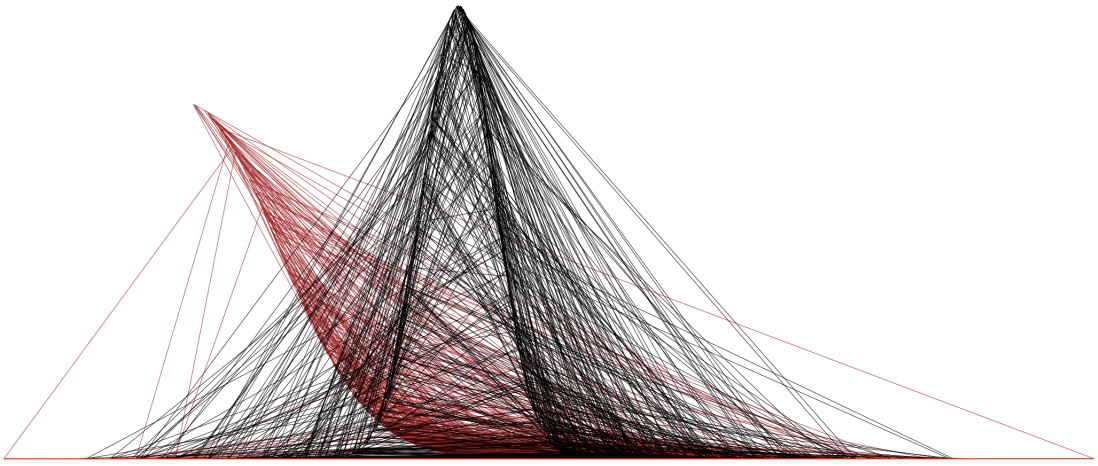
Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων

ΔΠΜΣ στις Εφαρμοσμενες Μαθηματικες Επιστημες

Μπεϋζιανή Στατιστική και MCMC



Μάιος 30, 2017

Φοιτητής:
Παντελής Καρατζας
email: pantelispanka@gmail.com

Διδάσκων:
ΔΗΜΗΤΡΗΣ ΦΟΥΣΚΑΚΗΣ
email: fouskakis@math.ntua.gr

Άσκηση 1

Τα δεδομένα της άσκησης μας δίνουν την διάρκεια ζωής (σε ώρες) 20 ηλεκτρονικών εξαρτημάτων.
Οι τιμές είναι οι παρακάτω:

$$60, 119, 100, 130, 43, 227, 23, 91, 128, 199, 85, 125, 40, 26, 141, 212, 238, 94, 111, 67$$

Αναλυτικός υπολογισμός ύστερης κατανομής

Για να υπολογίσουμε αναλυτικά την *posterior* κατανομή χρειαζόμαστε την συνάρτηση πιθανοφάνειας της κατανομής που ακολουθούν τα δεδομένα, (εδώ X $\sim N(\mu, \sigma^2)$ με άγνωστη μέση τιμή και τυπική απόκλιση $\sigma = 60$), καθώς και την prior. Γνωρίζουμε ότι η πρότερη πληροφορία μας δίνεται πάλι από μια Κανονική κατανομή με μέση τιμή $\mu = 0$ και διασπορά $\sigma^2 = 1000$, "περικομένη" στο διάστημα $(0, 500)$. Εφόσον και οι δυο κατανομές είναι της ίδιας οικογένειας (εκθετικές κατανομές) και άρα έχουμε μία *conjugateprior* κατανομή το αποτέλεσμα θα είναι μια κανονική κατανομή πάλι. Μένει να υπολογίσουμε τις μ και σ της *posterior* κατανομής.

Δεδομένα : $N(\mu, \sigma^2) : f(x_n \mid \mu.\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ - \frac{(x-\mu)^2}{2\sigma^2} \Big\} \quad n = 1, 2, ..., 20$

Likelihood : $N(\mu, \sigma^2) : L(X \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ - \frac{(x-\mu)^2}{2\sigma^2} \Big\} \quad using \qquad \tau = \frac{1}{\sigma^2}$

$$= \big(\tfrac{\tau}{2\pi}\big)^{\frac{n}{2}} exp\Big\{ - \tfrac{1}{2}\tau \textstyle\sum_1^n (x_n - \mu)^2 \Big\} = \big(\tfrac{\tau}{2\pi}\big)^{\frac{n}{2}} exp\Big\{ - \tfrac{1}{2}\tau \Big(\textstyle\sum_1^n (x_n - \bar{x})^2 - n(\bar{x} - \mu)^2 \Big) \Big\}$$

Prior : $$f(\mu) = \begin{cases} \frac{1}{\sigma_0\sqrt{2\pi}\big[\Phi((\beta-\mu_0)/\sigma_0) - \Phi((\alpha-\mu_0)/\sigma_0)\big]} \exp\Big\{ - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} \Big\} & \text{if } \alpha \le \mu \le \beta \\ 0 & \text{else.} \end{cases}$$

Posterior : $p(\mu \mid \mathrm{X}) \propto p(X \mid \mu) p(\mu) \propto const * exp\Big\{ - \frac{1}{2} \Big(\tau \Big(\sum_1^n (x_n - \bar{x})^2 - n(\bar{x} - \mu)^2 \Big) + \tau_0 \big(\mu - \mu_0\big)^2 \Big) \Big\}$

$$\propto exp\Big\{ - \tfrac{1}{2} \Big(n\tau(\bar{x} - \mu)^2 + \tau_0 \big(\mu - \mu_0\big)^2 \Big) \Big\}$$

$$= exp\Big\{ - \tfrac{1}{2}(n\tau + \tau_0) \Big(\mu - \tfrac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0} \Big)^2 + \tfrac{n\tau\tau_0}{n\tau + \tau_0} (\bar{x} - \mu_0) \Big\}$$

$$\propto exp\Big\{ - \tfrac{1}{2}(n\tau + \tau_0) \Big(\mu - \tfrac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0} \Big)^2 \Big\} \text{ if } \alpha \le \mu \le \beta$$

Από τον αναλυτικό υπολογισμό της ύστερης κατανομής έχουμε : $p(\mu \mid X) \sim N(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}, \frac{1}{n\tau + \tau_0})$. Ο μέσος των δεδομένων μας ακολουθεί κανονική κατανομή με παραμέτρους όπως έχουν υπολογιστεί. Για τον αναλυτικό υπολογισμό και αυτών έχουμε $\tau = \frac{1}{\sigma^2}$, άρα $t_0 = 0.001$ και $t = 0.0002777$. Με αντικατάσταση των τιμώς στους τύπους του μέσου και διασποράς πέρνουμε τις τιμές $\mu = 95.57308$ και $\sigma^2 = 153.8462$

Διαγράμματα Πιθανοφάνειας, πρότερης και ύστερης κατανομής

Παρακάτω εμφανίζεται ένα συγκεντρωτικό διάγραμμα.
Σε αυτό βλέπουμε το διάγραμμα των:

· **Διάγραμμα πιθανοφάνειας** (χρησιμοποιείται το κόκκινο χρώμα)

· **Διάγραμμα** *Prior* **κατανομής** (χρησιμοποιείται το καφέ χρώμα)

· **Διάγραμμα** *Posterior* **κατανομής** (χρησιμοποιείται το μαύρο χρώμα)

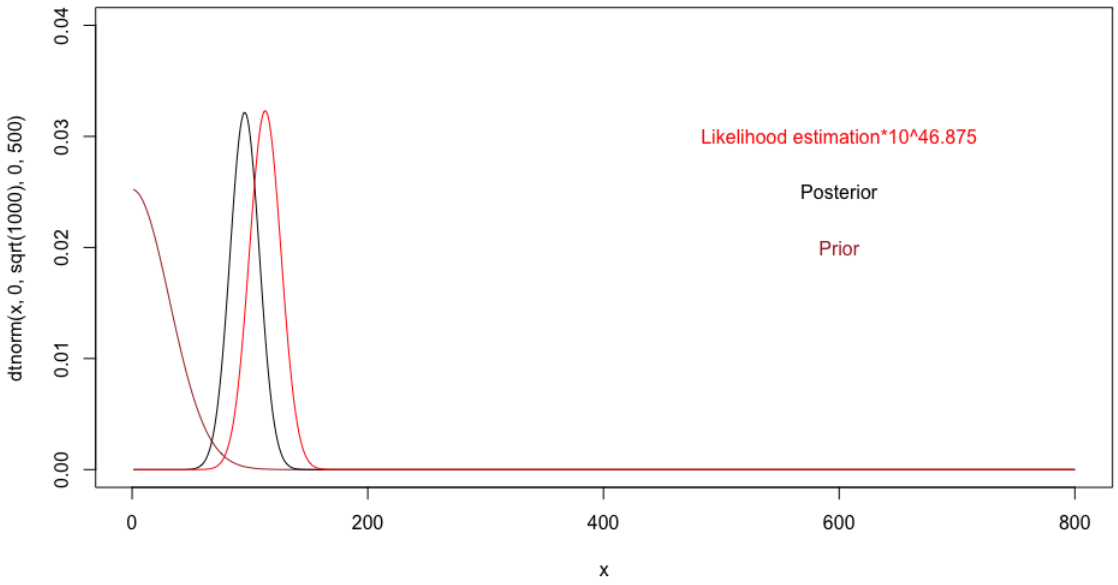


Figure 1: Διάγραμμα Πιθανοφάνειας, πρότερης και ύστερης κατανομής

Η σύγκριση που πρέπει να γίνει εδώ είναι αυτή των διαγραμμάτων εκτίμησης του μέσου με την μέθοδο της μέγιστης πιθανοφάνειας χωρίς την εισαγωγή "πρότερης" γνώσης στην διαδικασία μας και στο διάγραμμα της *posterior* κατανομής που δημιουργήθηκε με την εισαγωγή πρότερης γνώσης. Το τρίτο διάγραμμα της *prior* κατανομής αντιπροσωπεύει το κομμάτι που ως παλαιότερη πληροφορία είναι γνωστό και χρησιμοποιούμε για να φτιάξουμε την *posterior* κατανομή. Εύκολα μπορεί κανείς να διαπιστώσει το πώς αλλά και πόσο αυτή η πληροφορία επιδρά στον τρόπο υπολογισμού ενός μοντέλου για την μέση τιμή της διάρκειας ζωής των ηλεκτρονικών εξαρτημάτων που τώρα μελετάμε.

Για να καταφέρουμε να συγκρίνουμε το διάγραμμα της συνάρτησης πιθανοφάνειας που προκύπτει με τις άλλες δύο κατανομές πολλαπλασιάσαμε την likelihood με $10^{46.875}$ για να δημιουργήσουμε μια ίδια κλίμακα και να μπορέσει να προκύψει σύγκριση της μορφής των likelihood, prior, posterior

Η Posterior που προκύπτει έχει μικρότερη μέση τιμή από την εκτιμημένη με την μέθοδο της μέγιστης πιθανοφάνειας ενώ φαίνεται η τυπική απόκλιση νε μην επηρεάζεται κατα πολύ από την εισαγωγή της Prior καθότι η διασπορά της prior κατανομής είναι αρκετά μεγάλη ώστε να μην επηρεάζει την διασπορά της Posterior κατανομής.

Παρακάτω θα προχωρήσουμε και στην υλοποίηση του Αλγορίθμου Metropolis Hastings, που θα μας επιβεβαιώσει ή όχι την υπόθεσή μας.

