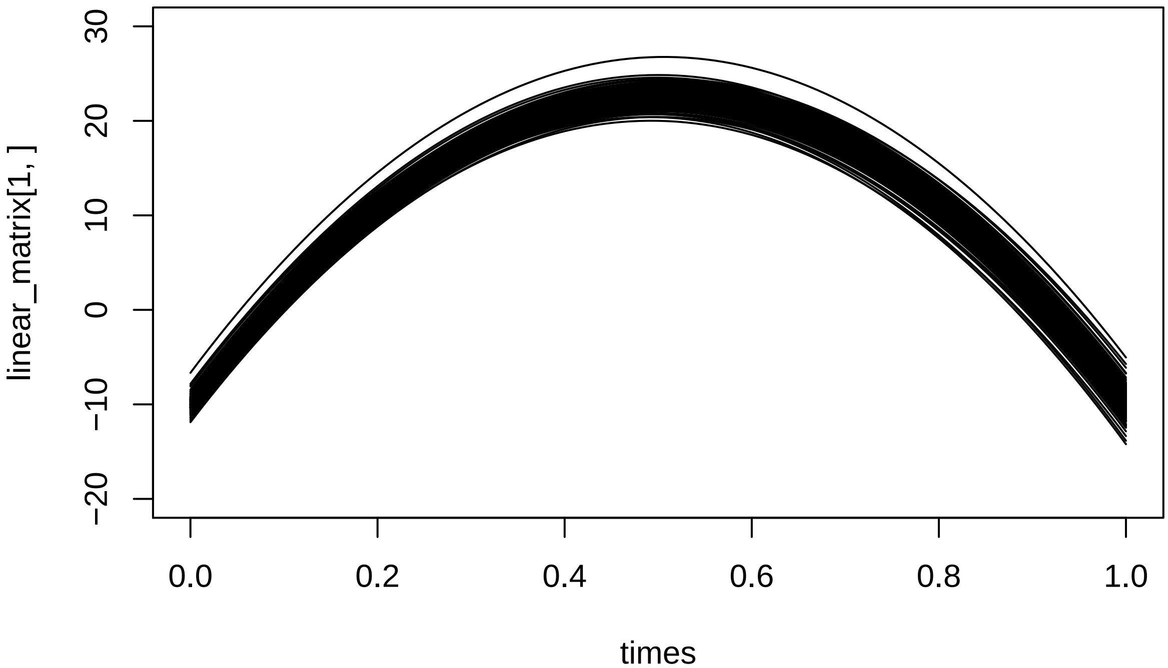
Lab 2 - TDDE07

Axel Holmberg (axeho681), Wilhelm Hansson (wilha431)

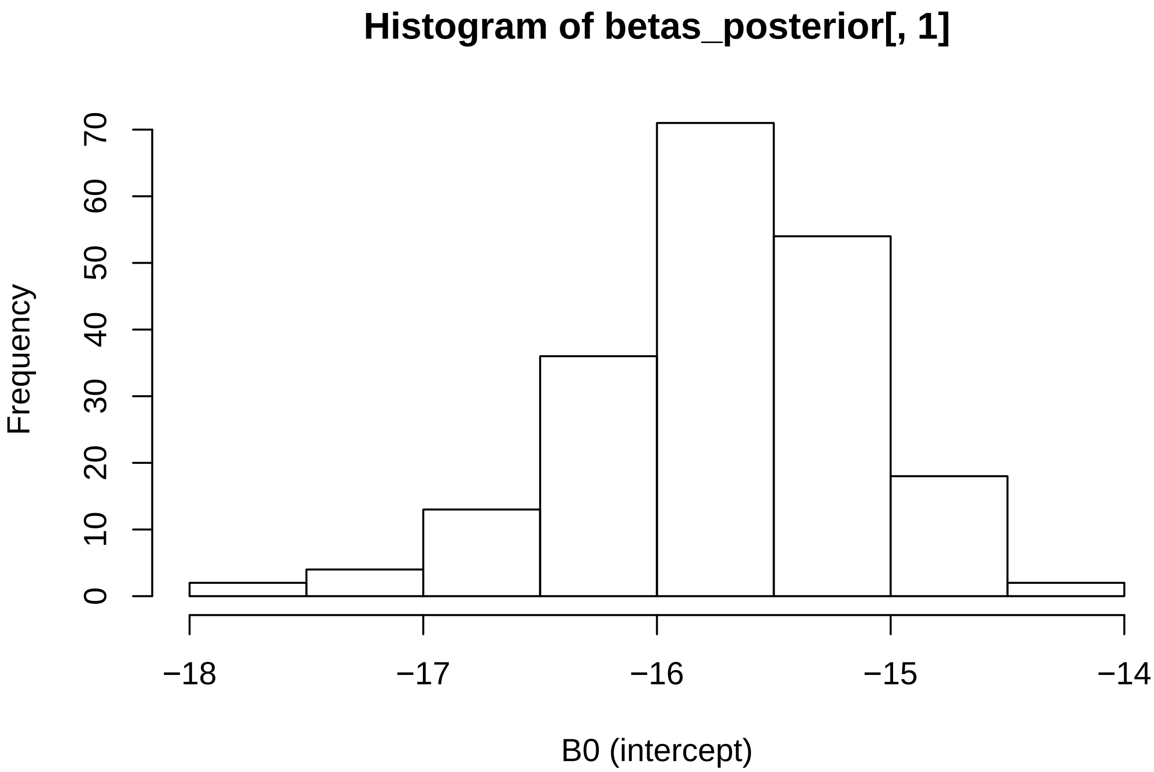
1.

