CPMA 573 — FINAL EXAM

Due Tuesday, May 5, 2015

Professor John Kern

This is a 200-point take-home examination, due by noon on Thursday, May 1. The work you submit for this exam must be your own; no group work of any kind is permitted.

NAME _____

Score: ___/200

Problem 1 (40 points):

Generate 25,000 independent values from the following probability density, and plot them in a histogram:

$$f(x) = \begin{cases} \frac{x-2}{2} & \text{if } 2 \leq x \leq 3 \\ \frac{2-x/3}{2} & \text{if } 3 \leq x \leq 6 \end{cases}.$$

Be sure to describe the method you used, and include all relevant code. Then, use the simulations to find the expected value of this distribution. Compare this expected value with the theoretical value.

Rejection sampling #2 notes/#3 HW

A = 2.5

proposal density - uniform

denom = A × 0.25

Lisa Over
Final Exam Problem 1
May 5, 2015

METHOD

The 25,000 independent values from the following probability density were obtained using rejection sampling with a uniform proposal density.

$$f(x) = \begin{cases} \frac{x-2}{2} & if\ 2 \leq x \leq 3 \\ \frac{2 - x/3}{2} & if\ 3 \leq x \leq 6 \end{cases}$$

To generate the 25,000 independent values from $f(x)$, a random uniform value, u, was drawn, transformed to be within the domain of $f(x)$, i.e., $x = u * 4 + 2$, and then "proposed" as being a value from the target density $f(x)$. This value was accepted as being from $f(x)$ if the probability that it was from $f(x)$ was greater than the probability represented by another independent uniform random value.

The probability that the value x was from $f(x)$ was calculated as one of the following two ratios depending on the value of x:

$p = \frac{\frac{x-2}{2}}{A*0.25}$ when $2 \leq x \leq 3$ and where A = 2

$p = \frac{\frac{2-x/3}{2}}{A*0.25}$ when $3 < x \leq 6$ and where A = 2

The height of the uniform proposal density over the 4-unit interval [2, 6] is 0.25. To guarantee that the calculated ratio is a value between (0, 1), the denominator must always be greater than or equal to any value in the numerator, i.e., any value in the range of $f(x)$. The factor required to raise the height of the proposal density to 0.5, which is the maximum height of $f(x)$, is A = 2.

A summary of the acceptance procedure and criteria for x follow:

Draw a second random uniform value, u2, to act as a comparison probability. If the probability that x is a value from the target density, p, is greater than u2, accept x as being a value from $f(x)$. Otherwise, reject x.

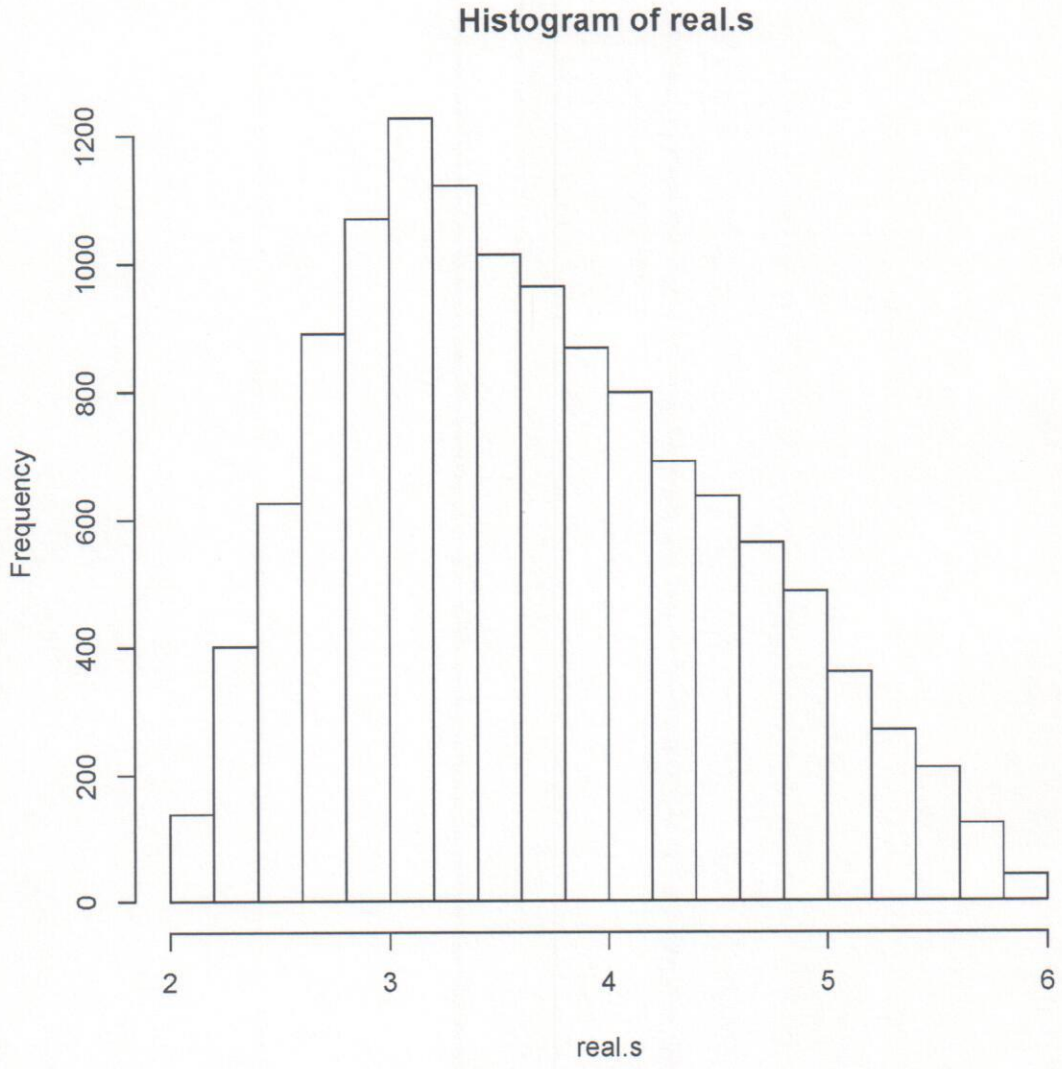
Lisa Over
Final Exam Problem 1
May 5, 2015

CODE

```
#The proposal density is uniform over a 4-unit interval so the height of the proposal density is 0.25. The maximum height of the target density is 0.5 so multiply 0.25 by 2 (A=2) so the height of the proposal density is greater than the height of the target density for all f(x).
A = 2
real.s = NULL
N = 25000

for(i in 1:N) {
        #Draw a uniform independent random variable and transform it to be within the domain of the given density, 2 <= x <= 6.
        x = runif(1) * 4 + 2

        #Determine what part of the piecewise function to use based on the value of x and set the numerator equal to the appropriate function and the denominator equal to A times the height of the uniform density over the four unit interval.
        if (x <= 3) {
                num = (x - 2)/2
                denom = A*0.25
                }
        else {
                num = (2 - (x/3))/2
                denom = A*0.25
                }
        #Calculate the probability that the previously generated uniform is from the target density
        prob <- num/denom

        #Generate a uniform independent random variable to use as a comparison probability: if the probability, prob, that the transformed x uniform could be from the given distribution is greater than the probability represented by this newly generated uniform random variable, select the x transformed uniform and store in real.s
        if(runif(1) < prob) {
                real.s <- c(real.s,x)
        }
}
hist(real.s)
mean(real.s)
#[1] 3.664185
```

Lisa Over
Final Exam Problem 1
May 5, 2015

RESULTS



The mean of the 25,000 independent realizations from $f(x)$ is 3.664185.
The theoretical mean of $f(x)$ is 3.66 (see notes).