Κώδικας που χρησιμοποιήθηκε

Για να δημιουργηθούν τα παραπάνω διαγράμματα και να υπολογιστούν οι απαραίτητες τιμές χρησιμοποιήσαμε τον παρακάτω κώδικα στην γλώσσα προγραμματισμού R.

```
#The Data of the problem
data <- c(60, 119, 100, 130, 43, 227, 23, 91,
          128, 199, 85, 125, 40, 26, 141, 212, 238, 94, 111, 67)

#Creating x values
x <- seq(1,800,by=1)

#Plotting Prior Distribution (truncated Normal N(0,1000) in 0-500)
plot(x, dtnorm(x,0,sqrt(1000), 0 , 500), type = "l", ylim = c(0,0.004))

#Adding the Posterior with mean = 95.573, variance = 153.8462 As calculated above
lines(x, dtnorm(x,95.573,sqrt(153.8462), 0 , 500), type = "l")

#Likelihood function of normal dist with known variance
likelihood.normal.mu <- function(mu, sig2, x) {
     # mu mean of normal distribution for given sig
     # sig2 variance of normall distribution
     # x vector of data
     n = length(x)
     a1 = (2*pi*sig2)^-(n/2)
     a2 = -1/(2*sig2)
     y = (x-mu)^2
     ans = a1*exp(a2*sum(y))
     return(ans)
 }


#Finding the standard deviance of the posterior
sqrt(153.8462)

#Intitialize new empty array for the results of the likelihood
likelihood_results <- c()

#Gathering results from the likelihood estimator
for (i in 1:length(x)){ likelihood_results[i] <-
                likelihood.normal.mu(mu = x[i], sig2=3600, data)}

#Plotting Likelihood , Posterior, Prior
plot(x, dtnorm(x,0,sqrt(1000), 0 , 500), type = "l", ylim = c(0,0.04), col = "brown")
lines(x, dtnorm(x,95.573,sqrt(153.8462), 0 , 500), type = "l")
lines(x, likelihood_results*10^46.875, type = "l", col = "red")

#Adding text for better visualization
text(x=600,y=0.03, label = "Likelihood estimation*10^46.875", col = "red")
text(x=600,y=0.025, label = "Posterior")
text(x=600,y=0.02, label = "Prior", col = "brown")
```

Metropolis Hastings Algorithm

Με την βοήθεια του αλγόριθμου Metropolis Hastings θα προσομοιάσουμε τιμές από την ύστερη κατανομή. Χρησιμοποιούμε αυτόν τον τρόπο όταν ο αναλυτικός υπολογισμός είναι δύσκολος λόγο της ύπαρξης του ολοκληρώματος που προκύπτει από τόν τύπο του Bayess: $\frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta}$. Με αυτόν τον αλγόριθμο δημιουργούμε μια αλυσίδα (Μαρκοβιανή αλυσίδα) που αποτελείται από τιμές της ύστερης κατανομής. Για να γίνει αυτό παίρνουμε τιμές από μια κατανομή εισήγησης και δοκιμάζουμε αν αυτές γίνονται αποδεκτές από την κατανομή που θέλουμε να προσεγγίσουμε. Έτσι βρίκουμε τιμές της ύστερης και άρα μπορούμε να προχωρήσουμε σε συμπεράσματα για αυτή.

Στο συγκεκριμένο πρόβλημα θεωρούμε ώς κατανομή εισήγησης μια κανονική κατανομή με μέση τιμή την τελευταία τιμή που έχει γίνει αποδεκτή στην αλυσίδα και τυπική απόκλιση τ. Για το αν θα γίνουν αποδεκτές οι τιμές που θα προκύπτουν χρησιμοποιούμε ως πιθανότητα αποδοχής το :

$$\alpha_{MH}(\theta^* \mid \theta_t, y) = min\left\{1, \frac{\frac{p(\theta^*|y)}{g(\theta^*|\theta_t,y)}}{\frac{p(\theta_t|y)}{g(\theta_t|\theta^*,y)}}\right\}$$

Συγκεκριμένα εδώ:

$$p_{acceptance} = posterior_{candidate} * suggestion_{candidate|last}/posterior_{proposal} * suggestion_{last|candidate}.$$

Σε συνέχεια με τον κώδικα που ήδη έχει παραχθεί για τον υπολογισμό της Likelihood παρατίθεται ο αλγορίθμος Metropolis Hastings:

```
metropolis.hastings <- function(iter, sigma){
    accept <- 0
    met <- numeric(iter)
    last<- 120
    sig2 <- sigma^2
    for(i in 1:iter){
        cand <- rnorm(1,last,sigma)
        alpha <- (likelihood.normal.mu(mean(data), sig2=153.8, cand)
            * dtnorm(cand, 0, sqrt(1000), 0, 500) * dnorm(last,cand,tau) )
            /(likelihood.normal.mu(mean(data), sig2=153.8, last)
            *dtnorm(last, 0, sqrt(1000), 0, 500)* dnorm(cand,last,tau) )
        if(runif(1) < min(alpha,1)){
            last<-cand
            accept <- accept+1
        }
        met[i] <- last
    }
    list(mean=met,accept=accept/iter)
}
```

Δέχεται σαν παράμετρους τον αριθμό των επαναλήψεων και την τυπική απόκλιση για την οποία χρησιμοποιούμε και διάφορες τιμές μέχρι να επιτύχουμε ποσοστό αποδοχής τιμών 25%. Μας επιστρέφει τις τιμές που έχουν γινει αποδεκτές και σε μορφή λίστας και το ποσοστό αποδοχής.

```
metropolis.run.1 <- metropolis.hastings(5000, 10)  # Acceptance Rate 73.44%
metropolis.run.2 <- metropolis.hastings(5000, 20)  # Acceptance Rate 54.34%
metropolis.run.3 <- metropolis.hastings(5000, 30)  # Acceptance Rate 42.54%
metropolis.run.4 <- metropolis.hastings(5000, 40)  # Acceptance Rate 34.16%
metropolis.run.5 <- metropolis.hastings(5000, 50)  # Acceptance Rate 27.88%
metropolis.run.6 <- metropolis.hastings(5000, 55)  # Acceptance Rate 25.4%
```

Επιτυγχάνουμε το επιθυμητό αποτέλεσμα για τυπική απόκλιση ανάμεσα στις τιμές περίπου 55.

Παρακάτω βλέπουμε το διάγραμμα της Μαρκοβιανής αλυσίδας που δημιουργήθηκε από τον αλγόριθμο:

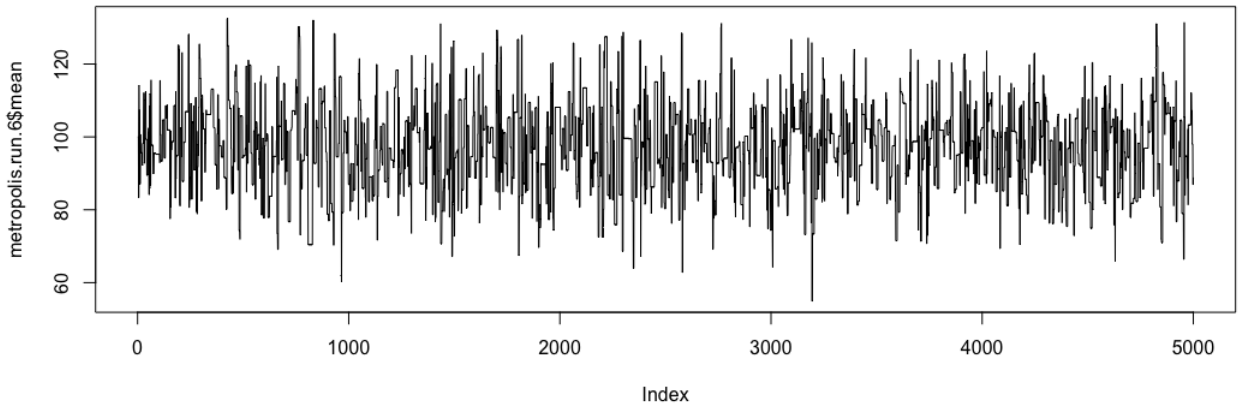


Figure 2: Διάγραμμα τιμών Μαρκοβιανής αλυσίδας

Φαίνεται πως η αλυσίδα κινείται κανονικά χωρίς πολλά στάσιμα σημεία ούτε και κάποια κλίση / κατεύθυνση. Έχοντας επιλέξει αρχική τιμη για την μέση τιμή 120 δεν είναι εύκολο να δεί κανείς την σύγκλιση μετά από κάποιο σημείο. Εδώ θεωρούμε αυτό το 1000οστό σημείο. Μετά από αυτό μπορούμε να πούμε με μεγαλύτερη σιγουριά πως η αλυσίδα έχει συγκλίνει και να λάβουμε υπόψη τις τιμές. Αυτό θα φαινόταν πιο καλά αν είχαμε επιλέξει κάποια άλλη αρχική τιμή πιο μακριά από την 120. Επιλέξαμε αυτή την μέση τιμή έχοντας κάποια ήδη πληροφορία για την ύστερη από τα αρχικά διαγράμματα. Αυτό θα καθόριζε και τον αριθμό των βημάτων που θα απαιτούνταν για την σύγκλιση της σειράς.

Το επόμενο διάγραμμα μας δίνει το ιστόγραμμα των τιμών της προσομοιομένης ύστερης κατανομής.

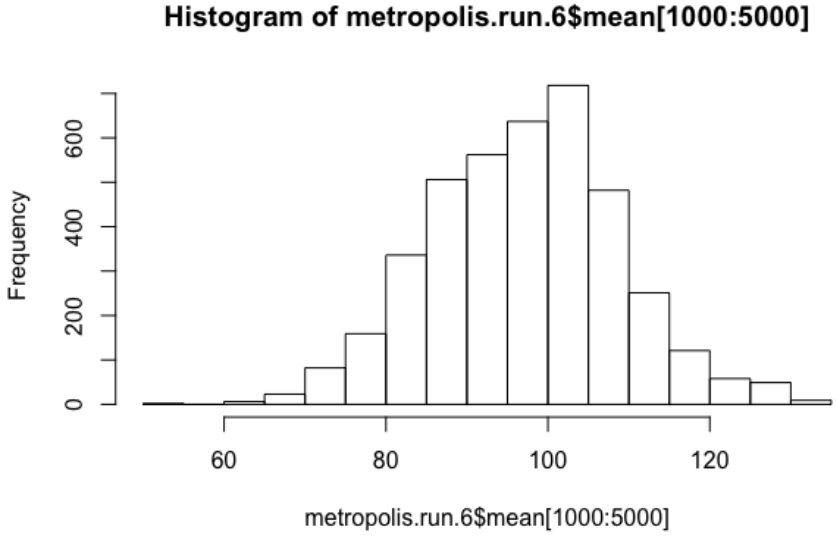


Figure 3: Ιστόγραμμα τιμών Μαρκοβιανής αλυσίδας

Παρακάτω (Figure 4) έχουμε το ιστόγραμμα των τιμών της αναλυτικά υπολογισμένης ύστερης κατανομής.

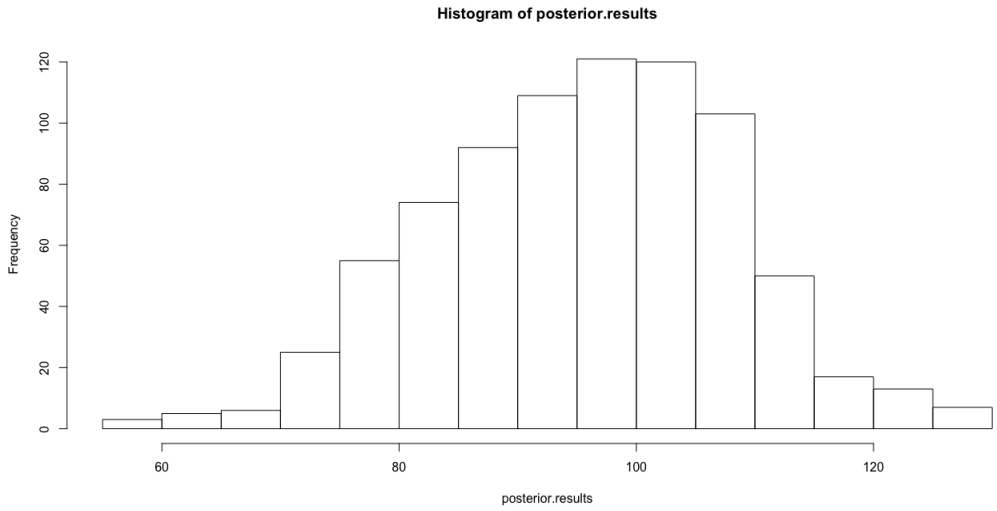


Figure 4: Ιστόγραμμα τιμών Posterior κατανομής

Παρακάτω γίνεται μια σύγκριση των τιμών που παίρνουμε από την θεωρητική προσέγγιση της posterior κατανομής και των προσομοιωμένων τιμών από τον αλγόριθμο MCMC (Figure 5).

