TDDE07 - Lab 2

Pontus Svensson (ponsv690) & Kristian Sikiric (krisi211)

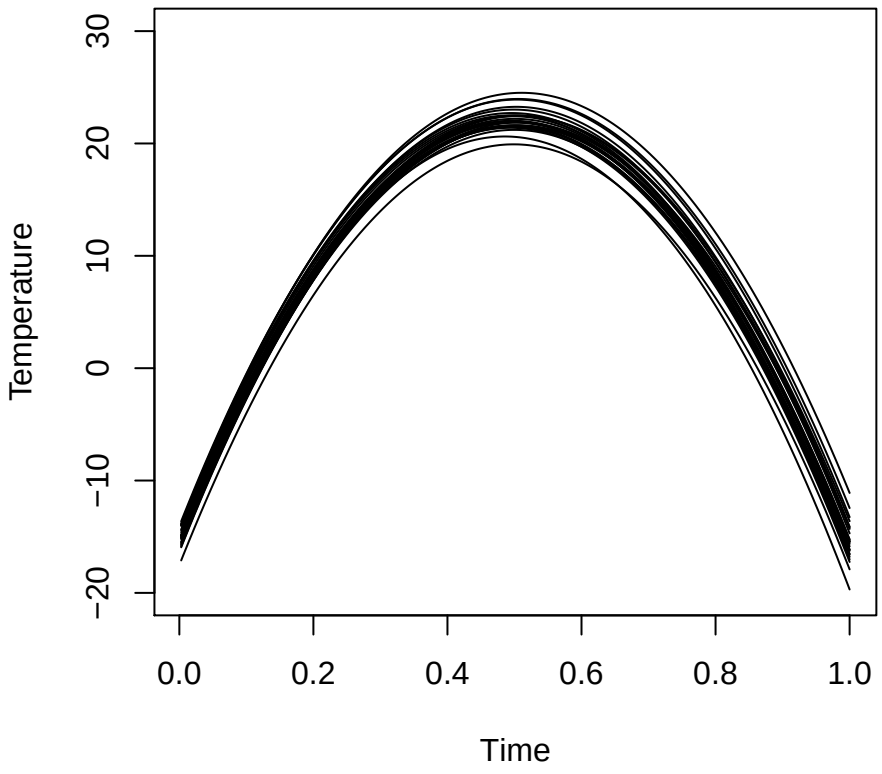
Assignment 1

Linear and polynomial regression

In this assignent a data set of temperatures in Malmslätt, Linköping was given. The task was to perfrom a Bayesian analysis of a quadratic regression $temp = \beta_0 + \beta_1 * time + \beta_2 * time^2$.

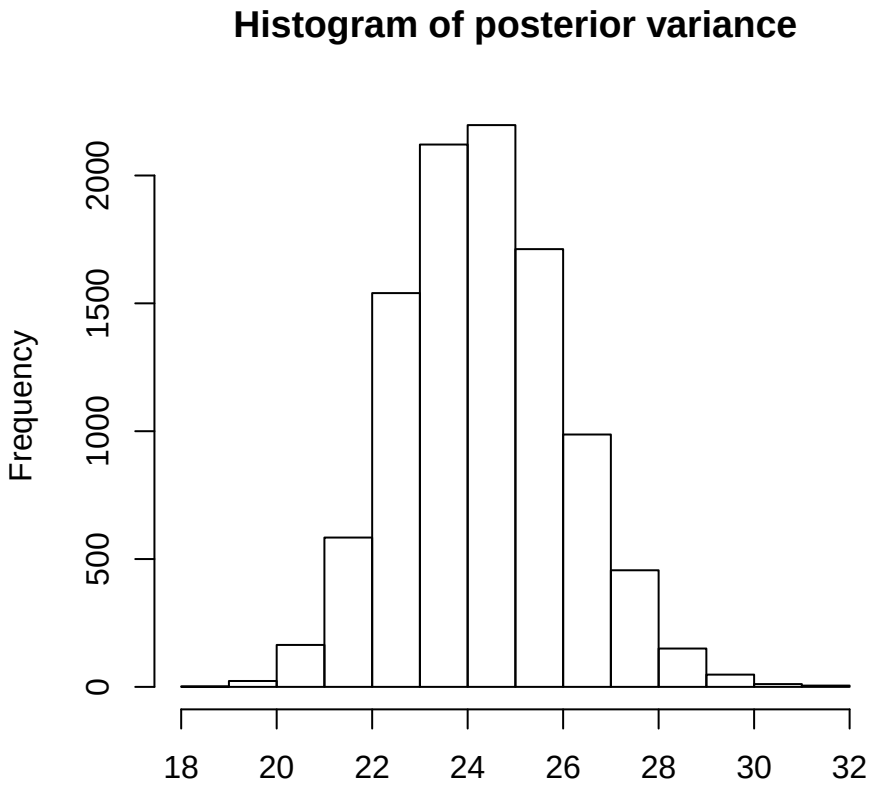
First the model parameters for the prior distribution were to be determined. To do this, a conjugate prior for the linear regression model was used. At first we used some given values for the hyperparameters. These resulted in a large variance between different simulations.