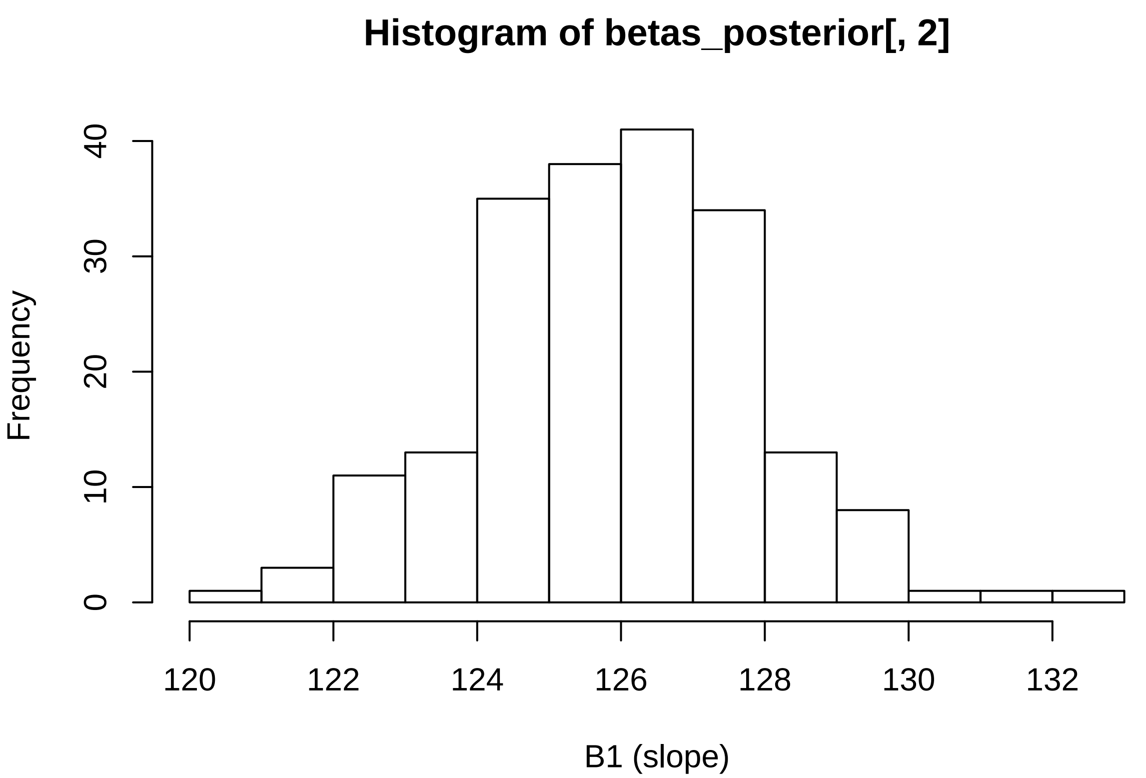
a)