Lisa Over

Problem 1

Theoretical Expected Value

$$f(x) = \begin{cases} \frac{x-2}{2} & 2 \le x \le 3 \\ \frac{2-x/3}{2} & 3 \le x \le 6 \end{cases}$$

$$E[X] = \int_2^3 x \cdot \frac{x-2}{2}\,dx + \int_3^6 x \cdot \frac{2-x/3}{2}\,dx$$

$$= \frac{1}{2}\left[\int_2^3 x^2 - 2x\,dx + \int_3^6 2x - \frac{x^2}{3}\,dx\right]$$

$$= \frac{1}{2}\left[\frac{x^3}{3} - x^2\Big|_2^3 + x^2 - \frac{x^3}{9}\Big|_3^6\right]$$

$$= \frac{1}{2}\left(\frac{22}{3}\right) = \boxed{3.66}$$

Problem 2 (90 points):

A physiologist wants to investigate the impact of exercise on the human immune system. The physiologist theorizes that the amount of immunoglobulin y in blood (called IgG, an indicator of long-term immunity) is related to the maximal oxygen uptake x (a measure of aerobic fitness level) of a person by the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \, ,$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Note that this is equivalent to saying

$$y \sim \mathrm{N}(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2).$$

Values of x and y measured for each of 30 human subjects can be found in

`www.mathcs.duq.edu/~kern/exim.dat`

Use these data in combination with independent, uniform$(-2^{11}, 2^{11})$ prior densities on $\{\beta_0, \beta_1, \beta_2\}$, and independent prior $\pi(\sigma^2) \propto \sigma^{-2}$ to complete the following:

i. Write down the joint posterior (up to a proportionality constant) for the four parameters $\{\beta_0, \beta_1, \beta_2, \sigma^2\}$.

ii. Identify the full conditional distribution for these parameters and then use Gibbs sampling to sample from the marginal posterior density of each of the four parameters. Include all relevant computer code, as well as trace plots and autocorrelation plots to verify convergence and independence.

iii. Plot the data, and superpose three quadratics that correspond to three randomly selected $\{\beta_0, \beta_1, \beta_2\}$ coordinates from your simulation. These are three "possible" regression curves that model these data. When plotting these curves, use dashed-line command `lty=2`.

iv. Use the `apply` command in Splus to superpose the *pointwise* mean regression curve to the plot in **iii**. Make this curve bold with `lwd=2.5`.

v. Provide a histogram of draws from the posterior predictive distribution of IgG measurements for the next person who enters with uptake $x = 60$.

Problem 2

$Y_i \sim N(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2)$ → Likelihood

Y_i is normal with mean $\beta_0 + \beta_1 x + \beta_2 x^2$ and constant variance σ^2.

Joint Posterior with uniformed priors

i. $$\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}\right] \frac{1}{\sigma^2} \cdot \frac{1}{10000} \cdot \frac{1}{10000} \cdot \frac{1}{10000}$$

$$\propto \left[\frac{1}{\sigma^2}\right]^{\frac{n}{2}+1} e^{\frac{-1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}$$

ii. Full conditional for σ^2

$$\alpha = n/2 \quad \lambda = \frac{\sum (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}{2}$$

$$(\sigma^2 | \beta_0, \beta_1, \beta_2, \vec{y}, \vec{x}, \vec{x^2}) \sim IG\left(\frac{n}{2}, \frac{\sum (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}{2}\right)$$

Full conditional for β_0

$$\prod_{i=1}^{n} e^{\frac{-1}{2\sigma^2}(y_i - \beta_1 x_i - \beta_2 x_i^2 - \beta_0)^2}$$

$$= \prod_{i=1}^{n} e^{\frac{-1}{2\sigma^2}[(y_i - \beta_1 x_i - \beta_2 x_i^2)^2 - 2\beta_0 (y_i - \beta_1 x_i - \beta_2 x_i^2) + \beta_0^2)]}$$

$$= e^{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_2 x_i^2)^2 - 2\beta_0 \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_2 x_i^2) - n\beta_0^2\right]}$$

divide by n →

$$= e^{-\frac{n}{2\sigma^2}\left[\frac{\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 x_i^2)^2}{n} - 2\beta_0\frac{\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 x_i^2)}{n} + \beta_0^2\right]}$$

$$\propto e^{-\frac{n}{2\sigma^2}\left(\beta_0 - \frac{\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 x_i^2)}{n}\right)^2}$$

$$(\beta_0 \mid \sigma^2, \beta_1, \beta_2, \vec{y}, \vec{x}, \vec{x^2}) \sim N\left(\frac{\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 x_i^2)}{n}, \frac{\sigma^2}{n}\right)$$

Full conditional for β_1

$$\prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_2 x_i^2-\beta_1 x_i)^2}$$

$$= \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_2 x_i^2)^2 - 2\beta_1 x_i (y_i-\beta_0-\beta_2 x_i^2) + (\beta_1 x_i)^2)}$$

$$= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_2 x_i^2) - 2\beta_1\sum_{i=1}^{n} x_i (y_i-\beta_0-\beta_2 x_i^2) + \beta_1^2\sum_{i=1}^{n} x_i^2\right)}$$

$$= e^{-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\left(\frac{\sum(y_i-\beta_0-\beta_2 x_i^2)}{\sum x_i^2} - \frac{2\beta_1\sum x_i (y_i-\beta_0-\beta_2 x_i^2)}{\sum x_i^2} + \beta_1^2\right)}$$

$$\propto e^{-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\left(\beta_1 - \frac{\sum x_i (y_i-\beta_0-\beta_2 x_i^2)}{\sum x_i^2}\right)^2}$$

$$(\beta_1 \mid \sigma^2, \beta_0, \beta_2, \vec{y}, \vec{x}, \vec{x^2}) \sim N\left(\frac{\sum x_i (y_i-\beta_0-\beta_2 x_i^2)}{\sum x_i^2}, \frac{\sigma^2}{\sum x_i^2}\right)$$

