnorm(0.025,mean=mean(Small_beta_hat[,7]),sd=beta_std[7]),qnorm(0.975,mean=mean(Sm

## find the CI for NSmallChild by simulating from the Post
set.seed(12345)
Post_beta_hat = rmvnorm(n=1000, mean=beta_hat, sigma = beta_hes)

CI_NSmallChild <- c(qnorm(0.025,mean=mean(Post_beta_hat[,7]),sd=beta_std[7]),qnorm(0.975,mean=mean(Post_

kable(data.frame(NSmallChild=CI_NSmallChild))
#-----------------------------------
# Q2b.

woman <- c(constant=1,husbandIC=10,educYear=8,expYear=10,expYear2=1,age=40,NSmallChild=1,NBigChild=1)

posterior_betas <- rmvnorm(10000,mean=beta_hat,sigma=beta_hes)

predict_logistic <- function(betas, X){
  yNew <- (exp(X%*%betas)) / (1+exp(X%*%betas))
  return(yNew)
}

woman=as.matrix(woman,nrow=1)

#predict_logistic(posterior_betas, woman) in rows
working_pred <- apply(posterior_betas,1,predict_logistic,woman)

hist(working_pred,freq = F)

cat("The expected value for this woman is: ", mean(working_pred), ", thus the model predicts that she i
#-----------------------------------
# Q2c.

pred = mean(working_pred) #the Probability of working
pred_bino = rbinom(10000,size=10, prob=pred) #10 women
hist(pred_bino,main="Prediction on number of working women", xlab="number of working women",freq=F)
```