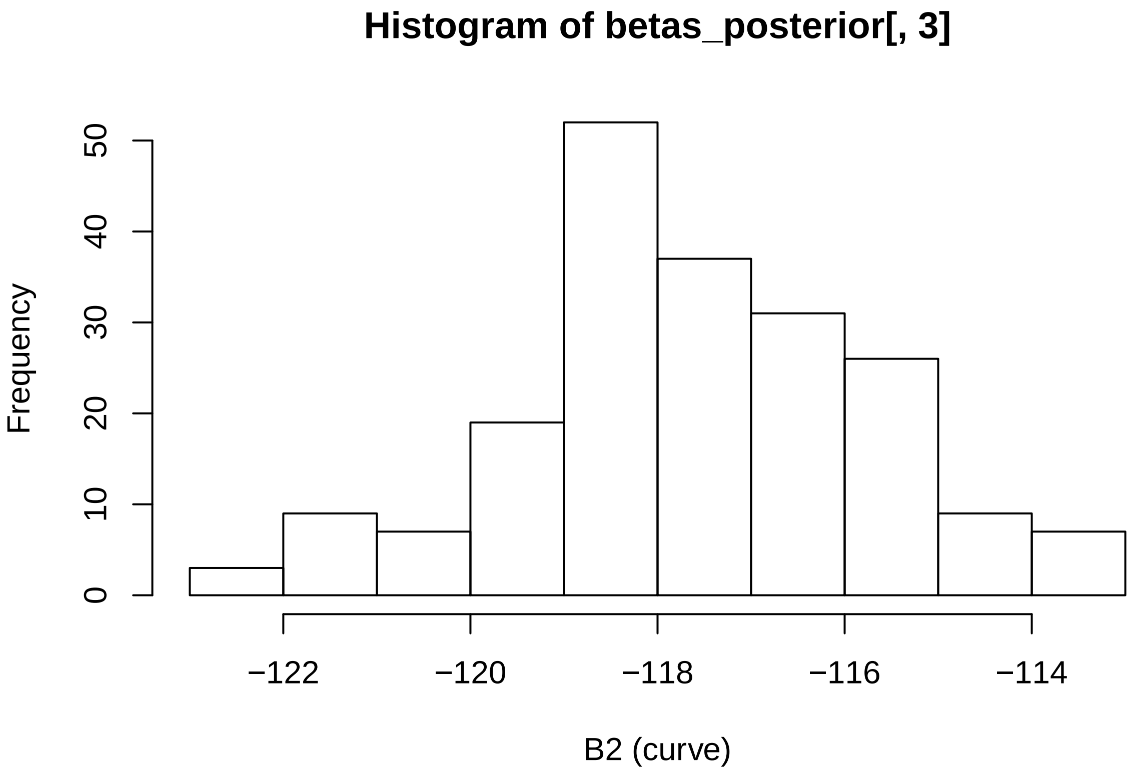


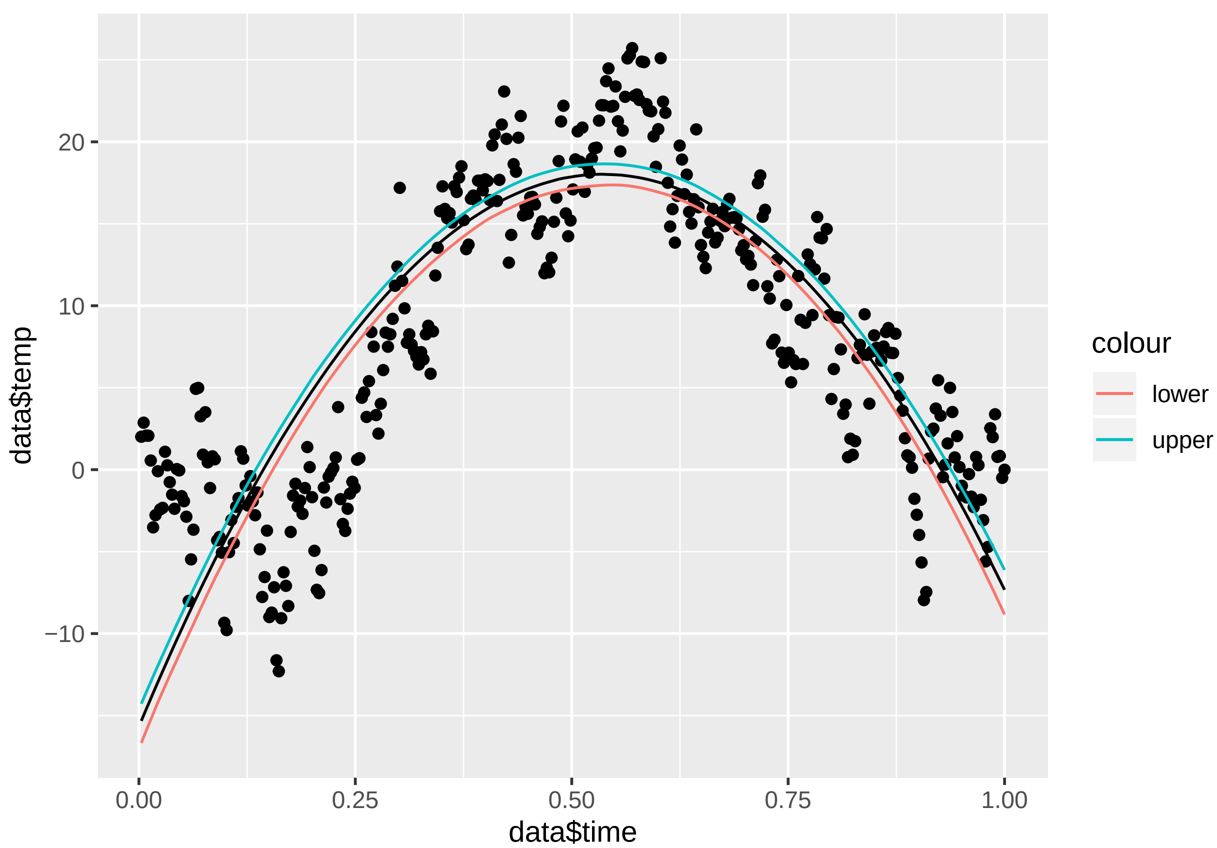
The plot above shows the regression curve for the temperature. The tuning of the joint prior parameters was decided by looking at the values of the actual data. From the data the mean, min and max values were identified. Further the knowledge that 0 = January and 1 = December was used to unnderstand when the peak of the quadratic curve wsa suppose to be (around 0.5 ~ June). So, variance between the individual curves were allowed to avoid to much bias.

b)


Histogram of betas_posterior[, 1]
Frequency
0
10
20
30
40
50
60
70
−18
−17
−16
−15
−14
B0 (intercept)