Full Conditional for β_2

$$\prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i-\beta_2 x_i^2)^2}$$

$$= \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}\left[(y_i-\beta_0-\beta_1 x_i)^2 - 2\beta_2 x_i^2(y_i-\beta_0-\beta_1 x_i) + (\beta_2 x_i^2)^2\right]}$$

$$= e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 - 2\beta_2\sum_{i=1}^{n} x_i^2(y_i-\beta_0-\beta_1 x_i) + \beta_2^2\sum_{i=1}^{n} x_i^4\right]}$$

$$= e^{-\frac{\sum x_i^4}{2\sigma^2}\left[\frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{\sum x_i^4} - \frac{2\beta_2\sum_{i=1}^{n} x_i^2(y_i-\beta_0-\beta_1 x_i)}{\sum x_i^4} + \beta_2^2\right]}$$

$$\propto e^{-\frac{\sum x_i^4}{2\sigma^2}\left[\beta_2 - \frac{\sum x_i^2(y_i-\beta_0-\beta_1 x_i)}{\sum x_i^4}\right]^2}$$

$$(\beta_2 \mid \sigma^2, \beta_0, \beta_1, \vec{y}, \vec{x}, \vec{x}^2) \sim N\left(\frac{\sum_{i=1}^{n} x_i^2(y_i-\beta_0-\beta_1 x_i)}{\sum_{i=1}^{n} x_i^4}, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}\right)$$

Lisa Over
Final Exam Problem 2
May 5, 2015

CODE

```
#Read file and attach
data = read.csv(file.choose(), header=TRUE)
attach(data)

#Set gibbs parameters

y = data$IgG
x = data$Max.O2.Uptake

output = summary(glm(y ~ x + I(x^2)))

N = 25000
lag = 500
burnin = 0
b0 = output$coef[1,1]
b1 = output$coef[2,1]
b2 = output$coef[3,1]

#Gibbs function receives eight parameters: "x" and "y" data values; "b0", "b1", and "b2" initial regression coefficients, "N" for the number of realizations, "lag" for determining how many realizations to skip between saves, and "burnin" for determining how many realizations to skip before starting to save. Gibbs generates N independent random normal values for each regression coefficient and N independent inverse gamma values for the regression variance.

gibbs <- function(x,y,b0,b1,b2,N,lag,burnin) {

        #obtain length of data
        n = length(x)
        #Set N to be N*lag+burnin
        N <- N*lag + burnin

        #Initialize vectors to hold the beta coefficients and sigsq realizations
        b0s = NULL
        b1s = NULL
        b2s = NULL
        s2s = NULL

        for(i in 1:N) {

                #Generate a sigsq, s2, based on current regression coefficients b0, b1, b2
                s2 = 1/rgamma(1, n/2, sum((y-b0-b1*x-b2*x^2)^2)/2)
                #Generate coefficients based on current values of coefficients and sigsq, s2
                b0 = rnorm(1, sum(y-b1*x-b2*x^2)/n, sqrt(s2/n))
                b1 = rnorm(1, sum(x*(y-b0-b2*x^2))/sum(x^2), sqrt(s2/sum(x^2)))
                b2 = rnorm(1, sum(x^2*(y-b0-b1*x))/sum(x^4), sqrt(s2/sum(x^4)))
```

Lisa Over
Final Exam Problem 2
May 5, 2015

```
        #if i is greater than burnin and if i is a multiple of the lag, store alpha, beta, and siqsq
        if(i > burnin) {
         if(i %% lag == 0) {
                b0s <- c(b0s,b0)
                b1s <- c(b1s,b1)
                b2s <- c(b2s,b2)
                s2s <- c(s2s,s2)
   }
  }
        }

        vectors <- list("beta0s" = b0s, "beta1s" = b1s, "beta2s" = b2s, "sigsqs" = s2s)
        return(vectors)

}

v = gibbs(x,y,b0,b1,b2,N,lag,burnin)

beta0s = v$beta0s
beta1s = v$beta1s
beta2s = v$beta2s
sigsqs = v$sigsqs

mean(beta0s)
mean(beta1s)
mean(beta2s)
mean(sigsqs)

#Plot each parameter with ts.plot, hist, and acf

par(mfrow=c(3,1)) #split plotting window into 3 rows and 1 column
ts.plot(beta0s,xlab="Iterations")
hist(beta0s,probability=T, cex.lab=1.5, cex.axis=1.5)
acf(beta0s,lag.max=500)

par(mfrow=c(3,1)) #split plotting window into 3 rows and 1 column
ts.plot(beta1s,xlab="Iterations")
hist(beta1s,probability=T, cex.lab=1.5, cex.axis=1.5)
acf(beta1s,lag.max=500)

par(mfrow=c(3,1)) #split plotting window into 3 rows and 1 column
ts.plot(beta2s,xlab="Iterations")
hist(beta2s,probability=T, cex.lab=1.5, cex.axis=1.5)
acf(beta2s,lag.max=500)

par(mfrow=c(3,1)) #split plotting window into 3 rows and 1 column
ts.plot(sigsqs,xlab="Iterations")
hist(sigsqs,probability=T, cex.lab=1.5, cex.axis=1.5)
```

Lisa Over
Final Exam Problem 2
May 5, 2015

```
acf(sigsqs,lag.max=500)

#Create a matrix of regreesion coefficients - N rows and 3 columns
k=1
mcoef <- matrix(, nrow = N, ncol = 3)
for(i in 1:N) {
                coord = c(beta0s[i], beta1s[i], beta2s[i])
                mcoef[k ,] = coord
                k = k + 1
 }

#Create a matrix of yhat values - N rows and 30 columns (number of x values)
yhats <- matrix(, nrow = N, ncol = length(x))
for(i in 1:N) {
        yhat <- mcoef[i,1] + mcoef[i,2]*x + mcoef[i,3]*x^2
        yhats[i ,] = yhat
}
#Calculate the means of the columns (2) of the yhats matrix
means = apply(yhats, 2, mean)


#Select 3 random numbers between 0 and 25000 to represent the indices of the 3 randomly selected
coefficient coordinates from the mcoef matrix
rLines = sample(1:25000, 3, replace=F)

#Create a 3-row matrix of yhat values for each x using the randomly selected coordinates
yhats3 <- matrix(, nrow = 3, ncol = length(x))
for(i in 1:3) {
        yhat <- mcoef[rLines[i],1] + mcoef[rLines[i],2]*x + mcoef[rLines[i],3]*x^2
        yhats3[i ,] = yhat
}

#Plot x and y with possible fitted curves after spline smoothing
smooth1 = smooth.spline(x,yhats3[1,], spar=0.35)
smooth2 = smooth.spline(x,yhats3[2,], spar=0.35)
smooth3 = smooth.spline(x,yhats3[3,], spar=0.35)
smooth4 = smooth.spline(x, means, spar=0.35)

plot(x,y)
lines(smooth1, lty=2)
lines(smooth2, lty=2)
lines(smooth3, lty=2)
lines(smooth4, lwd=2.5)
```