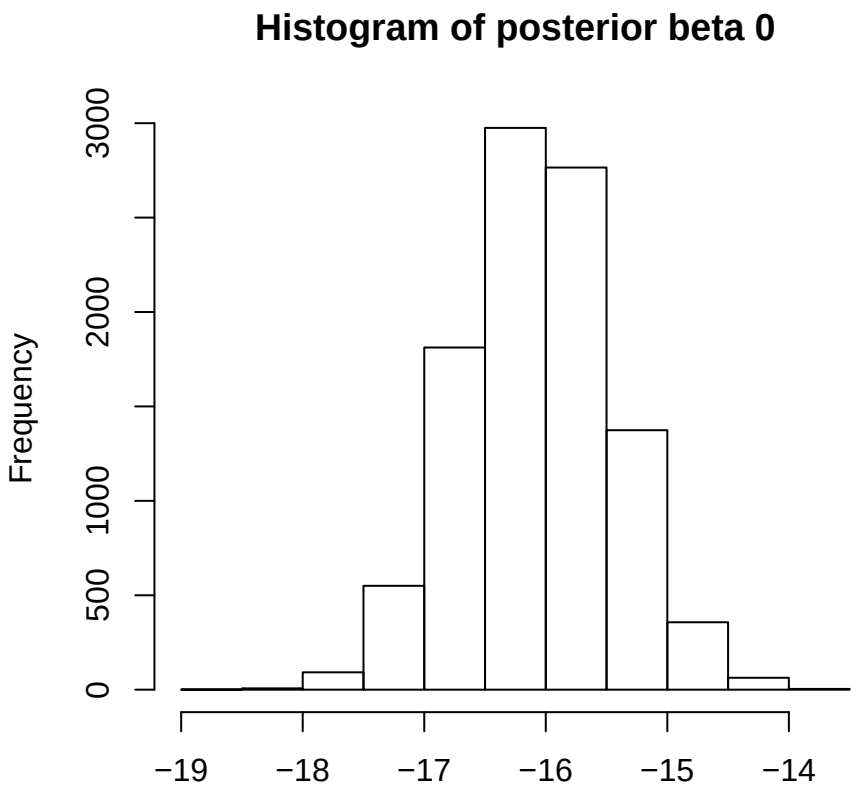
With some trial and error we ended up using the following tuned hyperparameters: $\mu_0 = (-15, 150, -150), \Omega_0 = I_3, \nu_0 = 40, \sigma_0^2 = 1$. This resulted in the following regression curves.