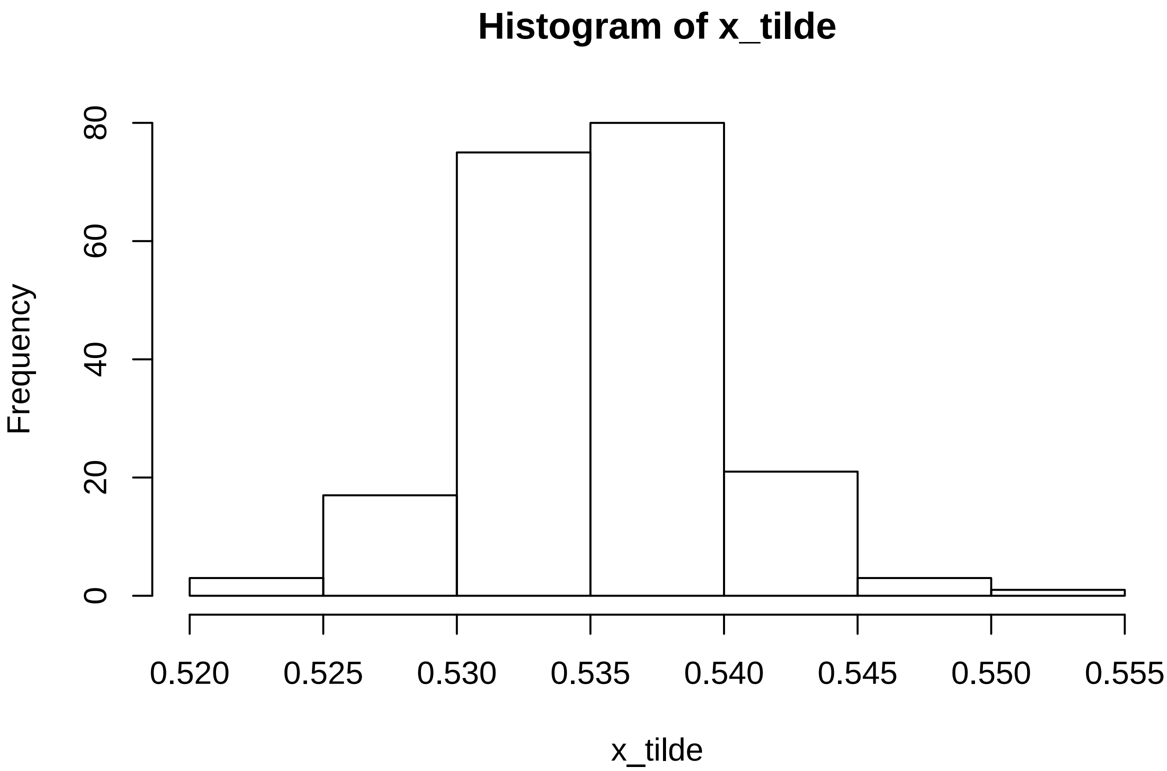

Histogram of betas_posterior[, 2]
Frequency
0
10
20
30
40
120
122
124
126
128
130
132
B1 (slope)





The band does not contain most of the data points. This should be the case as it is the confidence interval of our posterior and not of the data points.

c)



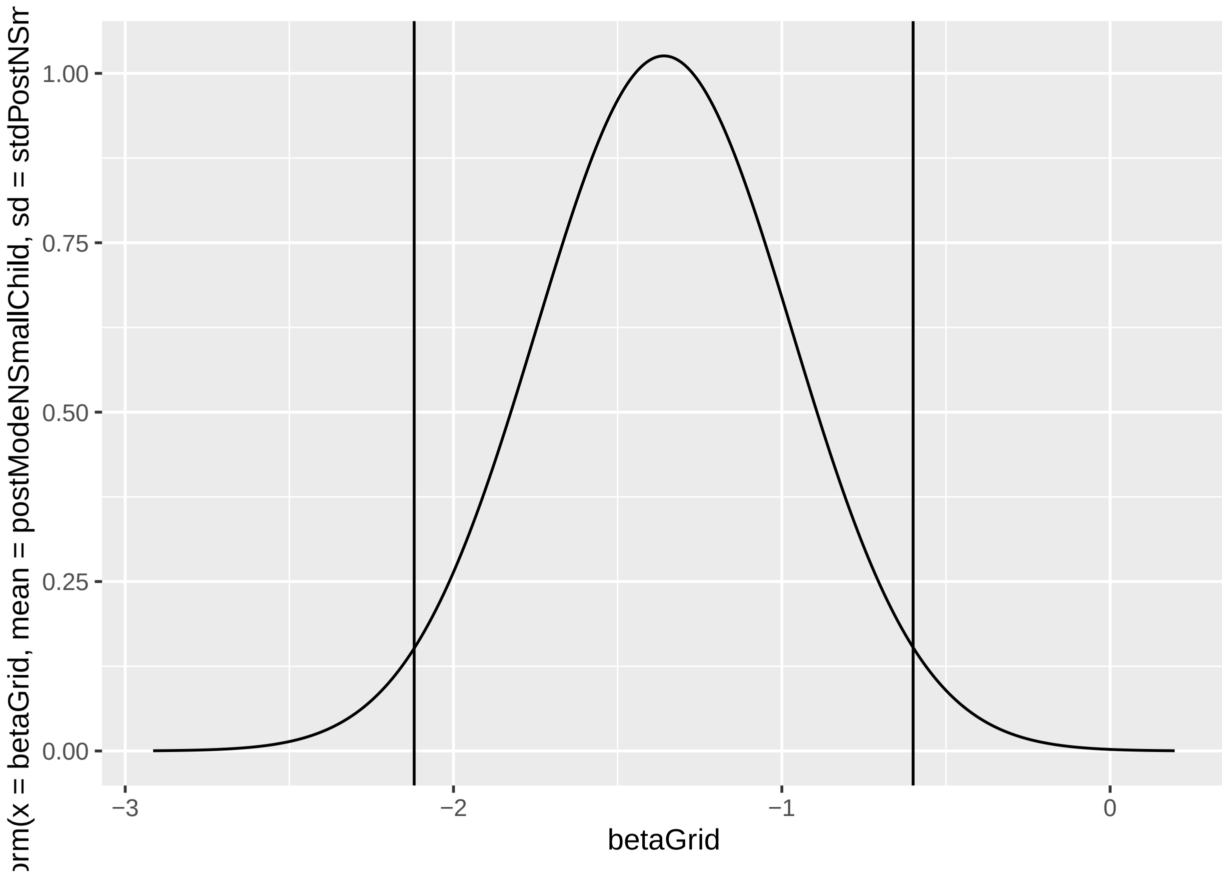
The histogram above shows how $\tilde{x}$ is distributed and it shows that the hottest day is sometime in the middle of July.