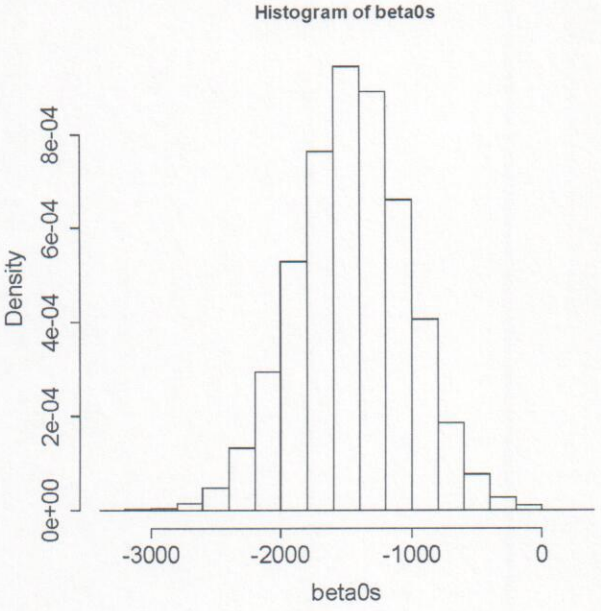
Lisa Over
Final Exam Problem 2
May 5, 2015

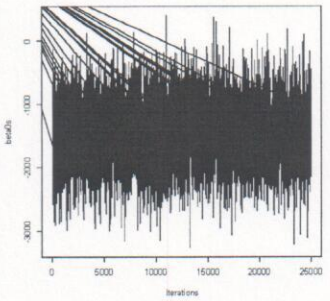
```
#Create a vector of yhat values with O2 uptake = 60 and plot a histogram of the values
yhats60 = NULL
for(i in 1:N) {
        yhat60 <- mcoef[i,1] + mcoef[i,2]*60 + mcoef[i,3]*60^2
        yhats60 = c(yhats60, yhat60)
}
hist(yhats60)

mean(yhats60)
quantile(yhats60, 0.025)
quantile(yhats60, 0.975)
```

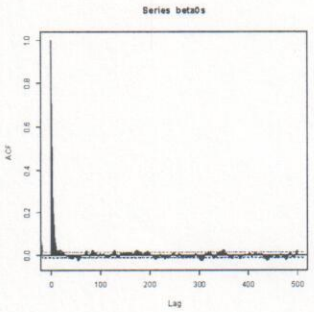
Lisa Over
Final Exam Problem 2
May 5, 2015