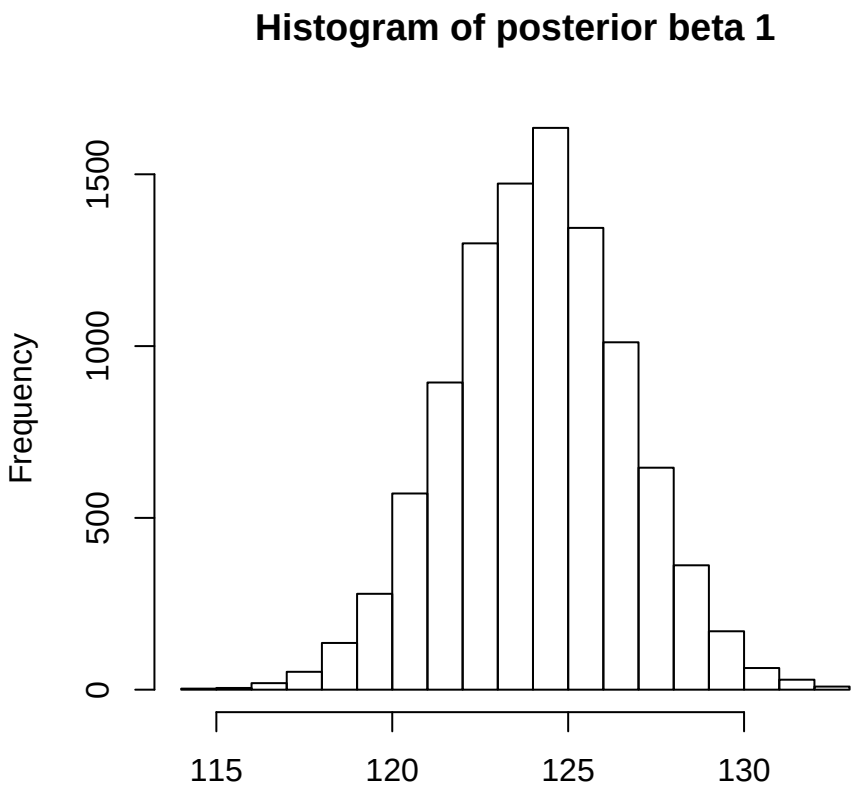


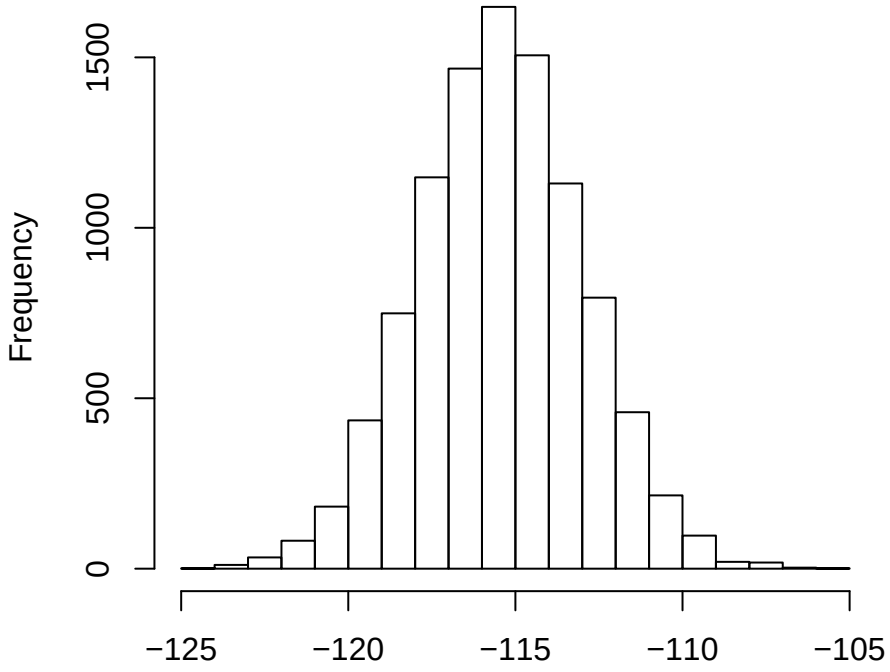
These curves looks reasonable because we see that it is hottest in the middle of the year, i.e during the summer with a temperature around 20. The curves lies roughly around the same temperature, showing that they do not vary too much. Also we see that the mean temperature for the whole year is around 8 degrees, which seems reasonable.