d)

A suitable prior to mitigate the potential problem of overfitting would be to set parameters of μ_0 and Ω_0 to values the would result in a $\beta = 0$ for the higher order terms. In this case the $\mu_{4...7} = 0$, and since it is Ω^{-1} that is used the values of Ω should be small.

2.

a)



As the variable for NSmallChild has a 95 % confidence interval between -2.1197916 and -0.600082 we would consider this feature as an important determinant for the probability wether a woman works or not.

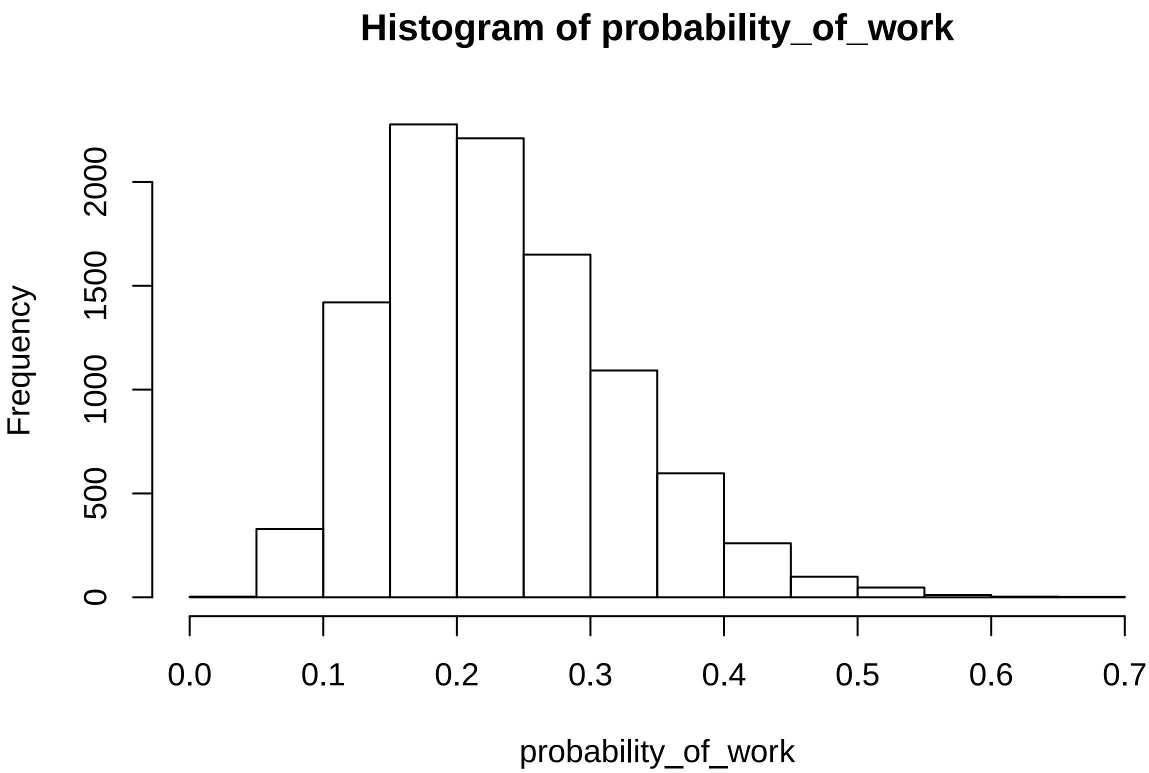
The posterior covariance matrix is:

```
##              [,1]          [,2]          [,3]          [,4]          [,5]
## [1,]  2.266027789  3.338901e-03 -6.545163e-02 -1.179144e-02  0.0457802049
## [2,]  0.003338901  2.528054e-04 -5.610272e-04 -3.125417e-05  0.0001414875
## [3,] -0.065451630 -5.610272e-04  6.218248e-03 -3.558226e-04  0.0018963418
## [4,] -0.011791442 -3.125417e-05 -3.558226e-04  4.351727e-03 -0.0142491717
## [5,]  0.045780205  1.414875e-04  1.896342e-03 -1.424917e-02  0.0555792108
## [6,] -0.030293445 -3.588556e-05 -3.242407e-06 -1.340862e-04 -0.0003299400
## [7,] -0.188747337  5.066946e-04 -6.134698e-03 -1.468969e-03  0.0032082956
## [8,] -0.098024192 -1.444244e-04  1.752747e-03  5.437010e-04  0.0005120558
##               [,6]          [,7]          [,8]
## [1,] -3.029344e-02 -0.1887473366 -0.0980241916
## [2,] -3.588556e-05  0.0005066946 -0.0001444244
## [3,] -3.242407e-06 -0.0061346975  0.0017527466
## [4,] -1.340862e-04 -0.0014689689  0.0005437010
## [5,] -3.299400e-04  0.0032082956  0.0005120558
## [6,]  7.184605e-04  0.0051841702  0.0010952946
## [7,]  5.184170e-03  0.1512628879  0.0067688186
## [8,]  1.095295e-03  0.0067688186  0.0199722751
```