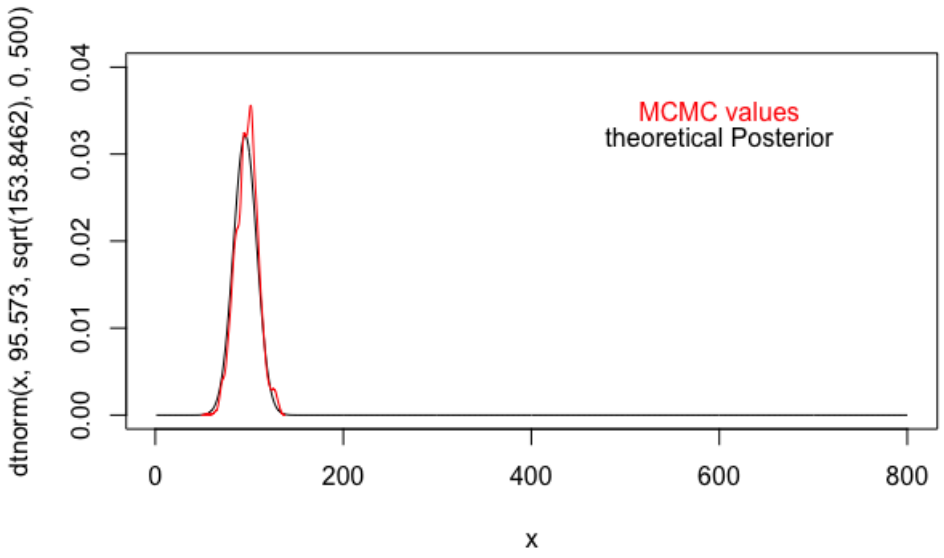


Figure 5: Σύγκριση τιμών Μαρκοβιανής αλυσίδας και θεωρητικής ύστερης

Από το διάγραμμα της αυτοσυσχέτισης (Figure 6), βλέπουμε πως δεν υπάρχει από κάποιο σημείο και μετά αυτοσυσχέτιση.

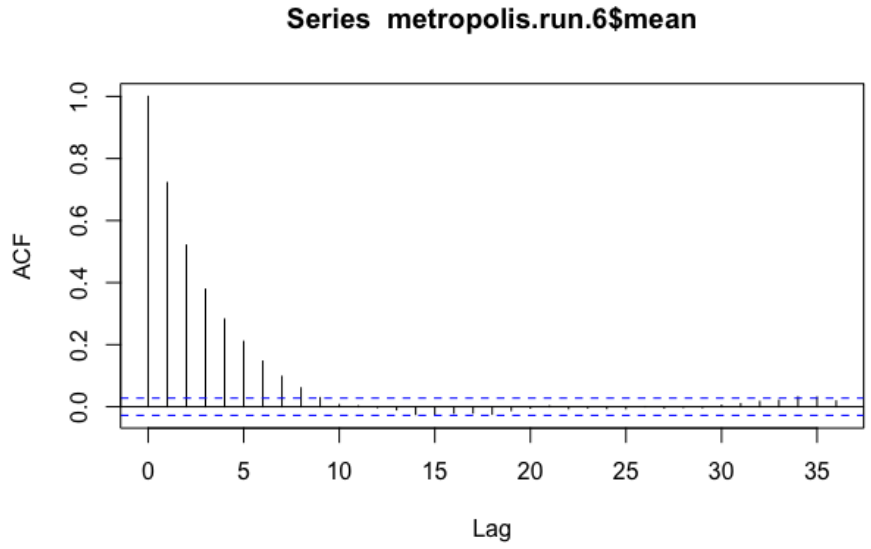


Figure 6: Ιστόγραμμα τιμών Μαρκοβιανής αλυσίδας

Για να δημιουργήσουμε το διάγραμμα του εργοδικού μέσου φτιάχνουμε την συνάρτηση που θα υπολογίσει τον εργοδικό μέσο:

```
erg.mean<-function(x){
        n<-length(x)
        result<-cumsum(x)/cumsum(rep(1,n))
        }
```

και με την αντίστοιχη εντολή *plot(erg.mean(metropolis.run.6$mean), type = "l")* βλέπουμε το διάγραμμα αυτού:

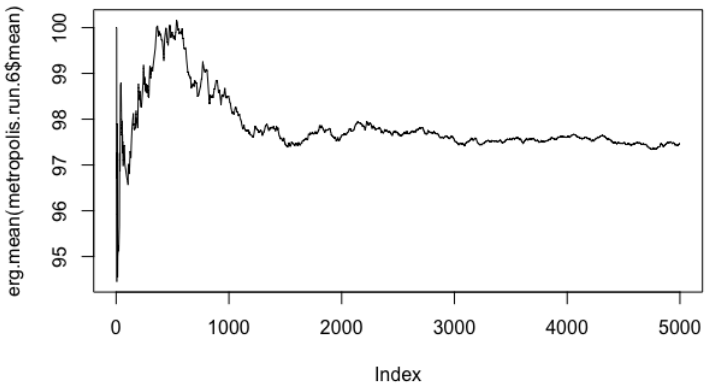


Figure 7: Διάγραμμα Εργοδικού μέσου

Όλα τα διαγράμματα μας δείχνουν πως η αλυσίδα έχει συγκλίνει χωρίς κάποιο πρόβλημα. Από το διάγραμμα του εργοδικού μέσου παρατηρούμε πως η αλυσίδα συγκλίνει από την 1000ωστή τιμή και μετά. Μπορούμε να την εμπιστευτούμε την αλυσίδα για να υπολογίσουμε την εκ των ύστερων μέση τιμή και διασπορά της κατανομής του μ.

Η μέση τιμή των προσομοιωμένων τιμών δίνεται με την εντολή :

```
mean(metropolis.run.6$mean[1000:5000])
```

και η αριθμητική τιμή της είναι 97.19635.

Η διασπορά των προσομοιωμένων τιμών δίνεται με την εντολή :

```
var(metropolis.run.6$mean[1000:5000])
```

και η αριθμητική τιμή της είναι 134.893.

Συνολικός κώδικας που χρησιμοποιήθηκε

```
metropolis.hastings <- function(iter, sigma){
     accept <- 0
     met <- numeric(iter)
     last<- 120
     sig2 <- sigma^2
     for(i in 1:iter){
          cand <- rnorm(1,last,sigma)
          alpha <- (likelihood.normal.mu(mean(data), sig2=153.8, cand)
                 * dtnorm(cand, 0, sqrt(1000), 0, 500) * dnorm(last,cand,tau) )
                 /(likelihood.normal.mu(mean(data), sig2=153.8, last)
                 *dtnorm(last, 0, sqrt(1000), 0, 500)* dnorm(cand,last,tau) )
          if(runif(1) < min(alpha,1)){
              last<-cand
              accept <- accept+1
          }
          met[i] <- last
     }
     list(mean=met,accept=accept/iter)
}

metropolis.run.1 <- metropolis.hastings(5000, 10)
metropolis.run.2 <- metropolis.hastings(5000, 20)
metropolis.run.3 <- metropolis.hastings(5000, 30)
metropolis.run.4 <- metropolis.hastings(5000, 40)
metropolis.run.5 <- metropolis.hastings(5000, 50)
metropolis.run.6 <- metropolis.hastings(5000, 60)


#Histogram of posterior values
posterior.results <- c()
posterior.results <- rtnorm(x,95.573,sqrt(153.8462))
hist(posterior.results)


#Plotting  Metropolis Hastings values
plot(metropolis.run.6$mean,type = "l")

hist(metropolis.run.6$mean[1000:5000])

#Computinh mean and variance of MCMC values
mean(metropolis.run.6$mean[1000:5000])
var(metropolis.run.6$mean[1000:5000])

#Comparing Theoritical / MCMC values
plot(x, dtnorm(x,95.573,sqrt(153.8462), 0 , 500), type = "l", ylim = c(0,0.04))
lines(density(metropolis.run.5$mean), col="red")
text(x=600,y=0.035, label = "MCMC values", col = "red")
text(x=600,y=0.032, label = "theoretical Posterior")
#Autocorellation plot
acf(metropolis.run.6$mean)

# compute ergodic mean
erg.mean<-function(x){
  n<-length(x)
  result<-cumsum(x)/cumsum(rep(1,n))
}

plot(erg.mean(metropolis.run.6$mean), type = "l")
```

MCMC winbugs

Αυτή την φορά θα χρησιμοποιήσουμε το WINBugs και θα τρέξουμε την προηγούμενη διαδικασία θεωρώντας επίσης άγνωστη την τυπική απόκλιση του πληθυσμού. Για *prior* κατανομή χρησιμοποιούμε μια *inverseGamma* μη πληροφοριακή.

Παρακάτω δίνεται ο κώδικας - μοντέλο που χρησιμοποιήθηκε στο winbugs για να μας δημιουργήσει την *posterior* κατανομή του μ.

Θέλουμε η διασπορά να ακολουθεί μία Inverse-Gamma κατανομή. Εφόσων το winbugs χρησιμοποιεί ακρίβεια και όχι διασπορά, συμπληρώνουμε: "var dgamma(0.001, 0.001)" (κανονική Gamma κατανομή και τότε η *likelihood* y[i] dnorm(mu,var) θα υπολογίσει τις τιμές σαν να έπερνε 1/var δηλαδή την inverse gamma.

```
model  trun1;
{

#prior distribution
mu~dnorm(0,0.001)I(0, 500)
#var <- 1/prec

var ~ dgamma(0.001, 0.001)

#Likelihood
for(i in 1:n) { y[i] ~ dnorm(mu, var) }


}
#Inits
list(mu = 10, var = 1)

#Data
list( y = c(60,119,100,130,43,227,23,91,128
        ,199,85,125,40,26,141,212,238,94,111,67),n=20)

}
```

Στο παρακάτω πινακάκι (table 1) εμφανίζονται τα βασικά στατιστικά στοιχεία των τιμώ που έχουν προσομοιωθεί με την βοήθεια του winbugs.

Έχουμε επιλέξει να κρατήσουμε τον αριθμό των "runs" στις 5000, χρησιμοποιώντας 1000 runs ώς burn in. Το αν οι αλυσίδες έχουν συγκλίνει καλά και άρα αν μπορούμε να χρησιμοποιήσουμε τα αποτελέσματα για εκτίμηση της Posterior θα το δούμε παρακάτω .

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu	89.24	16.38	0.2941	53.72	90.41	118.2	1001	4000
var	2.125E-4	7.68E-5	1.423E-6	9.003E-5	2.031E-4	3.892E-4	1001	4000

Table 1: Στατιστικά Αλυσίδων

Ήδη βλέπουμε κάποια ανησυχητικά μέτρα όπως το MC error που θα θέλαμε να είναι στο τρίτο δεκαδικό να είναι αρκετά μεγαλύτερο.

Στο Figure 10 εμφανίζεται το trace plot των μαρκοβιανών αλυσίδων όπως παρήχθησαν. Εδώ δεν βλέπουμε κάτι το ανησυχητικό όπως πιθανά trends ή στάσιμα σημεία.

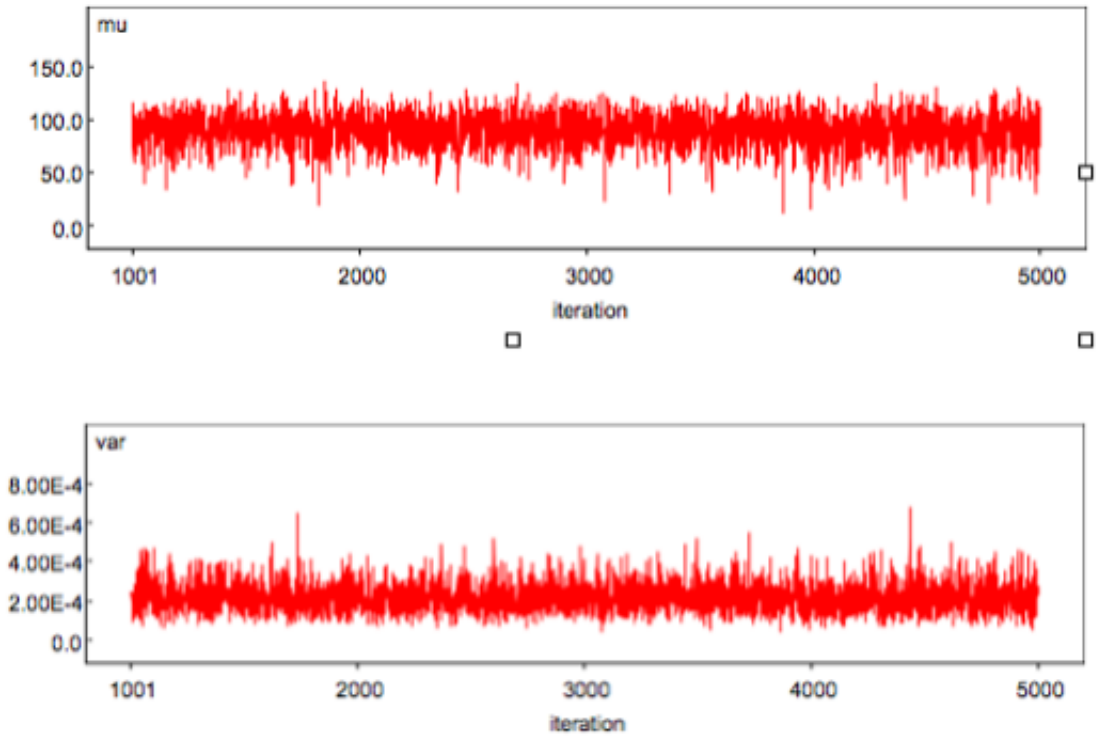


Figure 8: Μαρκοβιανή αλυσίδας για μέση τιμή και ακρίβειας της κατανομής του μ

Παρακάτω (Figure 11) βλέπουμε τα διαγράμματα της συνάρτησης πυκνότητας για την μέση τιμή και διασπορά, καθώς και της Αυτοσυχέτισης για τις αλυσίδες της μέσης τιμής και της διασποράς.