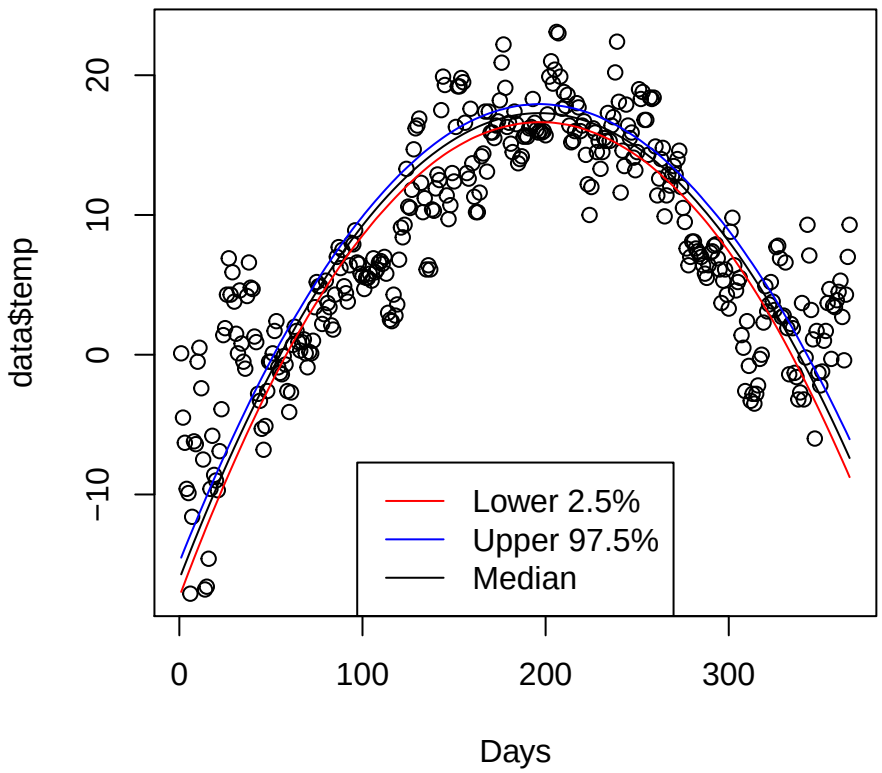
After this the distributions of $\beta_0, \beta_1, \beta_2$ and σ^2 were to be simulated. Below are the histograms of these distributions aswell as a scatter plot of the temperature data.




Histogram of posterior beta 0
Frequency
0
500
1000
2000
3000
−19
−18
−17
−16
−15
−14