RESULTS

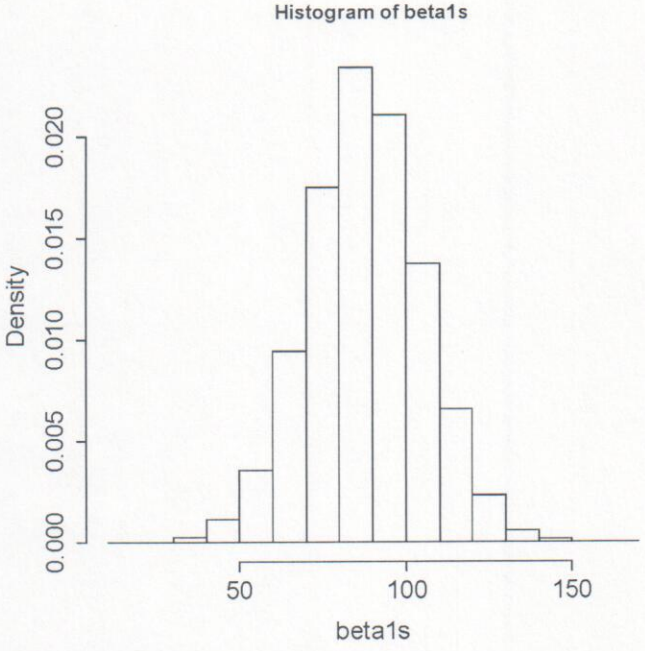
Beta0 mean: **-1452.651**

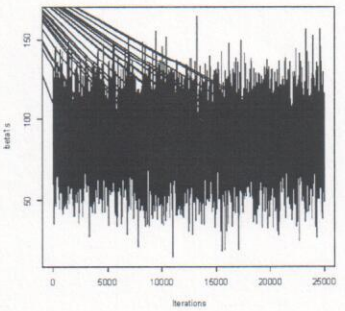


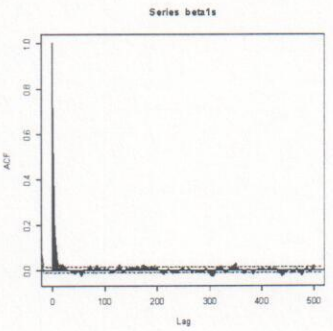




Beta1 mean: **87.85156**

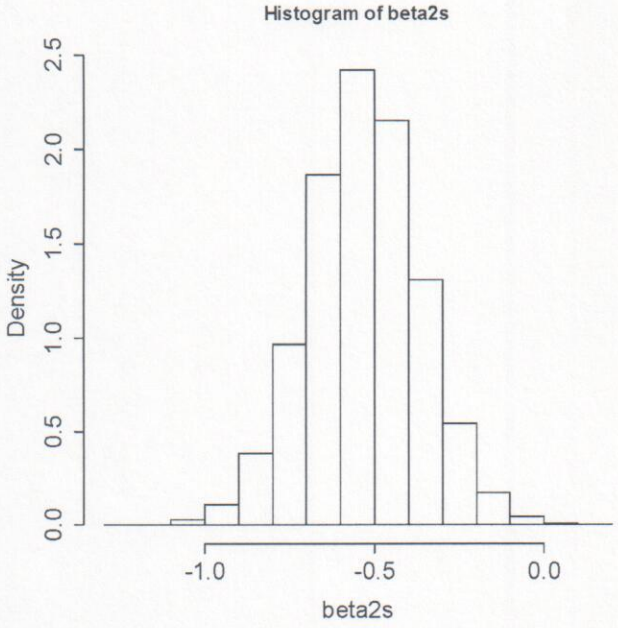


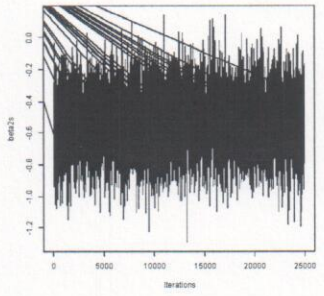


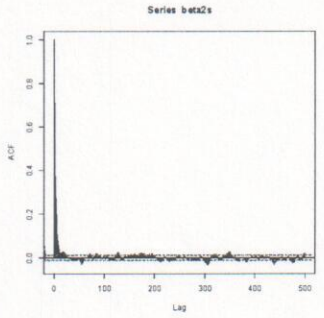


Lisa Over
Final Exam Problem 2
May 5, 2015

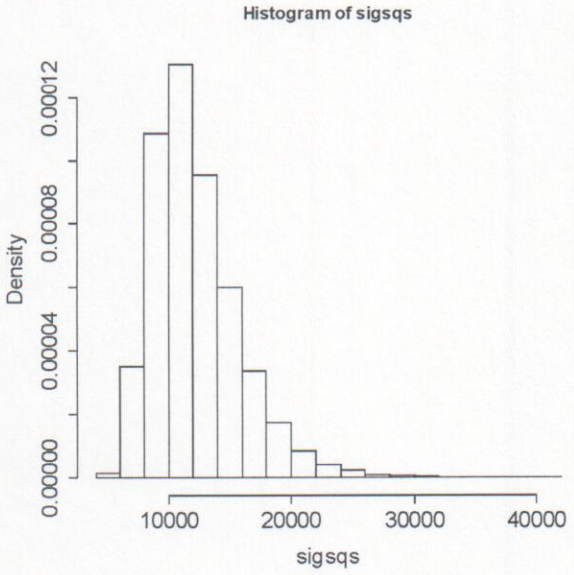
Beta2 mean: **-0.5320706**

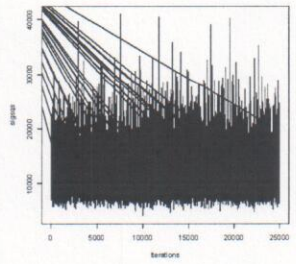


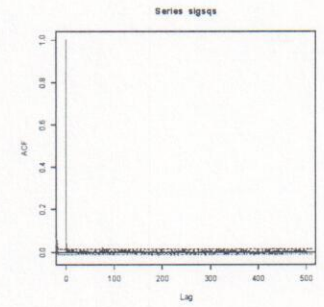




Sigma Sq mean: **12219.16**

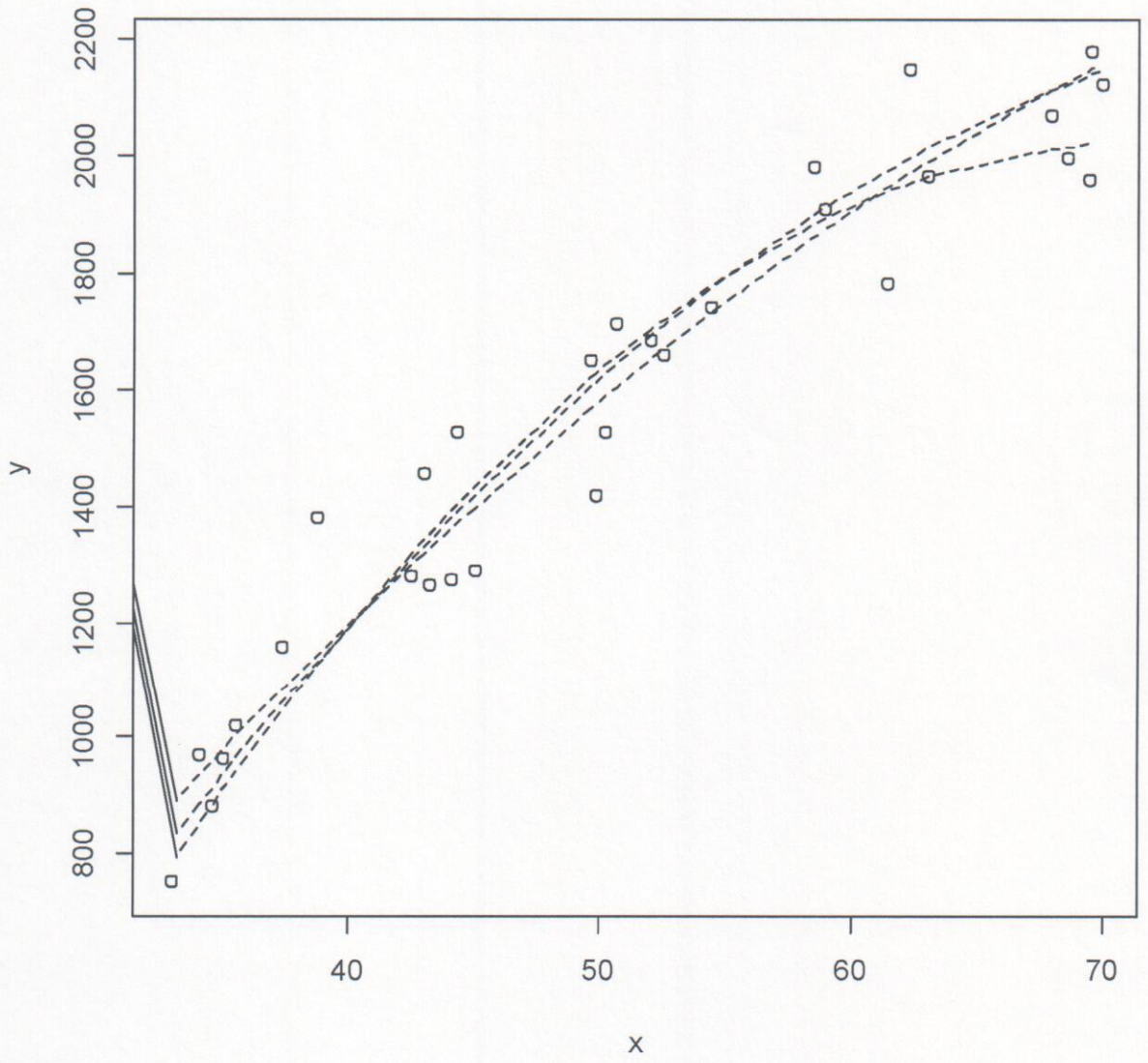






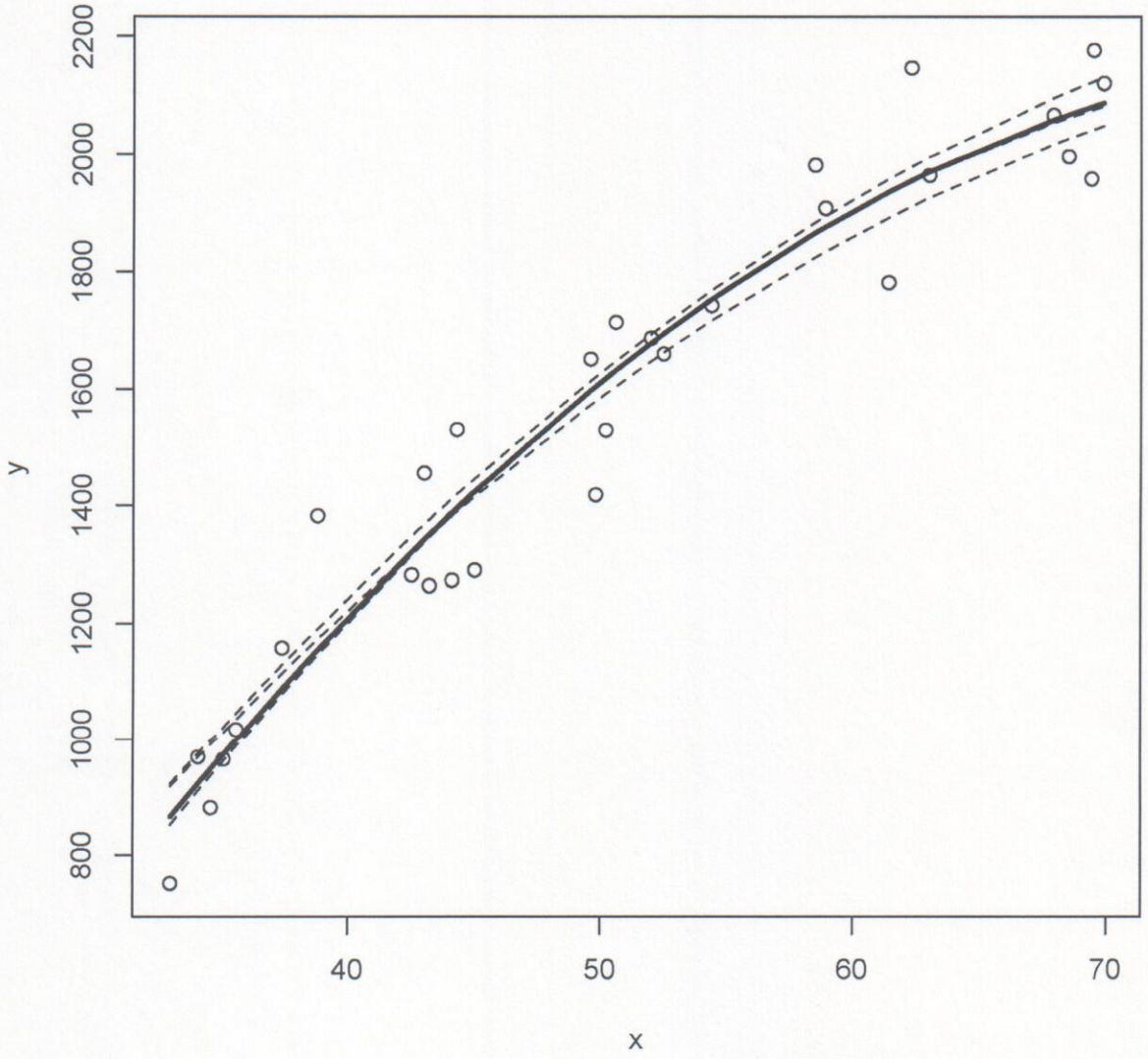
Lisa Over
Final Exam Problem 2
May 5, 2015