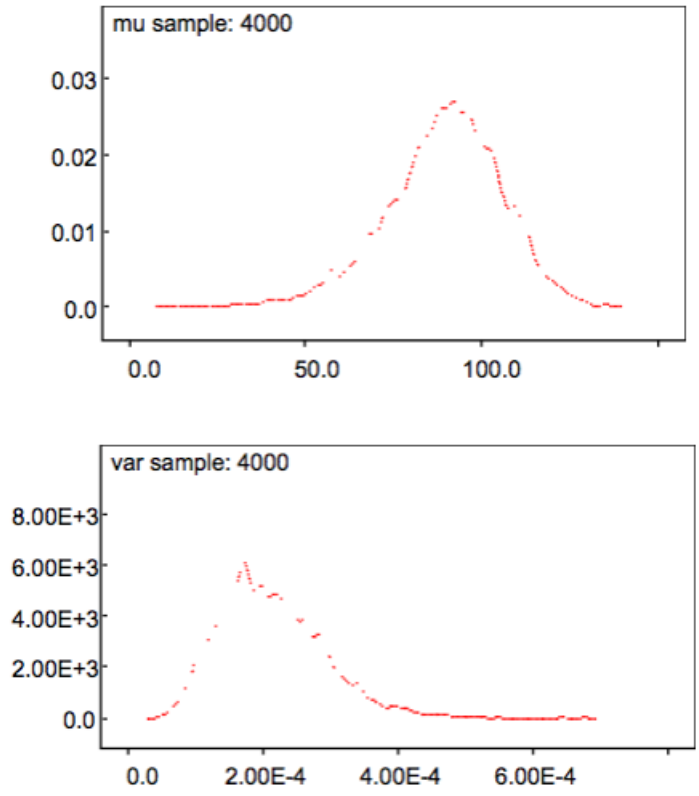


Figure 9: Density plots για την μέση τιμή και διασπορά της Posterior

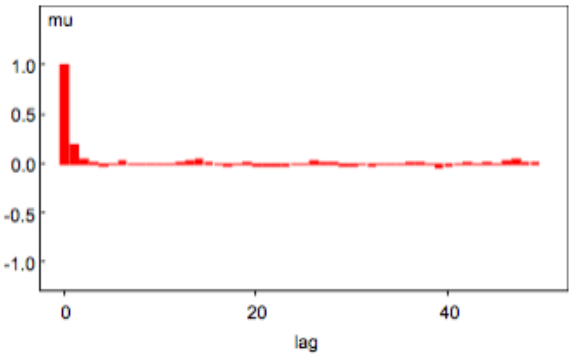


Figure 10: Διάγραμμα αυτοσυσχέτισης της αλυσίδας για την μέση τιμή

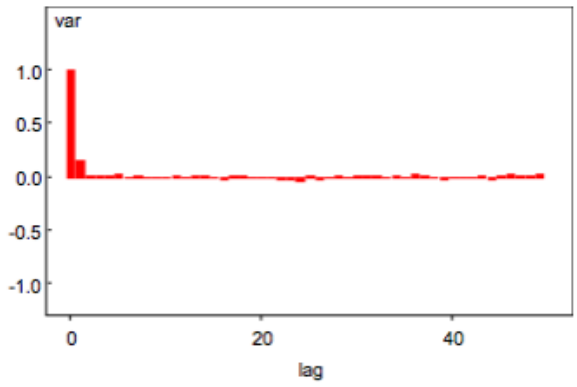


Figure 11: Διάγραμμα αυτοσυσχέτισης της αλυσίδας για την Διασπορά

Αντίστοιχα τα επόμενα πιο συμαντικά διαγράμματα αυτά των προσομοιωμένων τιμών (Dynamic trace):

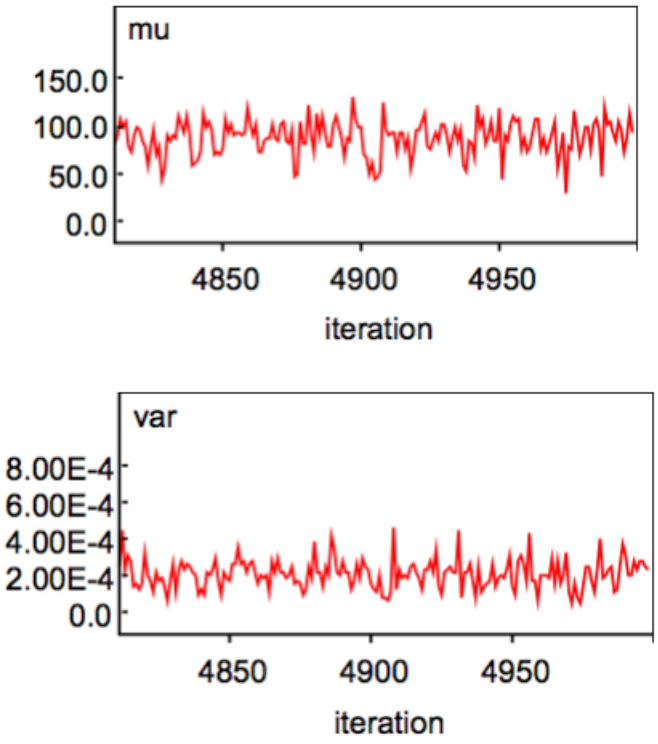


Figure 12: Διάγραμμα αυτοσυσχέτισης της αλυσίδας για την Διασπορά

Οι τιμές που προκύπτουν για την μέση τιμή και διασπορά της *Posterior* κανονικής κατανομής του μ είναι $mean = 89,94$ και διασπορά $\sigma^2 = 16.38^2$, $\sigma^2 = 268.3044$.

Στο συγκεκριμένο παράδειγμα έχουμε τρέξει την αλυσίδα 5000 φορές. Αποτέλεσμα αυτού είναι ένα αρκετά μεγάλο MC error. Η αλυσίδα δεν έχει συγκλίνει καλά και θα έπρεπε πιθανά να προχωρήσουμε σε διαγνωστκά τεστ για αυτή. Για να καταφέρουμε να πετύχουμε ένα MC error της στο τρίτο δεκαδικό στοιχείο (Δηλ MC error μικρότερο από 0.00) θα πρέπει να τρέξουμε την αλυσίδα σχεδών 100.000 φορές. Αυτό γίνεται λόγο της εισαγωγής της μη πληροφοριακής πρότερης η οποία έχει τεράστια διασπορά.

Άσκηση 2

Στο πρώτο κομμάτι της άσκησης θα αναλύσουμε τα δεδομένα και μόνο αυτα, χρησιμοποιώντας τα γενικευμένα γραμμικά μοντέλα και συγκεκριμένα το μοντέλο της λογιστικής παλλινδρόμησης. Στην συνέχεια θα χρησιμοποιήσουμε πρότερη πληροφορία και συγκεκριμένα αποτελέσματα παραδείγματος χάρη σε παλαιότερη μελέτη, και συγκεκριμένα το γεγονός ότι από 1725 ποντίκια χωρίς καμιά φαρμακευτική αγωγή παρατηρήθηκε όγκος σε 263 από αυτά.

Τα δεδομένα της τωρινής μελέτης είναι πώς σε 14 ποντίκα χωρίς φαρμακευτική αγωγή αναπτύχθηκε όγκος σε 4 από αυτά, σε 34 με 1 δόση συγκεκριμένου φαρμάκου παρατηρήθηκε όγκος σε 4 από αυτα και τέλος σε 34 με διπλή δόση παρατηρήθηκε όγκος σε 2 από αυτά.

Σκοπός της μελέτης είναι να δούμε με ποιόν τρόπο επιδρά το φάρμακο στην ανάπτυξη όγκου ή μη στον υπό μελέτη πληθυσμό.

Μελέτη απλής λογιστικής παλινδρόμησης

Στο πρώτο κομμάτι της ανάλυσης των δεδομένων χρησιμοποιούμε το μοντέλο της λογιστικής παλινδρόμησης, Ένα γενικευμένο γραμμικό μοντέλο δηλαδή με συνδετική συνάρτηση $log(p_i/(1-p_i) = \beta_0 + \beta_1 X_i))$. Πιο συγκεκριμένα θα δούμε το πώς επιδρά το φάρμακο και ο αριθμός δόσεων (Επεξηγηματική X) στην πιθανότητα (p_i) να αναπτύξει όγκο μια μονάδα του πληθυσμού i, πρός την πιθανότητα να μην αναπτύξει ($1-p_i$).

Πρώτα περνάμε τα δεδομένα της άσκησης σε ένα *Data.frame()* , μια δομή δεδομένων που κάνει εύκολη την ανάλυση αυτών με την χρήση της R.

```
Y <- rep(0,10)
Y <- append(Y, rep(1,4))
Y <- append(Y, rep(0,30))
Y <- append(Y, rep(1,4))
Y <- append(Y, rep(0,32))
Y <- append(Y, rep(1,2))
X <- rep(0,14)
X <- append(X, rep(1,34))
X <- append(X, rep(2,34))


rats.dataframe <- data.frame(Y,X)
```

Πιο συγκεκριμένα δημιουργούμε έναν πίνακα με δυο διανύσματα. Το πρώτο της τυχαίας μεταβλητής Y που αποκρίνεται στα ποντίκια ανα γκρούπ με τιμές 0 για αυτά χωρίς όγκο και 1 για αυτά με όγκο και το δεύτερο της επεξηγηματικής μεταβλητής X με τιμές ανάλογες της δόσης 0, 1, 2.

Στην συνέχεια προσαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης

```
classical.log <- glm(Y~X, data = rats.dataframe, family = binomial())
```

Για να δούμε τα δεδομένα του προσαρμοσμένου μοντέλου χρησιμοποιούμε την εντολή summary (classical.log) *Όπου classical.log το αντίστοιχα προσαρμοσμένο μοντέλο. Η R μας επιστρέφει τον παρακάτω πίνακα (table 2):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9767	0.5428	-1.80	0.0720
X	-0.9451	0.4778	-1.98	0.0479

Table 2: Πίνακας παραμέτρων μοντέλου λογιστικής παλινδρόμησης

Μία γρήγορη ματιά μας δείχνει πώς η κατηγορική μεταβλητή X είναι στατιστικά σημαντική, και πρέπει να ληφθεί υπόψη. Επίσης βλέπουμε ότι όσο αυξάνεται η τιμή της τυχαίας μεταβλητής πέφτει. Για να μπορέσουμε να κάνουμε μία καλύτερη ανάλυση αντικαθιστούμε στο μοντέλο τις τιμές και έχουμε

$$log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -0.9767 - 0.9451X$$

Μια πιο καθαρή άποψη για το πως μεταβάλεται ο λόγος $log\Big(\frac{\hat{p}}{1-\hat{p}}\Big)$ θα είχαμε αν υψώνοντας το αποτέλεσμα στην e. Ας λάβουμε υπόψη ότι έχουμε ένα ποντίκι το οποίο δεν λαμβάνει κάποια φαρμακευτική αγωγή τότε από το μοντέλο που έχει προκύψει θα είχαμε $log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -0.9767$ και άρα $\frac{\hat{p}}{1-\hat{p}} = e^{-0.9767} \Rightarrow \frac{\hat{p}}{1-\hat{p}} = 0.3765517$
$\Rightarrow \hat{p} = 0.2735471$, Δηλαδή η πιθανότητα να αναπτυχθεί όγκος σε ένα ποντίκι είναι 0.2735 % με βάση τα δεδομένα που έχουμε στην διάθεσή μας.