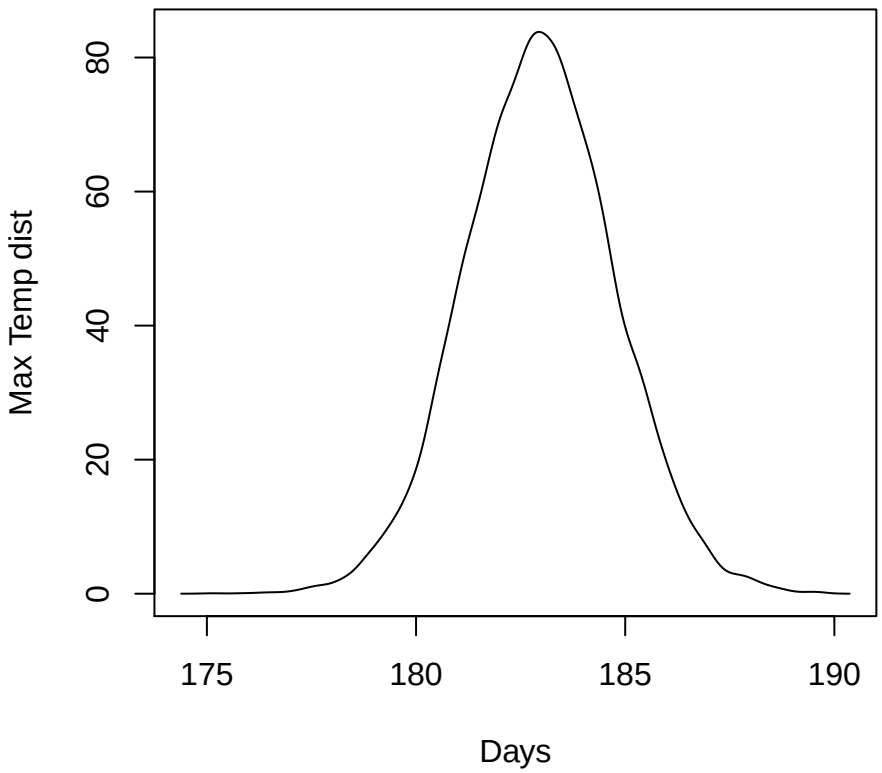

Histogram of posterior beta 1
Frequency
0
500
1000
1500
115
120
125
130


Histogram of posterior beta 2
Frequency
0
500
1000
1500
−125
−120
−115
−110
−105



On the scatter plot, the 95% credible interval is plotted around the median. The interval bands does not contain most data points because we look at the median temperature for each day, which is just a single value.

Below we have plotted the expected highest temperature from the distributions simulated above. We can see that the highest temperature is excpected to be between day 180 and 185, i.e in July, which often is the hottest month of the year.



If we were to estimate a polynomial model of order 7, but we suspect that higher order terms are not needed, we would use the following specifications:

μ_0: vector of length 7, set the last terms to zero so that the higher order terms have no effect.

ω_0: matrix of dim 7x7, set the last terms in the diagonal to a very high value so that the variance of the higher order terms becomes small.

Assignment 2

Posterior approximation for classification with logistic regression

In this assignment a dataset with variables describing parts of women's life situation. This data will be used to predict if a woman works or not.

At first we fitted the logistic regression using maximum likelihood estimation (see code).

Next we wanted to approximate the posterior distribution of the 8-dim parameter vector $\beta|y, X \sim N(\tilde{\beta}, J_y^{-1}(\tilde{\beta}))$. $\tilde{\beta}$ and $J(\tilde{\beta})$ were calculated using the Optim function in R.

The numerical values for $\tilde{\beta}$ is

```
## [1] "0.62673 -0.01979 0.18022 0.16757 -0.1446 -0.08207 -1.35913 -0.02468"
```