The following scatterplot shows the data with three possible regression curves that model the data. The regression coefficients for these curves were selected randomly from a matrix of possible combinations of coefficients, β_0, β_1, and β_2, with each combination of coefficients representing a possible regression curve. The coefficients were drawn from their respective posterior predictive distributions.



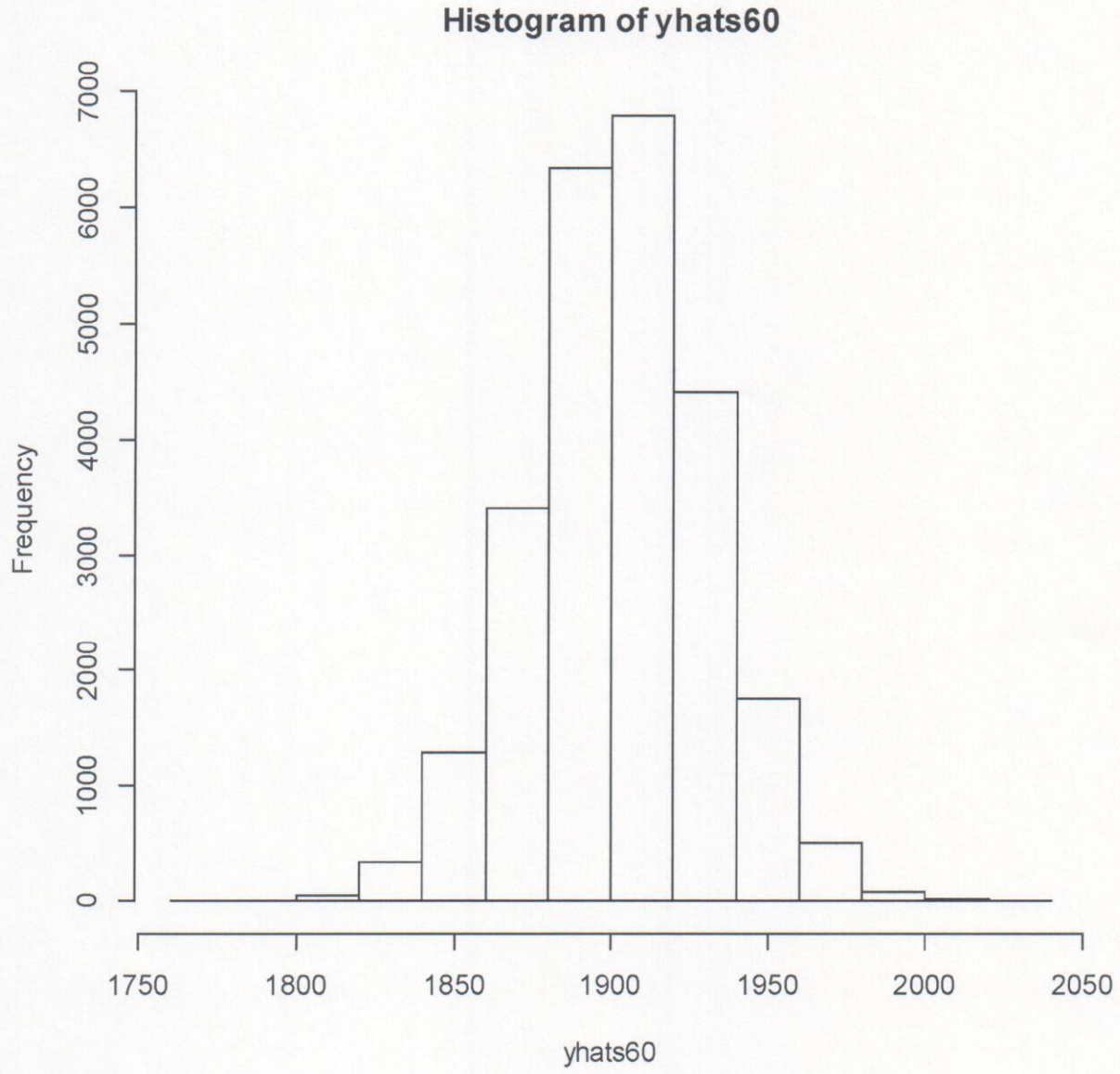
Lisa Over
Final Exam Problem 2
May 5, 2015

The following scatterplot shows data with the three randomly selected regression curves superimposed as dotted lines and the mean regression curve superimposed as a bold line. The mean regression curve was obtained by averaging each column of coefficients, β_0, β_1, and β_2, in the coefficient matrix.