Στην συνέχεια υπολογίζουμε την τιμή του μοντέλου για X=1. $log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -1.9218 \Rightarrow \frac{\hat{p}}{1-\hat{p}} = 0.1463433 \Rightarrow$ και συνεπώς $\hat{p} = 0.127661$. Άρα η εκτιμώμενη πιθανότητα να αναπτύξει όγκο ένα ποντίκι που λαμβάνει μια δόση από το φάρμακο είναι 0.127661. Άρα μπορούμε να εξηγήσουμε με την βοήθεια των πιθανοτήτων με ποιόν τρόπο επιδρά το φάρμακο στον υπό μελέτη πληθυσμό.

Στα αποτελέσματα έχουμε χρησιμοποιήσει τα Odds. Για να χρησιμοποιήσουμε τα Odds ratios χρησιμοποιούμε

$\frac{log\Big(\frac{\hat{p}}{1-\hat{p}}\Big)}{log\Big(\frac{\hat{p}}{1-\hat{p}}\Big)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta 1} = 0.3886407$ Για μια δόση φαρμάκου και αντίστοιχα 0.1510416 για δύο δόσεις φαρμάκου.

Συνεπώς μπορούμε να πούμε ότι η πιθανότητα ένα ποντίκι να εμφανίσει όγκο αν λαμβάνει μια δόση του φαρμάκου μειώνεται κατά 2.573071 φορές και 6.620693 αν λαμβάνει 2 δόδεις φαρμάκου.

Εισαγωγή παλαιότερης γνώσης

Αν εξάγουμε αποτελέσματα από τα δεδομένα που έχουμε ήδη μπορούμε να να δούμε κάποιες αντιφάσεις. Σαν δεδομένο παίρνουμε ότι από τα 14 ποντίκια που δεν χορηγήθηκε φάρμακο στα 4 αναπτύχθηκε όγκος και αυτό μας δίνει ένα πολύ μεγάλο ποσοστό στον αριθμό των ποντικιών που βρίσκονται θετικά στην παρατήρηση της ύπαρξης κάποιου όγκου. Έτσι έχουμε πολύ μεγάλη πιθανότητα στην ανάπτυξη όγκου αν στον πληθυσμό δεν δωθεί κάποια φαρμακευτική αγωγή. Σχεδόν ένα στα τέσσερα ποντίκια θα αναπτύξει όγκο. Αυτό είναι κάτι που δεν μπορεί να είναι αποδεκτό κάτι που και διαισθητικά μπορεί να καταλάβει κανείς. Έτσι λοιπόν στην συνέχεια θα εισάγουμε πληροφορία στο μοντέλο η οποία θα μας δώσει μια πιο πλήρη εικόνα της πραγματικότητας.

Συγκεκριμένα , έχουμε πληροφορία πως από 1725 ποντίκια έχει παρατηρηθεί όγκος σε 263 από αυτά. Ήδη από αυτή την πληροφορία έχουμε πως το ποσοστό των ποντικιών με όγκο είναι περίπου 0,15% . Η μεταβλητή που θα φέρει την πληροφορία αυτή στην εκτίμηση του μοντέλου είναι η εκτίμηση του β_0 στο γενικευμένο γραμμικό μοντέλο.

Η κατανομή βήτα μπορούμε να πούμε ότι αντιπροσωπεύει μια κατανομή πιθανοτήτων. Αντιπροσωπεύει όλες τις πιθανές τιμές μιας πιθανότητας όταν δεν γνωρίζουμε ποιά είναι η πιθανότητα αυτή. Αν θέλαμε να επιλέξουμε μια μη πληροφοριακή Βήτα κατανομή θα επιλέγαμε την $Beta(0.5, 0.5)$. Με αυτόν τον τρόπο συγκεντρώνουμε τις τιμές της κατανομής γύρω από το 0 και 1. Αυτή θα ήταν μια καλή κατανομη για να εισάγουμε το μοντέλο μας αν δεν ξέραμε τίποτα για την πιθανότητα του να διαλέξουμε από έναν πληθυσμό μια μονάδα που θα είναι θετική στην παρατήρηση όγκου. Στο δικό μας πείραμα γνωρίζουμε ότι κατα περίπου 15 % ένα ποντίκι είναι θετικό στην παρατήρηση όγκου και αυτή την πληροφορία θα προσπαθήσουμε να εισάγουμε στο μοντέλο.

Παρακάτω δίνεται η γραφική παράσταση μιας μη πληροφοριακής Βήτα κατανομής (Figure 16):

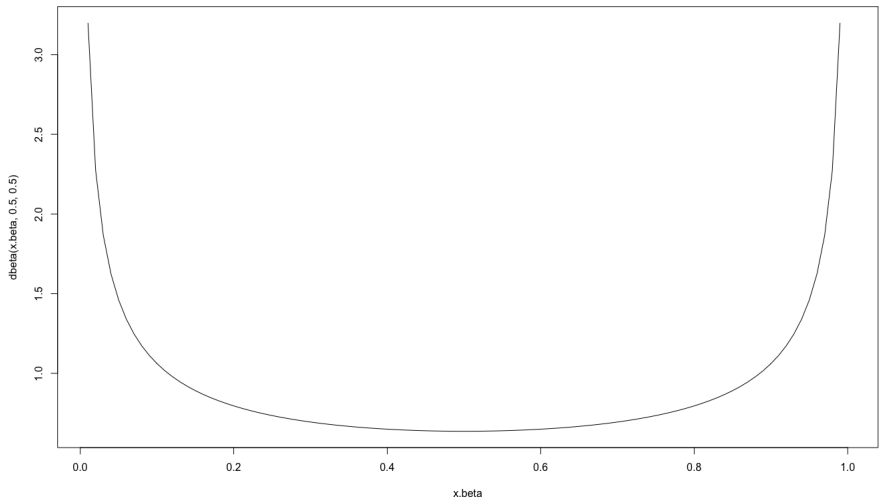


Figure 13: Μη πληροφοριακή κατανομή Βήτα

Στο δικό μας πρόβλημα πρέπει να δημιουργήσουμε μια κατανομή βήτα η οποία θα είναι πληροφοριακή. Ξέρουμε ένα ποια είναι η πιθανότητα ένα επιλεγμένο τυχαία ποντίκι να είναι θετικό ως πρός την εμφάνιση όγκου. Αυτή η πιθανότητα είναι περίπου 15 %. Θα πρέπει να βρούμε παραμέτρους της βήτα κατανομής που να περιέχουν αυτή την πληροφορία.

Οι παράμετροι α, β που θα επιλεχθούν για να αντιπροσωπεύσουν αυτή την πληροφορία είναι οι: $\alpha = 264, \beta = 1462$, $Beta(s+1, (n-s)-1)$. Παρακάτω δίνεται το γράφημα της συνάρτησης αυτής (figure 17) :

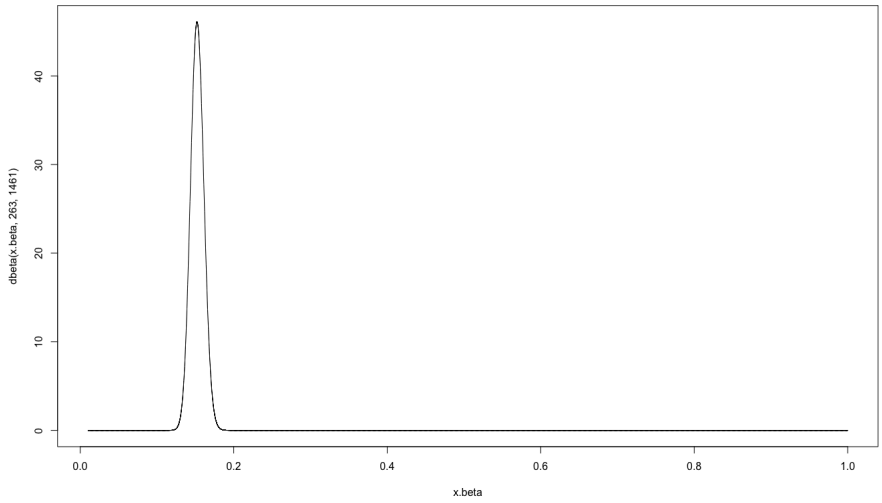


Figure 14: Πληροφοριακή κατανομή Βήτα ώς προς τα δεδομένα

Στην συνέχεια θα εισάγουμε την πληροφορία αυτή στο μοντέλο με την βοήθεια του Winbugs.
Ο κώδικας που χρησιμοποιήθηκε (μαζί με κάποιες αποτυχημένες προσπάθειες) :

```
model rats;
{

        for(i in 1:N){
        y[i] ~ dbin(p[i] , n[i])
        logit(p[i] ) <- beta0 + beta1*dose[i]
}

#beta0~dnorm(0,1.0E-6)
#beta0~dbeta(0.001,0.001)
#beta0~dbeta(264,1462)


p0 ~ dbeta(264,1462)
beta0 <- logit(p0)
beta1 ~ dnorm(0,1.0E-6)

}

#Initial
list(p0 = 0.5, beta1 = -0.2)

#data
list(N = 3, n = c(14,34,34), y=c(4,4,2), dose=c(0,1,2))
```

Καταρχήν βλέπουμε κάποιες δοκιμές για να δούμε το μοντέλο ($beta0$ $dnorm(0, 1.0E-6)$) οι οποίες έχουν σχολιαστεί και δεν παίζουν κανένα ρόλο στο τελικό μοντέλο. Επίσης λόγω της αδυναμίας του Winbugs να επιτύχει τιμές δοκιμάστηκαν διάφορες τιμές στην βήτα κατανομή ($beta0$ $dbeta(0.001, 0.001)$) οι οποίες επίσης δεν παίζουν κανένα ρόλο.

Τελικά βλέπουμε πως επιλέχθηκε ο μετασχηματισμός της βήτα με την βοήθεια της logit συνάρτησης για να καταφέρουμε να πάρουμε τιμές για την παράμετρο β_0 του μοντέλου. Η βήτα συνάρτηση παίρνει τιμές στο διάστημα (0,1) με τον μετασχηματισμό της Logit θα πάρουμε τιμές στο σύνολο R.

Παρακάτω βέπουμε τα στατιστικά των αλυσίδων των παραμέτρων του μοντέλου που μας ενδιαφέρουν (β_0, β_1), table3:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	-1.702	0.06726	7.289E-4	-1.835	-1.702	-1.573	1001	10000
beta1	-0.5418	0.3083	0.00274	-1.197	-0.5182	-7.927E-5	1001	10000

Table 3: Στατιστικά Αλυσίδων

Οι ίδιες οι αλυσίδες:

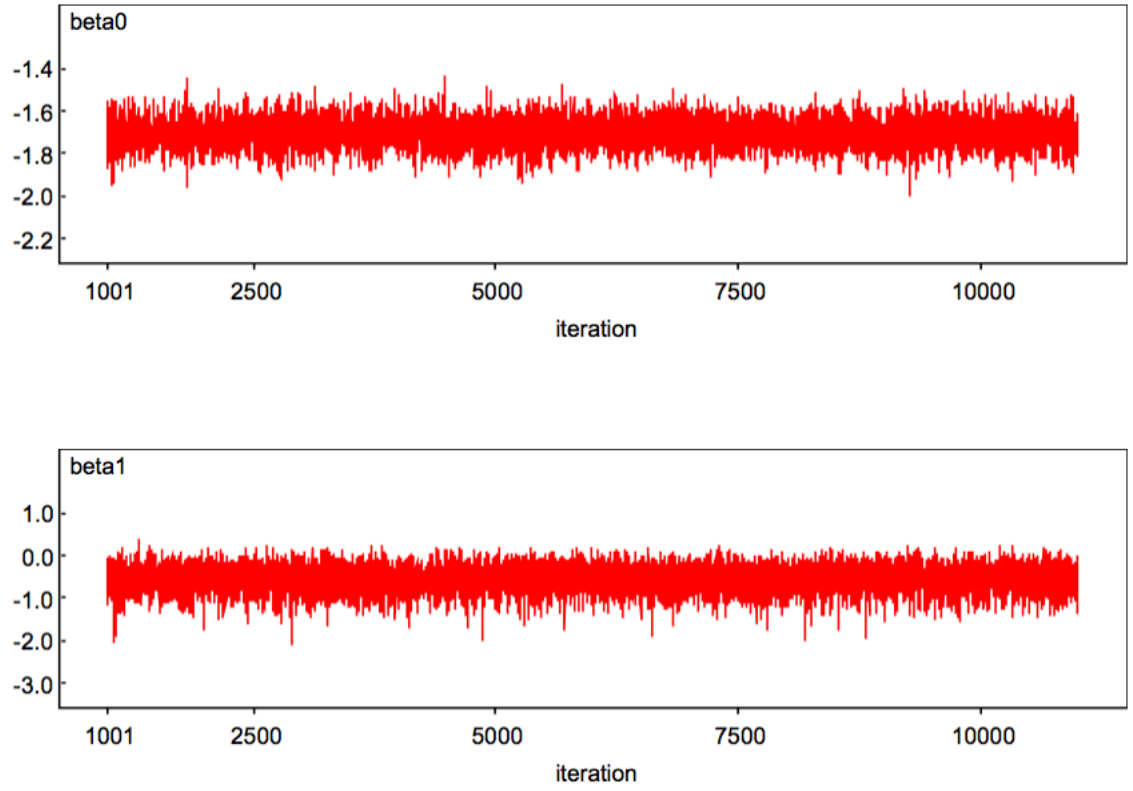


Figure 15: Αλυσίδα τιμών β_0, β_1

Καθώς και τα διαγράμματα πυκνότητας και αυτοσυσχέτισης των τιμών και των αλυσίδων αντίστοιχα.