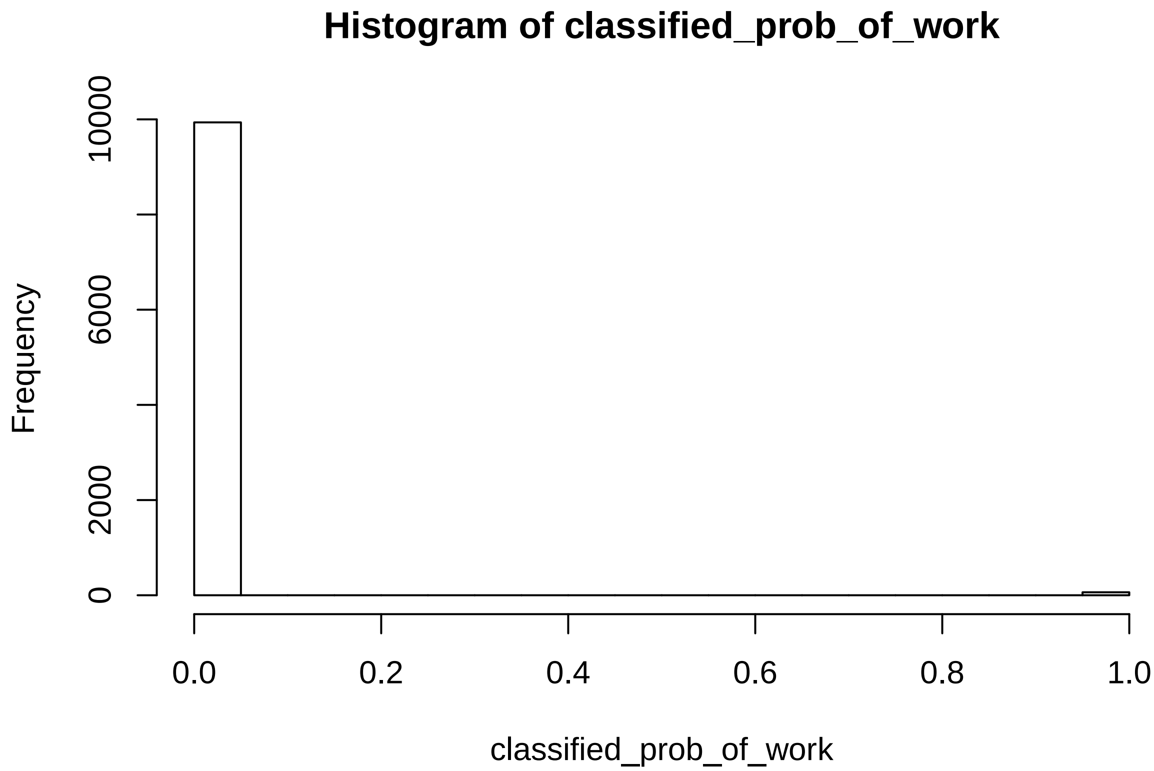
The posterior mode is:

```
##            [,1]        [,2]     [,3]      [,4]       [,5]        [,6]      [,7]
## [1,] 0.6266063 -0.01979155 0.180227 0.1675651 -0.1445918 -0.08206432 -1.359147
##             [,8]
## [1,] -0.02468011
```

b)