Lisa Over
Final Exam Problem 2
May 5, 2015

The following histogram shows the distribution of possible IgG measurements for a person with maximal oxygen uptake equal to 60. The possible values were computed using the regression curves generated from the posterior predictive distributions of the regression coefficients, β_0, β_1, and β_2, and hence from the posterior predictive distribution of IgG. The mean IgG is 1902.988 with a 95% credible interval (1845.851, 1959.284).



Mean: **1902.988**

95% Credible Interval: **(1845.851, 1959.284)**

Problem 3 (70 points):

Consider the following two-component Poisson mixture distribution:

$$f(x) = (1-p)\frac{b^x e^{-b}}{x!} + p\frac{(b+\mu)^x e^{-(b+\mu)}}{x!} \quad \text{for} \quad x = 0, 1, \ldots.$$

I've simulated 74 observations from this model with $b = 1$ and undisclosed values of $\mu > 0$ and $p \in (0, 1)$. You can find these data at

`www.mathcs.duq.edu/~kern/pmix.dat`

Your goal for this problem is twofold:

1. Provide an interval in which μ has a 95% chance of occurring.
2. Provide an interval in which p has a 95% chance of occurring.

Be sure to use the following in your analysis:

- An exponential prior on μ: $\pi(\mu) \propto e^{-\mu/10}$.
- A mixture model that uses latent variables $z_1, \ldots, z_{74}$ to indicate from which Poisson component the corresponding counts originate.

Please include all relevant computer code, as well as any plots or graphical summaries appropriate to support your work.

Problem 3

Let $b = 1$ and $Y = (b+\mu) = (1+\mu)$

$$f(x_1, x_2, \ldots, x_n \mid \vec{z}, b, \mu, p) =$$

$$\left[\frac{p(Y)^{x_1} e^{-Y}}{x_1!}\right]^{1-z_1} \cdot \left[\frac{(1-p) b^{x_1} e^{-b}}{x_1!}\right]^{z_1} \cdot \ldots \cdot$$

$$\left[\frac{p Y^{x_n} e^{-Y}}{x_n!}\right]^{1-z_n} \cdot \left[\frac{(1-p) b^{x_n} e^{-b}}{x_n!}\right]^{z_n}$$

$$= \frac{1}{\Sigma x_i!^{n-\Sigma z_i}} \frac{1}{\Sigma x_i!^{\Sigma z_i}} p^{n-\Sigma z_i} (1-p)^{\Sigma z_i} Y^{\Sigma x_i - \Sigma x_i z_i}_{i \exists z=1} b^{\Sigma x_i z_i}_{i \exists z=0} e^{-Y(n-\Sigma z_i)} e^{-b\Sigma z_i}$$

↳ joint density as a function of $x_1, x_2, \ldots, x_n$ and likelihood as a function of p, b, Y, and $\vec{z}$

μ has an exponential prior $\pi(\mu) \propto e^{-\mu/10}$
Convert $\pi(\mu)$ to $\pi(Y)$: $e^{-\mu/10} \cdot e^{b/10} \cdot e^{-b/10} = e^{\frac{-\mu-b}{10}} e^{-b/10} = e^{-Y} e^{-b/10}$

Joint Posterior

$$\pi(p, Y, \vec{z} \mid \vec{x}) \propto p^{n-\Sigma x_i} (1-p)^{\Sigma x_i} \frac{Y^{\Sigma x_i - \Sigma x_i z_i} e^{-Y(n-\Sigma z_i)} e^{-\Sigma z_i} e^{-Y/10}}{\Sigma x_i!^n}$$

$$\propto p^{n-\Sigma x_i} (1-p)^{\Sigma x_i} Y^{\Sigma x_i - \Sigma x_i z_i} e^{-Yn + Y\Sigma z_i - Y/10}$$

$$\propto p^{n-\Sigma x_i} (1-p)^{\Sigma x_i} Y^{\Sigma x_i - \Sigma x_i z_i} e^{-Y(n-\Sigma z_i + 1/10)}$$

Full Conditionals

p: Beta $(n - \Sigma z_i + 1, \Sigma z_i + 1)$

Y: Gamma $(\Sigma x_i - \Sigma x_i z_i + 1, n - \Sigma z_i + 1/10)$

z_i: Bernoulli $\left[\frac{(1-p) e^{-1}/x!}{\frac{p(Y^x e^{-Y})}{x!} + (1-p)(e^{-1})/x!}\right]$

Lisa Over
Final Exam Problem 3
May 5, 2015

CODE

```
#Select file and read data
data = read.csv(file.choose(), header=TRUE)
attach(data)

#Create x vector from "data"
x = X4

#Set Gibbs parameters
p = 0.5
b = 1
m = mean(x)
Y = m + b
N = 25000
lag = 1
burnin = 0

#Gibbs function receives six parameters: "x" for the sample data values, "p" for the probability that data value, xi, comes from distribution 1, "Y" for the mean of the distribution with mean (b+m) with b=1, "N" for number of realizations, "lag" for determining how many realizations to skip between saves, and "burnin" for determining how many realizations to skip before starting to save. Gibbs generates N independent z vectors of indicator variables, N independent means from a Poisson distribution, and N independent probabilities that are each a probability of drawing from distribution 1.

gibbs <- function(x,p,Y,N,lag,burnin) {

        #obtain length of x
        n = length(x)

        #Set N to be N*lag+burnin
        N <- N*lag + burnin

        #Initialize vectors to hold the z vector realizations, m2 realizations, and p realizations
        zs = NULL
        Ys = NULL
        ps = NULL

        for(i in 1:N) {
```