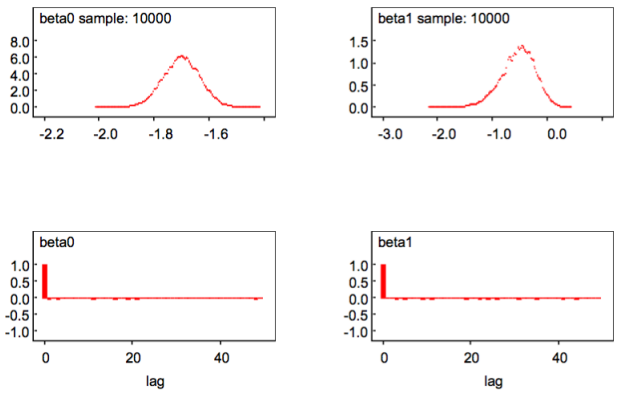


Figure 16: Διαγράμματα πυκνότητας και αυτοσυσχέτισης

Πλέον μπορούμε να προχωρήσουμε σε συμπεράσματα με βάση το νέο μοντέλο που δημιουργήθηκε με την εισαγωγή της πρότερης γνώσης σε αυτό.

$$log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -1.702 - 0.5418X$$

Για να δούμε την πιθανότητα να δημιουργηθεί όγκος σε ένα ποντίκι που δεν λαμβάνει φαρμακευτική αγωγή με τον ίδιο τρόπο όπως και πρίν θέτουμε όπου X = 1 και θα έχουμε $log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -1.702 \Rightarrow \frac{\hat{p}}{1-\hat{p}} = e^{-1.702} = 0.1823185$ $\Rightarrow \hat{p} = 0.1542042$.

Άρα η πιθανότητα να εμφανιστεί όγκος σε ένα ποντίκι που δεν λαμβάνει φαρμακευτικη αγωγή είναι 0.1542042 , που είναι κατα λίγο αυξημένη από την πιθανότητα με βάση τα αρχικά μας στοιχεία, 0.1524638. Αυτό συμβαίνει γιατί στα τωρινά στοιχεία έχουμε πολύ μεγαλύτερο ποσοστό ποντικιών που εμφάνισαν όγκο και άρα αυξάνεται κατα κάποιο ποσό η συγκεκριμένη πιθανότητα.

Στην συνέχεια θέτουμε το X = 1 για να δούμε πως θα αλλάξουν τα *Odds* . Συγκεκριμένα θα δούμε την πιθανότητα εμφάνισης όγκου στα ποντίκια αν αυτά λαμβάνουν 1 δόση από το φάρμακο. Έχουμε : $log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -1.702 - 0.5418 \Rightarrow \frac{\hat{p}}{1-\hat{p}} = 0.1060547 \Rightarrow \hat{p} = 0.09588621$. Η πιθανότητα να αναπτύξει όγκο ένα ποντίκι που λαμβάνει μια δόση φαρμακευτικής αγωγής είναι 0.09588621. Μειωμένη κατά 0.05657759 από το αν δεν λάμβανε καμιά φαρμακευτική αγωγή. Αντίστοιχα μπορούμε να δούμε και για X = 2 δόσεις και ούτω καθ' εξής.

Μια καλύτερη εκτίμηση για την επίδραση του φαρμάκου θα έχουμε αν υπολογίσουμε το *Odds Ratio*, που όπως και πριν θα μας δώσει την εκτίμηση της επίδρασης μόνο του φαρμάκου στον πληθυσμό. Αντίστοιχα όπως και πρίν θα έχουμε:

$\frac{log\Big(\frac{\hat{p}}{1-\hat{p}}\Big)}{log\Big(\frac{\hat{p}}{1-\hat{p}}\Big)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta 1} = 0.5817002$, Για μια δόση φαρμάκου και αντίστοιχα 0.3383752 για δύο δόσεις φαρμάκου.

Σε αυτό το παράδειγμα είδαμε πόσο πιο πραγματικά μπορεί να κάνει τα αποτελέσματα η εισαγωγή μιας πρότερης γνώσης. Τα αποτελέσματα με την εισαγωγή αυτής στο μοντέλο είναι πολύ πιο κοντά στην διαισθητική προσέγγιση του αριθμού των ποντικιών που μπορεί να αναπτύξουν όγκο και στον υπολογισμό της πιθανότητας να συμβεί αυτό. Ταυτόχρονα έχουμε μια καλύτερη εκτίμηση του πώς το φάρμακο επιδρά στον πληθυσμό.

Στην πρώτη περίπτωση η επίδραση του φαρμάκου στον πληθυσμό φαίνεται να είναι αρκετά ισχυρή και να μειώνει κατά πολύ περισσότερο την πιθανότητα να αναπτύσσει κάποιο είδος όγκου ένα ποντίκι (συγκεκριμένα σχεδόν κατα 70%). Με την εισαγωγή της πρότερης πληροφορίας αυτό το ποσοστό αλλάζει και είναι σχεδόν 50%, και έχουμε και μια καλύτερη εικόνα ως πρός την πιθανότητα να αναπτύξει κάποιο είδους όγκου ένα ποντίκι αν δεν λαμβάνει καμιά φαρμακευτική αγωγή,

Σύνολο κώδικα που χρησιμοποιήθηκε

```
#Generating the dataset
Y <- rep(0,10)
Y <- append(Y,rep(1,4))
Y <- append(Y,rep(0,30))
Y <- append(Y,rep(1,4))
Y <- append(Y,rep(0,32))
Y <- append(Y,rep(1,2))
X <- rep(0,14)
X <- append(X,rep(1,34))
X <- append(X,rep(2,34))


rats.dataframe <- data.frame(Y,X)

#Running a GLM
classical.log <- glm(Y~X, data = rats.dataframe, family = binomial())
summary(classical.log)


#Non Inf Beta
x.beta <- seq(0.01,1,by=0.01)
plot(x.beta, dbeta(x.beta, 0.5,0.5))


#Inf Beta
plot(x.beta, dbeta(x.beta, 263, 1461), type = "l")


#Winbugs
model rats;
{
        for(i in 1:N){
        y[i] ~ dbin(p[i] , n[i])
        logit(p[i] ) <- beta0 + beta1*dose[i]
}

#beta0~dnorm(0,1.0E-6)
#beta0~dbeta(0.001,0.001)
#beta0~dbeta(264,1462)

p0 ~ dbeta(264,1462)
beta0 <- logit(p0)
beta1 ~ dnorm(0,1.0E-6)
}

#Initial
list(p0 = 0.5, beta1 = -0.2)

#data
list(N = 3, n = c(14,34,34), y=c(4,4,2), dose=c(0,1,2))
```

Άσκηση 3

Τα δεδομένα της άσκησης αφορούν τον αριθμό των πλοίων που φθάνουν στο λιμάνι του Πειραιά σε 20 τυχαίες επιλεγμένες μέρες και είναι τα παρακάτω:

$$9,4,5,5,7,13,8,3,6,5,4,5,10,5,5,4,3,3,4,7$$

Τα δεδομένα ακολουθούν κατανομή Poisson δεδομένο ότι μιλάμε για " επιτυχίες " σε χρονικό διάστημα (εδώ πλοία ανά ημέρα). Θα προσπαθήσουμε λοιπόν να εκτιμήσουμε το λ της κατανομής που θα περιγράφει τα δεδομένα, με δύο τρόπους. Πρώτα με την κλασική στατιστική και έπειτα με την Μπευζιανή προσέγγιση και θα συγκρίνουμε τα δύο αποτελέσματα.

Εκτίμηση λ με κλασική στατιστική

Ξέρουμε ότι τα δεδομένα ακολουθούν την Poisson κατανομή και θέλουμε να εκτιμήσουμε το λ με την κλασική στατιστική. Για να γίνει αυτό πρέπει να φτιάξουμε την συνάρτηση πιθανοφάνειας του λ .

$$L(\lambda) = \prod_1^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod X_i!}$$

Στην συνέχεια λογαριθμίζουμε, παραγωγίζουμε ως προς λ και θέτουμε ως πρός 0. Έτσι θα βρούμε την τιμή του λ που θα μεγιστοποιεί την συνάρτηση πιθανοφάνειας.

$$l(\lambda) = -n\lambda + \sum_1^n x_i log(\lambda) - \sum_i^n log(x_i!)$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = 0 \implies -n + \frac{1}{\lambda} \sum_1^n X_i = 0 \implies \hat{\lambda} = \frac{1}{n} \sum_1^n X_i = \bar{X}$$

Άρα για την συγκεκριμένη περίπτωση $\lambda = \bar{X} = 5.75$

Παρακάτω (Figure 16) δίνεται το διάγραμμα της συνάρτησης πυκνότητας πιθανότητας της Poisson με παράμετρο Λάμδα 5.57 όπως αυτή υπολογίστηκε με την μέθοδο της μέγιστης πιθανοφάνειας.

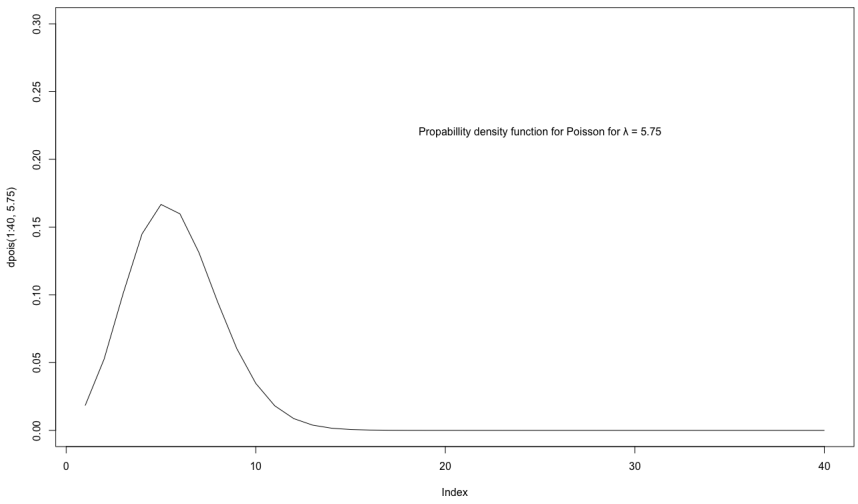


Figure 17: Διάγραμμα Πυκνότητας πιθανότητας για Poisson με $\lambda = 5.75$

Θα πρέπει να προχωρήσουμε στην κατασκευή διαστημάτων εμπιστοσύνης για την παράμετρο λ. Το ακριβές διάστημα εμπιστοσύνης δίνεται από τον τύπο : $\alpha = \bar{x} - t_{n-1,a}s/\sqrt{n}, \beta = \bar{x} + t_{n-1,a}s/\sqrt{n}$, η τιμή εδώ για την $t_{19,0.05}$, είναι 1.729. Από τα δεδομένα έχουμε $\bar{x} = 5.75$, και η διασπορά 6.723684, άρα $S = 2.593007$. Έτσι βρίσκουμε $\alpha = \bar{x} - 1.001919$ και το άνω όριο $\beta = \bar{x} + 1.001919$. 'Αρα το διάστημα εμπιστοσύνης, με συντελεστή εμπιστοσύνης 95% για το λ, είναι $(\alpha, \beta) = (4.748081, 6.751919)$. Αυτό σημαίνει πως αν συνεχίζουμε να κατασκευάζουμε διαστήματα εμπιστοσύνης της λ με ανεξάρτητα δείγματα μεγέθους 20 από αριθμο πλοίων που καταφθάνουν στο λιμάνι το 95% αυτών αναμένεται να έχουν τιμή λ στο διάστημα $(4.748081, 6.751919)$.