and for $J_y^{-1}(\tilde{\beta})$ is

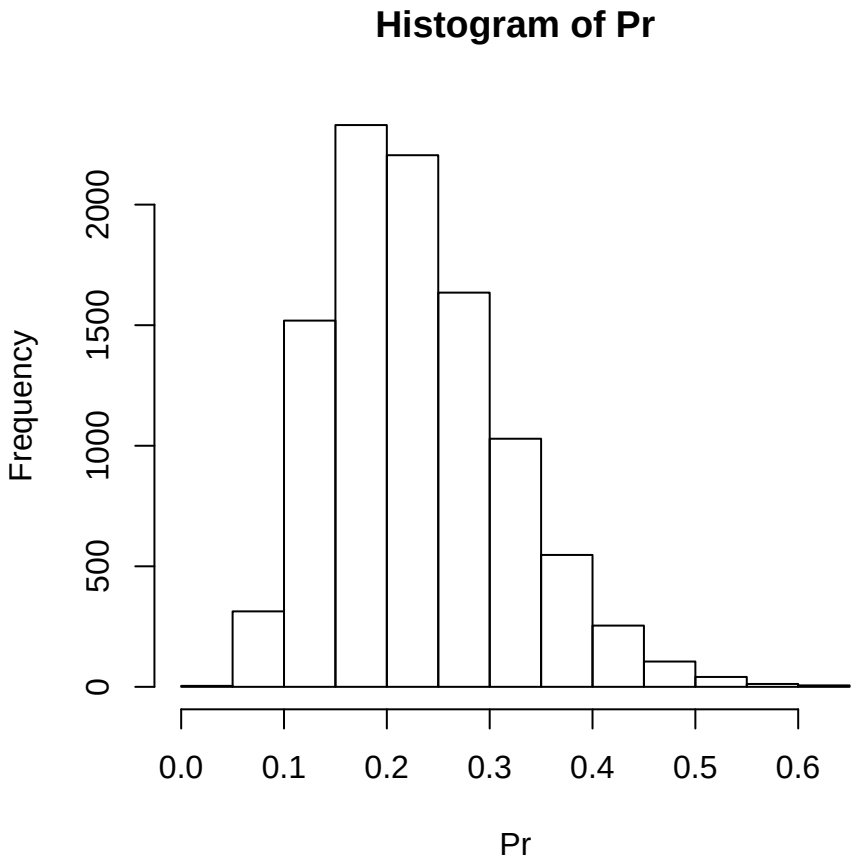
```
## 2.26602 , 0.00334 , -0.06545 , -0.01179 , 0.04578 , -0.03029 , -0.18875 , -0.09802
## 0.00334 , 0.00025 , -0.00056 , -3e-05 , 0.00014 , -4e-05 , 0.00051 , -0.00014
## -0.06545 , -0.00056 , 0.00622 , -0.00036 , 0.0019 , 0 , -0.00613 , 0.00175
## -0.01179 , -3e-05 , -0.00036 , 0.00435 , -0.01425 , -0.00013 , -0.00147 , 0.00054
## 0.04578 , 0.00014 , 0.0019 , -0.01425 , 0.05558 , -0.00033 , 0.00321 , 0.00051
## -0.03029 , -4e-05 , 0 , -0.00013 , -0.00033 , 0.00072 , 0.00518 , 0.0011
## -0.18875 , 0.00051 , -0.00613 , -0.00147 , 0.00321 , 0.00518 , 0.15126 , 0.00677
## -0.09802 , -0.00014 , 0.00175 , 0.00054 , 0.00051 , 0.0011 , 0.00677 , 0.01997
```

After this the credible interval for the variable NSmallChild was calculated, the interval can be seen below

```
##       2.5%      97.5%
## -2.1272365 -0.6170603
```

Since both values are negative, we can confidenlty say that this variable is not important to determine the probability that a woman works. Since it affects the probability negatively.

Now we wrote a function that simulates from the predictive distribution of the response variable in logisitic regression. We used our normal approximation from above, aswell as the equation $Pr(y = 1|x) = \frac{exp(x^T\beta)}{1+exp(x^T\beta)}$. We predicted on a 40 year old woman, with one small child and one big, 8 years of education, 10 years of experience and a husband with an income of 10. The following is our result.