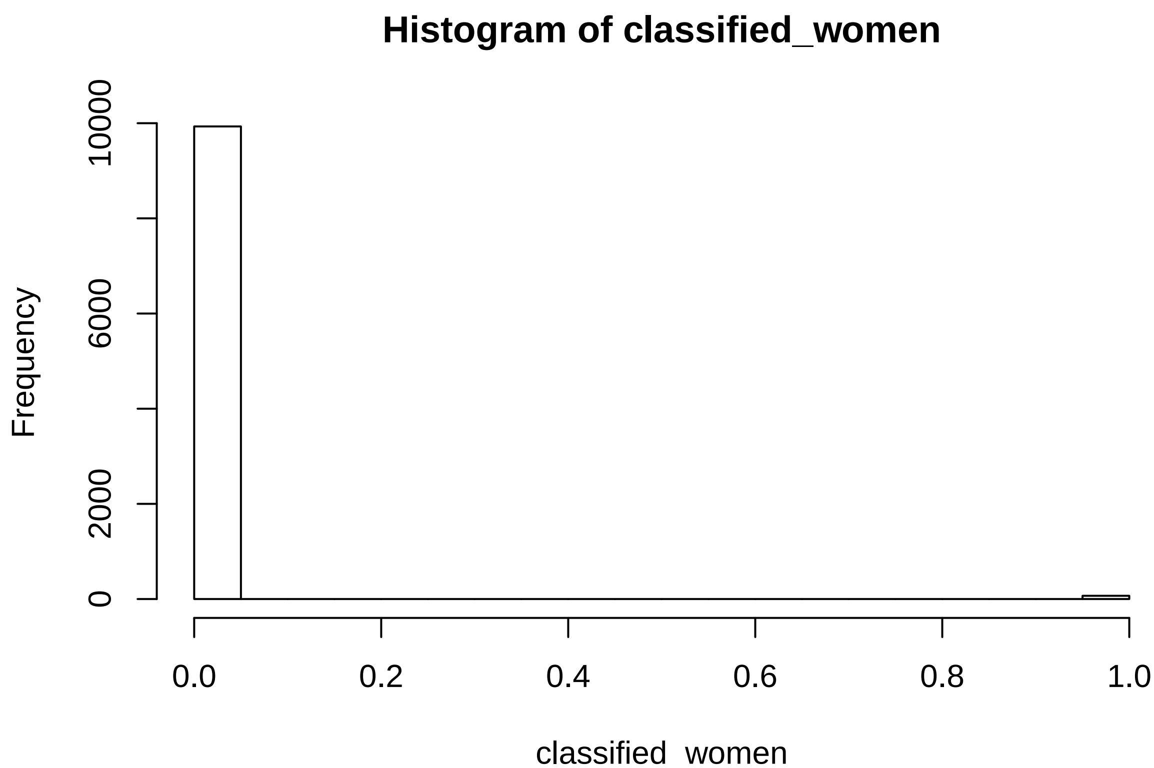


The histogram above shows the distributon for a women with the given feature-vector to work.



As one can see in the histogram above there is a low probability for a women to work given the feature vector.

c)



Considering our plot above it can be expected that in most of the cases none of the 10 women will be working.