Υπολογισμός κατανομής με εισαγωγή Prior

Στην συνέχεια υπολογίζουμε την κατανομή Poisson $P(\lambda)$ χρησιμοποιώντας ως πρότερη κατανομή για το λ μια $Gamma(\alpha, \beta)$, μη πληροφοριακή.

Αναλυτικός υπολογισμός ύστερης προβεπτικής κατανομής

Για να γίνει ο υπολογισμός της ύστερης χρειαζόμαστε την συνάρτηση πιθανοφάνειας της Poisson, όπως αυτή έχει υπολογιστεί προηγούμενα και την Prior, εδώ μια $Gamma(\alpha, \beta)$.

$$Likelihood: \qquad L(\lambda) = \prod_1^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\prod X_i!}$$

$$Prior: \qquad G(\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \qquad x \geq 0 \quad \alpha, \beta > 0$$

Η Gamma prior είναι μια conjugate prior της Poisson κατανομής, άρα σαν αποτέλεσμα περιμένουμε μια κατανομή της ίδιας οικογένειας.

$$\begin{aligned} Posterior: \qquad p(\lambda \mid x) &\propto p(\lambda)f(x \mid \lambda) \propto \frac{\beta^\alpha \lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)} * e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\prod X_i!} \\ &\propto \quad \lambda^{\alpha+\sum x_i - 1} * e^{-\beta\lambda - n\lambda} = \lambda^{\sum x_i + \alpha - 1} * e^{-\lambda(\beta+n)} \end{aligned}$$

'Αρα η Posterior κατανομή είναι πάλι μια Γάμμα (όπως και περιμέναμε) με παραμέτρους $\left(\sum x_i + \alpha, n + \beta\right)$. $p(\lambda \mid x) \propto Gamma(115 + 0.001, 20 + 0.001) = Gamma(115.001, 20.001)$

Για να υπολογίσουμε την προβλεπτική ύστερη κατανομή θα πρέπει να ολοκληρώσουμε. Θεωρούμε x_{new} τον αριθμό των πλοίων που θα καταφθάσουν την επόμενη μέρα στο λιμάνι.

Υπολογισμός Προβλεπτικής ύστερης κατανομής

$$\begin{aligned} p(x_{new} \mid x) &= \frac{(n+\beta)^{\sum x_i + \alpha}}{\Gamma(\sum x_i + \alpha) * \Gamma(x_{new}+1)} \int_0^\infty \lambda^{x_{new} + \sum x_i + \alpha + 1} e^{-(n+\beta+1)\lambda} \partial\lambda \\ &= \frac{(n+\beta)^{\sum x_i + \alpha}}{\Gamma(\sum (x_i + \alpha))\Gamma(x_{new}+1)} \frac{\Gamma(x_{new} + \sum x_i + \alpha)}{(n+\beta+1)^{x_{new} + \sum x_i + \alpha}} \\ &= \frac{\Gamma(x_{new} + \sum x_i + \alpha)}{\Gamma(\sum (x_i + \alpha))\Gamma(x_{new}+1)} \left(\frac{n+\beta}{n+\beta+1}\right)^{\sum x_i + \alpha} \left(\frac{1}{n+\beta+1}\right)^{x_{new}} \end{aligned}$$

Συγκρίνοντας με την συνάρτηση πυκνότητας πιθανότητας της Αρνητικής διωνυμικής:

$$\Gamma(x+n)/(\Gamma(n)x!)p^n(1-p)^x$$

Μπορούμε να πούμε πως η προβλεπτική ύστερη κατανομή ακολυθεί μία αρνητική διωνυμική με μέση τιμή $\mu = \frac{\sum x_i \alpha}{n+\beta}$ και διασπορά $\sigma^2 = \frac{\sum x_i + \alpha}{(n+\beta)^2}(n + \beta + 1)$

Παρακάτω (Figure 18) παρατίθεται το το συγκριτικό διάγραμμα ανάμεσα στην Predictive Posterior κατανομή και στην Poisson με παράμετρο λ, όπως υπολογίστηκε με την μέθοδο της μέγιστης πιθανοφάνειας.

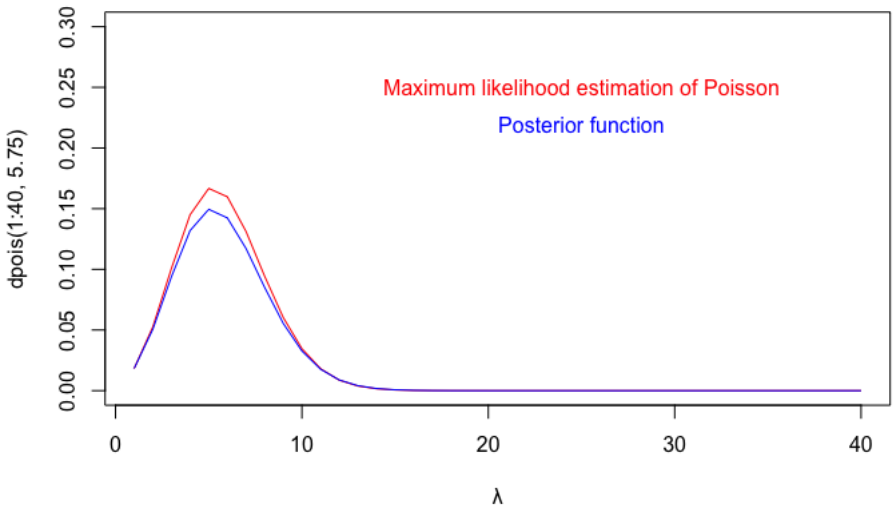


Figure 18: Συγκριτικό Διάγραμμα Posterior / Μέγιστης πιθανοφάνειας

Μία διαπίστωση που εύκολα κάνει κανείς είναι η μειωμένη τυπική απόκλιση της κατανομής που δημιουργήθηκε χρησιμοποιώντας την Prior Gamma. Ο λόγος που η διασπορά είναι μικρότερη είναι λόγω της μη πληροφοριακής Prior που χρησιμοποιήθηκε. Η πολύ μεγάλη διασπορά των τιμών (πλήρη άγνοια) έδωσε αντίστοιχη αβεβαιότητα στην Predictive Posterior κατανομή που προκύπτει.

Παρακάτω δίνεται μία πιό καθαρή εικόνα της Posterior κατανομής.

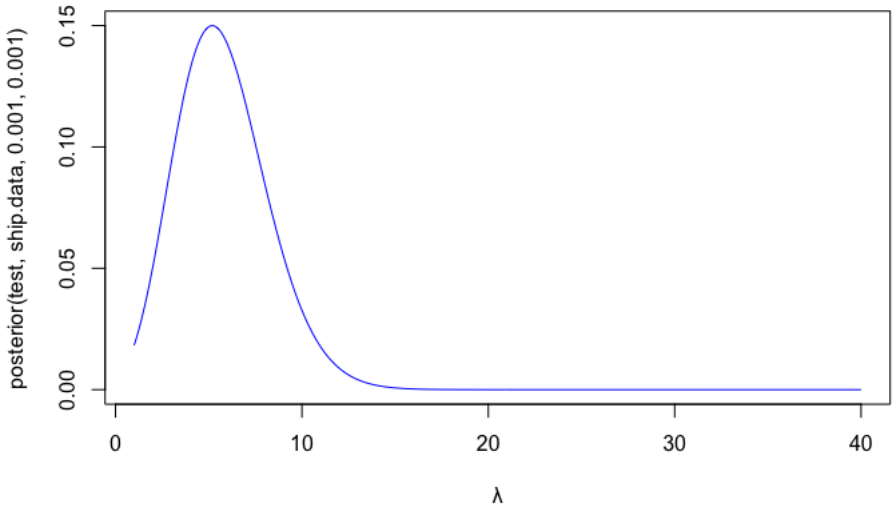


Figure 19: Συγκριτικό Διάγραμμα Posterior / Μέγιστης πιθανοφάνειας

Η διαδικασία υπολογισμού της Predictive Posterior πέρα από την εισαγωγή παλαιότερης γνώσης στο μοντέλο μας δίνει την δυνατότητα να εξάγουμε πιο εύκολα συμπεράσματα για τα διαστήματα εμπιστοσύνης της Παραμέτρου λ . Πλέον η παράμετρος λ είναι μια κατανομή από μόνη της και αυτό μας δίνει την δυνατότητα να πουμε ότι το 95% των τιμών της παραμέτρου είναι στο διάστημα (α, β) απλά κόβοντας ένα 2,5 % από το αριστερό και δεξί μέρος της ύστερης κατανομής δημιουργήσαμε για την παράμετρο λ. Με αυτό τον τρόπο δημιουργούμε ένα διάστημα εμπιστοσύνης 95%!.

Έχουμε δείξει ότι η Posterior κατανομή για το λ είναι μια $Gamma(115.001, 20.001)$. Πιο συγκεκριμένα δίνεται στο παρακάτω σχεδιάγραμμα:

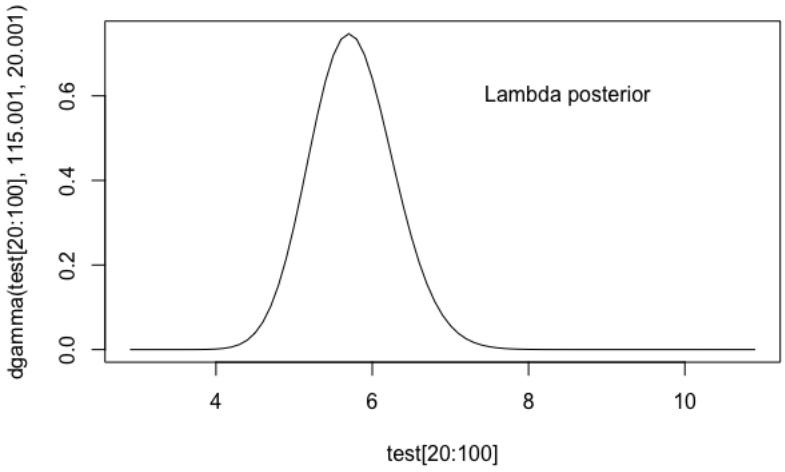


Figure 20: Posterior κατανομή της λ

Με την βοήθεια της R υπολογίζουμε το 95% διάστημα αυτής της κατανομής. Παρακάτω δίνεται το διάγραμμα του διαστήματος εμπιστοσύνης. Οι τιμές του είναι $(\alpha, \beta) = (4.7470, 6.8471)$.