Appendix

Assignment 1

```
data =read.delim("/home/krisi211/Desktop/TDDE07/Lab2/TempLinkoping.txt")

set.seed(235)

## 1a
Ndraws = 10000
mu_0 = t(data.frame(-15,150,-150))
Omega_0 = 1*diag(3)
nu_0 = 40
var_0 = 1


var<-(nu_0*var_0)/rchisq(Ndraws,nu_0)
hist(var,freq = FALSE)

beta = matrix(0,Ndraws,3)
library(mvtnorm)

sim.beta = function(sigma2, mu, omega){
  rmvnorm(n=1,mean=mu,sigma=sigma2*solve(omega))
}

betas = sapply(var,sim.beta,mu = mu_0, omega = Omega_0)
betas = t(betas)

reg_fun =function(betas,time){
  temp = betas[1] + betas[2]*time + betas[3]*time^2
  return(temp)
}

prior.temp = apply(betas[seq(1,25,1),],1,reg_fun, time = data$time)
x = dim(prior.temp)[1]
x.axis = (1:x) / x
plot(x.axis,prior.temp[,1],type = "l",
     xlab = "Time", ylab = "Temperature",ylim=c(-20,30))
for (i in 2:25) {
  lines(x.axis,prior.temp[,i])
}

## 1b
X = cbind(rep(1,length(data$time)),data$time,data$time^2)
X.prim.X = t(X)%*%X
beta.hat = solve(t(X)%*%X)%*%t(X)%*%data$temp
mu_n = solve((t(X)%*%X + Omega_0))%*%(t(X)%*%X%*%beta.hat + Omega_0%*%mu_0)
Omega_n = X.prim.X + Omega_0
nu_n = nu_0 + dim(X)[1]
nuvar_n = nu_0%*%var_0 + (t(data$temp)%*%data$temp +
                            t(mu_0)%*%Omega_0%*%mu_0 - t(mu_n)%*%Omega_n%*%mu_n)
var_n = as.numeric(nuvar_n/nu_n)
```

```
var.post = (nu_n*var_n)/rchisq(Ndraws,nu_n)
hist(var.post,main = "Histogram of posterior variance", xlab = "")

betas.post = sapply(var.post,sim.beta, mu = mu_n, omega = Omega_n)
betas.post = t(betas.post)

hist(betas.post[,1],main = "Histogram of posterior beta 0", xlab = "")
hist(betas.post[,2],main = "Histogram of posterior beta 1", xlab = "")
hist(betas.post[,3],main = "Histogram of posterior beta 2", xlab = "")

plot(data$temp)
post.temp = apply(betas.post[1:1000,],1,reg_fun,time = data$time)
post.temp.median = apply(post.temp,1,median)
lines(post.temp.median)

post.temp.q = apply(post.temp,1,quantile,probs = c(0.025,0.975),na.rm=TRUE)
lines(post.temp.q[1,],col = "red")
lines(post.temp.q[2,],col = "blue")
legend("bottom", col = c("red", "blue", "black"),
       legend = c("Lower 2.5%", "Upper 97.5%", "Median"), lty=1)

## 1c
time = -betas[,2]/(2*betas[,3])
plot(density(time)$x*366,density(time)$y,type="l",xlab = "Days",
     ylab = "Max Temp dist")

## 1d
# mu_0: vector of length 7, set the last terms to zero so that the higher
#order terms have no effect.

# omega_0: matrix of dim 7x7, set the last terms in the diagonal to a very high value
# so that the variance of the higher order terms becomes small.
```

Assignment 2

```
data = read.delim("/home/krisi211/Desktop/TDDE07/Lab2/WomenWork.dat",sep="")
#data = read.delim("/home/ponsv690/Documents/TDDE07/Lab2/WomenWork.dat",sep="")
set.seed(235)

## a
glm.model = glm(Work~0 +., data = data, family = binomial)

## b
tau = 10
library("mvtnorm")
y = as.vector(data[,1])
X = as.matrix(data[,2:9])
params = dim(X)[2]

mu = matrix(0,params,1)
Sigma = tau^2*diag(params)
```

```r
LogPostLogistic = function(betas, y, X, mu, Sigma) {
  params = length(betas)
  linPred = X%*%betas

  logLik = sum(linPred*y -log(1+exp(linPred)))
  if (abs(logLik) == Inf) logLik = -20001

  logPrior = dmvnorm(betas, mu, Sigma, log=TRUE)

  return(logLik + logPrior)
}

initVal = c(rep(0,dim(X)[2]))

OptimResults = optim(initVal,LogPostLogistic,gr=NULL,
                     y,X,mu,Sigma,method=c("BFGS"),
                     control=list(fnscale=-1),hessian=TRUE)

beta.tilde = OptimResults$par
inv.hessian = -solve(OptimResults$hessian)

beta = rmvnorm(10000,beta.tilde,inv.hessian)

quantile(beta[,7],probs = c(0.025,0.975))
CI.lo = beta.tilde[7] + 1.96*(inv.hessian[7,7])^0.5
CI.hi = beta.tilde[7] - 1.96*(inv.hessian[7,7])^0.5

## c
new.X = c(1, 10, 8,10,1,40,1,1)

mul = function(row){
  return(t(new.X)%*%row)
}

pred = apply(beta,1,mul)

Pr = exp(pred)/(1+exp(pred))

hist(Pr)
```