Appendix for code

```
setwd("~/Programming/TDDE07/Lab 2")
library(ggplot2)
data <-
    read.delim("TempLinkoping.txt",
                          header = TRUE,
                          sep = "\t",
                          dec = ".")

library(LaplacesDemon)
library(mvtnorm)
set.seed(12345)
my_0 <- matrix(c(-10, 130, -130), nrow = 3, ncol = 1)

omega_0 <- 3 * diag(3)

v_0 <- 10
sigma_sq_0 <- 2

number_of_draws <- 200

sigma_sq <- rinvchisq(number_of_draws, v_0, sigma_sq_0)

omega_0.inv <- solve(omega_0)


betas <- c()


for (sigma in sigma_sq) {
    beta <- rmvnorm(1, my_0, sigma * omega_0.inv)
    #tmp <- cbind(sigma, beta)
    betas <- rbind(betas, beta)
}


linear_matrix <- c()

times <- seq(0, 1, 0.01)
for (rownumber in 1:nrow(betas)) {
    tmprow <- c()
    for (time in times) {
        tmp <- betas[rownumber, 1] + betas[rownumber, 2] * time + betas[rownumber, 3] *
            time ^ 2
        tmprow <- cbind(tmprow, tmp)
    }
    linear_matrix <- rbind(linear_matrix, tmprow)
}

plot(times,linear_matrix[1,], type='l', ylim=c(-20,30))
for (i in 2:nrow(linear_matrix)) {
    lines(times, linear_matrix[i,])
```

```
}


# b)

library(Rmisc)
X <- cbind(rep.int(1, 365), data[, 1], I(data[, 1] ^ 2))
Y <- data[, 2]

beta_hat <- solve((t(X) %*% X)) %*% t(X) %*% Y
my_n <-
    solve(t(X) %*% X + omega_0) %*% (t(X) %*% X %*% beta_hat + omega_0 %*% my_0)
omega_n <- t(X) %*% X + omega_0
v_n <- v_0 + nrow(X)
v_n_sigma_sq_n <-
    v_0 * sigma_sq_0 + (t(Y) %*% Y + t(my_0) %*% omega_0 %*% my_0 - t(my_n) %*%
                                                 omega_n %*% my_n)
n_draws <- 200

sigmas_sq <- rinvchisq(n_draws, v_n, v_n_sigma_sq_n / v_n)
hist(sqrt(sigmas_sq))


betas_posterior <- c()
for (sigma_sq in sigmas_sq) {
    beta_posterior <- rmvnorm(1, my_n, sigma_sq * solve(omega_n))
    betas_posterior <- rbind(betas_posterior, beta_posterior)
}

hist(betas_posterior[, 1], xlab = "B0 (intercept)")
hist(betas_posterior[, 2], xlab = "B1 (slope)")
hist(betas_posterior[, 3], xlab = "B2 (curve)")


out <- c()
for (time in data$time) {
    ys <- c()
    for (i in 1:nrow(betas_posterior)) { #INCREASE


        b0m_tmp <- betas_posterior[i,1]
        b1m_tmp <- betas_posterior[i,2]
        b2m_tmp <- betas_posterior[i,3]

        y <-  b0m_tmp + b1m_tmp * time + b2m_tmp * time ^ 2
        ys <- cbind(ys, y)
    }

    out <- rbind(out,ys)

}

times_median_quantiles <- matrix(nrow=length(data$time), ncol=3)
```

```r
for (i in 1:nrow(out)) {
    ci <- quantile(out[i,], c(0.025,0.975))
    times_median_quantiles[i,1] = ci[1]
    times_median_quantiles[i,3] = ci[2]
    times_median_quantiles[i,2] = median(out[i,])

}

ggplot() + geom_point(aes(x = data$time, y = data$temp)) + geom_line(aes(x =
    data$time, y = times_median_quantiles[, 2])) + geom_line(aes(x = data$time, y = times_median_quantil
    "upper")) + geom_line(aes(x = data$time, y = times_median_quantiles[, 1], color = "lower"))

x_tilde = -1 * betas_posterior[,2] / (2* betas_posterior[,3])
hist(x_tilde)

################################### Ass 2 #####################################################
library("mvtnorm") # This command reads the mvtnorm package into R's memory. NOW we can use dmvnorm fun

data <-
    read.table("WomenWork.dat", header = TRUE)  # Spam data from Hastie et al.

y <- as.vector(data$Work)
X <- as.matrix(data[, 2:9])


tau <- 10
mu <- rep(0, 8)
sigma <- tau ^ 2 * diag(8)
initVal <- rmvnorm(1, mu, sigma)



LogPostLogistic <- function(betaVect, y, X, mu, Sigma) {
    nPara <- length(betaVect)

    linPred <- X %*% betaVect

    #print(betaVect)
    # evaluating the log-likelihood
    logLik <- sum(linPred * y - log(1 + exp(linPred)))
    #print(logLik)
    if (abs(logLik) == Inf)

        logLik = -20000
    # Likelihood is not finite, stear the optimizer away from here!

    # evaluating the prior
    logPrior <- dmvnorm(betaVect, matrix(0, nPara, 1), Sigma, log = TRUE)

    print(logLik + logPrior)
    #print(betaVect)
    # add the log prior and log-likelihood together to get log posterior
    return(logLik + logPrior)
```

```r
}


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

inverse_hessian <- solve(OptimResults$hessian)

postModeNSmallChild <- OptimResults$par[7]
stdPostNSmallChild <- sqrt(diag(-inverse_hessian)[7])



betaGrid <-
    seq(
        postModeNSmallChild - 4 * stdPostNSmallChild,
        postModeNSmallChild + 4 * stdPostNSmallChild,
        length = 1000
    )
ciBetaNsmallChild <-
    quantile(rnorm(10000, mean = postModeNSmallChild, sd = stdPostNSmallChild),
                     c(0.025, 0.975))
ggplot() + geom_line(aes(
    x = betaGrid,
    y = dnorm(x = betaGrid, mean = postModeNSmallChild, sd = stdPostNSmallChild)
)) + geom_vline(xintercept = ciBetaNsmallChild[1]) + geom_vline(xintercept = ciBetaNsmallChild[2])


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

glmSmallChildVar <- glmModel[["coefficients"]][["NSmallChild"]]

#b)
beta_posterior <- OptimResults$par

features <- c(1, 10.0, 8, 10, (10 / 10) ^ 2 , 40, 1, 1)

stdbeta <- (inverse_hessian)

func_working_women <- function(draws, optim_beta, optim_std) {
    sim_beta <-
        rmvnorm(draws,
```

```r
                        mean = optim_beta,
                        sigma = optim_std)
    return((exp(features %*% t(sim_beta))) / (1 + exp(features %*% t(sim_beta))))
}



probability_of_work <-  func_working_women(10000, beta_posterior, stdbeta)

classified_prob_of_work <- ifelse(probability_of_work>=0.5,1,0)

hist(probability_of_work)
hist(classified_prob_of_work)


### c)

classified_women <- c()


for (i in 1:10) {
    probability_of_work <-
        func_working_women(1000, beta_posterior, stdbeta)

    classified_prob_of_work <- ifelse(probability_of_work >= 0.5, 1, 0)

    classified_women <- rbind(classified_women, classified_prob_of_work)

}
hist(classified_women)
```