Lisa Over
Final Exam Problem 3
May 5, 2015

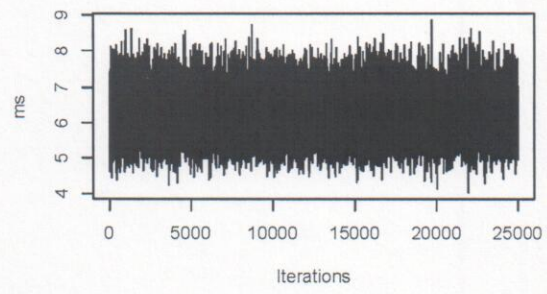
```
        #Calculate the vector of probabilities where each value represents the
probability that the corresponding value in the data vector x was drawn from
distribution 1, the lower mean distribution
        probz1 = (1-p)*exp(-1)/factorial(x)
        probz2 = p*Y^x*exp(-Y)/factorial(x)
        probz = probz1 + probz2
        pvec = probz1/probz

        #Calculate the z vector of latent indicator values where each value is either 0
or 1 to indicate which distribution the corresponding value in the data vector x is
from - "0" indicates distribution 2 and "1" indicates distribution 1
        z = rbinom(n,1,pvec)

        #Compute a mean for the mixed model from a gamma distribution using the
indicator variable vector z and the data vector x.
        Y = rgamma(1,(sum(x)-sum(x*z)+1),(n-sum(z)+1/10))

        #Compute a p for the probability that the data value comes from distribution
1
        p = rbeta(1,n-sum(z)+1,sum(z)+1)

        #if i is greater than burnin and if i is a multiple of the lag, store z, Y, and p
  if(i > burnin) {
        if(i %% lag == 0) {
                zs <- c(zs,z)
                Ys <- c(Ys,Y)
                ps <- c(ps,p)
   }
   }
}

  vectors <- list("zs" = zs, "Ys" = Ys, "ps" = ps)
  return(vectors)

}

vectors = gibbs(x,p,Y,N,lag,burnin)

#vectors$zs represents Gibbs sampler realizations for the z vector
#vectors$Ys represents Gibbs sampler realizations for the mean+1 of distribution 2
#vectors$ps represents Gibbs sampler realizations for the probability that an x
value comes from distribution 1
zs = vectors$zs
Ys = vectors$Ys
```

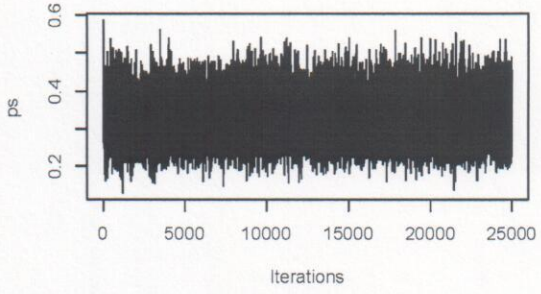
Lisa Over
Final Exam Problem 3
May 5, 2015

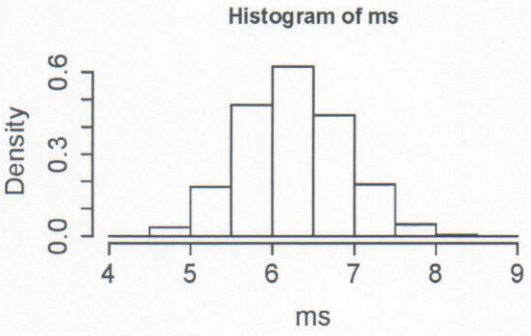
```
ps = vectors$ps

#Since Y = m + 1, subtract 1 from Ys values to create a vector of m values
ms = Ys - 1

par(mfrow=c(3,2)) #split plotting window into 2 rows and 2 columns
ts.plot(ms,xlab="Iterations")
ts.plot(ps,xlab="Iterations")
hist(ms,probability=T, cex.lab=1.5, cex.axis=1.5)
hist(ps,probability=T, cex.lab=1.5, cex.axis=1.5)
acf(ms,lag.max=500)
acf(ps, lag.max=500)

#Convert z to an n column matrix
zmat = matrix(zs, ncol=length(x), byrow=TRUE)
#Get column means of matrix zmat
zimeans = colMeans(zmat)

plot(x,zimeans)

mean(ms)
mean(ps)
quantile(ms, 0.025)
quantile(ms, 0.975)
quantile(ps, 0.025)
quantile(ps, 0.975)
```

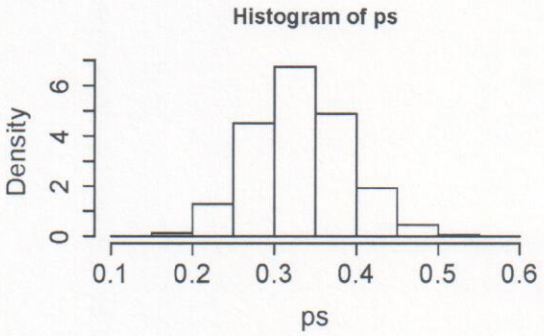
Lisa Over
Final Exam Problem 3
May 5, 2015

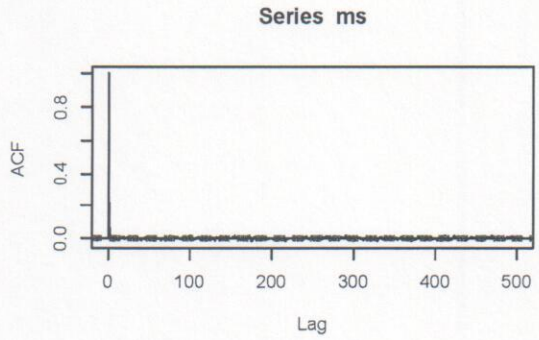
RESULTS

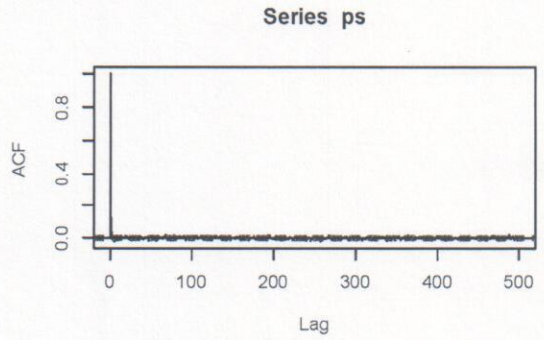












Lisa Over
Final Exam Problem 3
May 5, 2015

In the plot below, the proportion of 1s for each x_i, i.e., the means of the 25,000 z_i values generated for each x_i, is plotted against the corresponding x_i value in the sample data. The proportion of 1s indicates the probability that the corresponding value was drawn from distribution 1, the lower mean distribution. The x values that could only be drawn from distribution 1 are those forming a horizontal line at z=1. The x values that could only be drawn from distribution 2, the higher mean distribution, are those forming a horizontal line at z=0. The x values that correspond to $0 < z < 1$ are those that could have been drawn from either distribution 1 or 2. There is a negative relationship between the proportion of 1s and the x values. As the x values increase, they become less likely to have come from distribution 1 and more likely to have come from distribution 2.

The mean of the distribution of m values (calculated as Y – 1) is 6.26. The mean of the distribution of p values is 0.33. Therefore, the predicted mean of distribution 2 is 6.26 with a 95% credible interval of (5.09, 7.50), and the predicted proportion of values drawn from distribution 1 is 0.33 with a 95% credible interval of (0.2256, 0.4507). Therefore, the predicted proportion of values drawn from distribution 2 is 0.69 with a 95% credible interval of (0.5493, 0.7744).