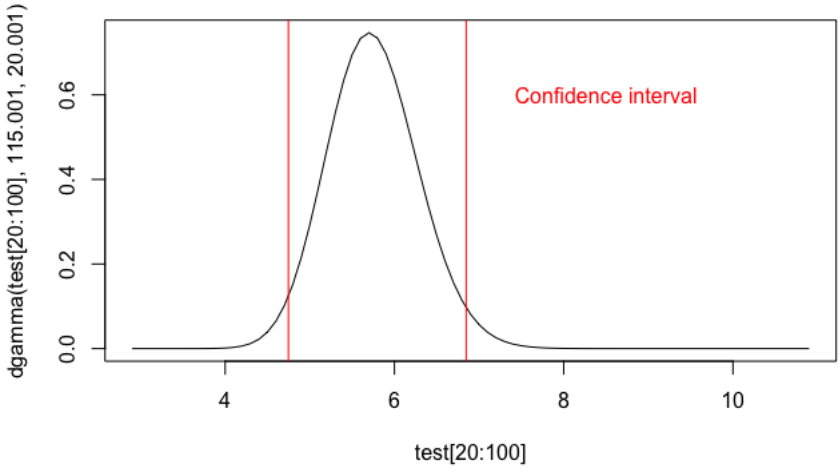


Figure 21: Διάστημα εμπιστοσύνης

Κώδικας που χρησιμοποιήθηκε

```
#The data of the problem
ship.data <- c(9,4,5,5,7,13,8,3,6,5,4,5,10,5,5,4,3,3,4,7)

#Ploting poisson pdf for 5.75 lambda
plot(dpois(1:40,5.75), type = "l", ylim = c(0,0.3))

#Poisson likelihood
poisson.likelihood <- function(lamda,data){
  exp(-lamda)*sum( (lamda^data)/factorial(data) )
}

#Empty plot to add intervals
plot(NULL, xlim=c(0,50), ylim=c(0,60), ylab="theta values",
                 xlab="Confidence inerval")

#Confidence interval for ship data
for(i in 1:length(ship.data)){
  conf.int <- poisson.test(ship.data[i], conf.level = 0.95)
  segments(x0 = conf.int$conf.int[1], y0 = ship.data[i],
  x1 = conf.int$conf.int[2], y1 = ship.data[i])
}

#Confidence interval for lambda 1 to 40
for(i in 1:40){
  conf.int <- poisson.test(i, conf.level = 0.95)
  segments(x0 = conf.int$conf.int[1], y0 = i, x1 = conf.int$conf.int[2], y1 = i)
}

#Adding estimated lambda value
segments(x0 = 5.75, y0 = 0, x1 = 5.75, y1 = 40)

#Gennerating the posterior function
posterior <- function(xnew,data,a,b){
  (gamma(sum(data)+a+xnew) / (gamma(sum(data + a))*gamma(xnew + 1) )) *
        ((length(data)+b)/(length(data) + b + 1))^(sum(data)+a)
        *(1/ (length(data) + b + 1)^xnew)
}

#Plot both Likelihood estimation of lambda and posterior
plot(dpois(1:40,5.75), type = "l", ylim = c(0,0.3), col="red")
lines(1:40, posterior(1:40,ship.data,0.001,0.001), type = "l", col = "blue")
text(x=25,y=0.25, label = "Maximum likelihood estimation of Poisson", col = "red")
text(x=25,y=0.22, label = "Posterior function", col ="blue")

#Clear Plot of the posterior
plot(seq(1,40, by=0.1), posterior(seq(1,40, by=0.1),ship.data,0.001,0.001),
        type = "l", col = "blue", xlab = "lambda")

#Confidence interval
plot(test[20:100], dgamma(test[20:100],115.001,20.001), type = "l")
text(x=8.5,y=0.6,label = "Confidence interval", col="red")
abline(v=qgamma(0.025,115.001,20.001), col = "red")
abline(v=qgamma(0.975,115.001,20.001), col = "red")
```

Άσκηση 4

Δίνεται ένα δείγμα 5375 τυχαία επιλεγμένων φοιτητών καπνιζόντων ή μη σε σχέση με το αν καπνίζει κάποιος από τους γονείς τους. Παρακάτω δίνεται ο πίνακας των τιμών.

	Ο φοιτητής κανίζει ($\Upsilon = 1$)	Ο φοιτητής δεν καπνίζει ($\Upsilon = 0$)
Τουλάχιστον ένας γονιός καπνίζει (X = 1)	816	3203
Κανένας γονιός δεν καπνίζει (X = 0)	188	1168

Σκοπός της έρευνας είναι να δούμε κατα ποιό τρόπο σχετίζεται το κάπνισμα των νεότερων σε σχέση με την καπνιστική συμπεροφορά των γονιών.

Δημιουργία μοντέλου κλασικής στατιστικής

Στο πρώτο κομμάτι δημιουργούμε το μοντέλο της λογιστικής παλινδρόμησης με την βοήθεια της κλασικής στατιστικής. Αυτό θα χρησιμοποιήσουμε για να προχωρήσουμε σε συγκρίσεις με το μοντέλο που θα προκύψει από το Winbugs.

Στον παρακάτω πίνακα (table 4) βλέπουμε τις τιμές β_0, β_1 όπως προκύπτουν για το προσαρμοσμένο μοντέλο.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8266	0.0786	-23.24	0.0000
X	0.4592	0.0878	5.23	0.0000

Table 4: Λογιστική παλινδρόμηση στα δεδομένα

Στην συνέχεια προχωράμε με την δημιουργία του μοντέλου με την χρήση μη πληροφοριακών πρότερων κατανομών. Χρησιμοποιούμε κανονικές κατανομές με μέση τιμή 0 και πολύ μεγάλη διασπορά. Για τις ανάγκες του winbugs χρησιμοποιούμε πολύ μικρή ακρίβεια. Για να καταφέρουμε να περάσουμε τα δεδομένα στο μοντέλο δημιουργούμε 2 διανύσματα. Ένα με τον αριθμό των γονιών που καπνίζει και αυτών που δεν καπνίζουν c(4019,1356). Ενα δεύτερο με τον αριθμό των επιτυχιών όσον αφορα τους φοιτητές που καπνίζουν στις αντίστοιχες ομάδες. 816 φοιτητές στους 2019 γονείς που καπνίζουν και 188 στους γονείς που δεν καπνίζουν. Αντίστοιχα τοποθετούμε την επεξηγηματική μεταβλητή κάπνισμα ή μη. Αυτό γίνεται πρός οικονομία χρόνου καθότι θα ήταν δύσκολο να ακολουθήσουμε την τακτική που ακολουθήσαμε στην R καθότι δεν έχουμε εργαλείο που να δημιουργεί διανύσματα για εμάς.

Ο κώδικας του μοντέλου δίνεται παρακάτω:

```
model smokers;
{
        for(i in 1:N){
        y[i] ~ dbin(p[i], n[i])
        logit(p[i]) <-beta0 + beta1*smoke[i]

}
beta0 ~ dnorm(0, 1.0E-6)
beta1 ~ dnorm(0, 1.0E-6)
}

list(beta0=0, beta1=0)

list(N=2, n=c(4019,1356), y=c(816,188), smoke=c(1,0))
```

Τα βασικά στοιχεία των αλυσίδων των τιμών β_0, β_1

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	-1.828	0.07765	0.001814	-1.983	-1.826	-1.681	1001	20000
beta1	0.4595	0.08702	0.002011	0.2898	0.4584	0.6324	1001	20000

Table 5: Στατιστικά Αλυσίδων με εισαγωγή μη πληροφοριακών πρότερων

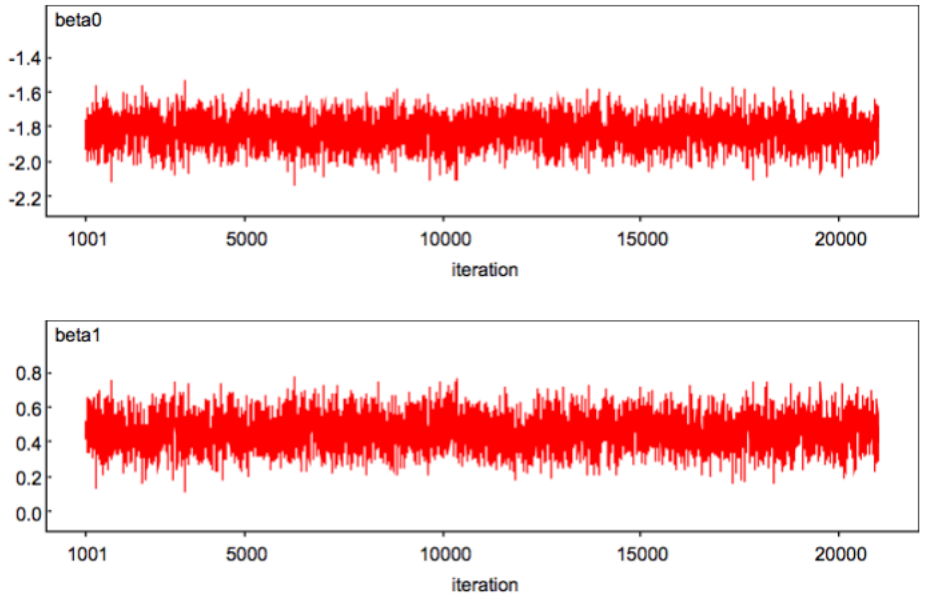


Figure 22: Αλυσίδες β_0, β_1

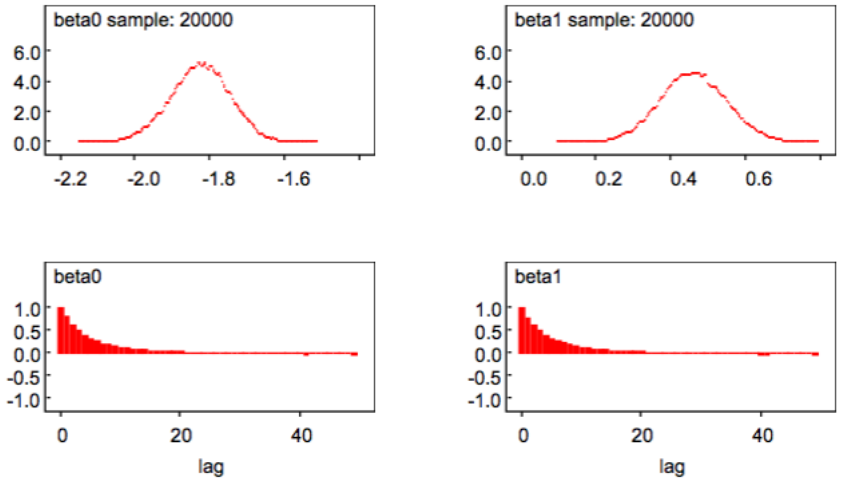


Figure 23: Αλυσίδες β_0, β_1

Από την τιμή του MC error και τα υπόλοιπα διαγράμματα κρίνουμε ότι οι αλυσίδες έχουν συγκλίνει και άρα μπορούμε να προχωρήσουμε σε εκτίμηση των παραμέτρων. Για την β_0 η τιμή θα είνα -1.828 και για την β_1 η τιμή θα είναι 0.4595 .

Αν κανείς δεί και συγκρίνει αυτές τις τιμές με τις τιμές που προκύπτουν από το μοντέλο που δέχεται σαν παράμετρο μόνο τα σημερινά δεδομένα θα διαπιστώσει μια σχεδόν πλήρης σύγκλιση. Αυτό συμβαίνει γιατί ο αριθμός των δεδομένων είναι ακετά μεγάλος (5374 δείγματα) και ταυτόχρονα οι δύο πρότερες κατανομές είναι αρκετά μη πληροφοριακές ώστε να μην επηρεάζουν το τελικό μοντέλο. Για να άλλαζαν οι τιμές θα έπρεπε οι πρότερες κατανομές να ήταν αρκετά πληροφοριακές ή τα δεδομένα να μην προέρχονταν από ένα τόσο μεγάλο δείγμα.

Ερμηνεία με χρήση Odds ratio

Χρησιμοποιούμε το μοντέλο $log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.828 + 0.4595 * smoke$

Η τιμή για το αν ένας από τους δύο γονείς καπνίζει είναι $log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.828 + 0.4595 * 1 = -1.3685$
$\Rightarrow \frac{\hat{p}}{1-\hat{p}} = 0.2544884$

Η τιμή για το αν κανένας από τους δύο γονείς δεν καπνίζει είναι $log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.828 + 0.4595 * 0 = -1.828$
$\Rightarrow \frac{\hat{p}}{1-\hat{p}} = 0.1607347$

Από τον υπολογισμό των odds παραπάνω μπορούμε να βρούμε τις πιθανότητες ο φοιτητής να καπνίζει ή όχι ανάλογα με το αν έχει η όχι καπνιστή έναν γονέα.

Για να καταφέρουμε να εκφράσουμε το αντίστοιχο συμπέρασμα με την χρήση του Odds Ratio είναι:

$$\frac{log\left(\frac{\hat{p}}{1-\hat{p}}\right)}{log\left(\frac{\hat{p}}{1-\hat{p}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta 1} = 1.583282$$

Probit μοντέλο

Στην συνέχεια θα χρησιμοποιήσουμε το Probit μοντέλο $Pr(Y = 1 \mid X) = \Phi(X^T\beta)$.

```
model smokers2;
{
        for(i in 1:N){
        y[i] ~ dbin(p[i], n[i])
        probit(p[i]) <-beta0 + beta1*smoke[i]

}
beta0 ~ dnorm(0, 1.0E-6)
beta1 ~ dnorm(0, 1.0E-6)
or <- exp(1.6*beta1)
}

list(beta0=0, beta1=0)

list(N=2, n=c(4019,1356), y=c(816,188), smoke=c(1,0))
```

Στον κώδικα έχει χρησιμοποιηθεί άλλη μια παράμετρος η OR την οποία θα χρησιμοποιήσουμε παρακάτω. Παρακάτω παρατίθενται τα σημαντικότερα στοιχεία των αλυσίδων που προκύπτουν από το Winbugs.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	-1.086	0.04234	7.396E-4	-1.17	-1.086	-1.004	1001	20000
beta1	0.2555	0.04781	8.485E-4	0.1631	0.255	0.3486	1001	20000
or	$e^{1.6\beta 1} = 0.4088$							

Table 6: Στατιστικά Αλυσίδων με εισαγωγή μη πληροφοριακών πρότερων

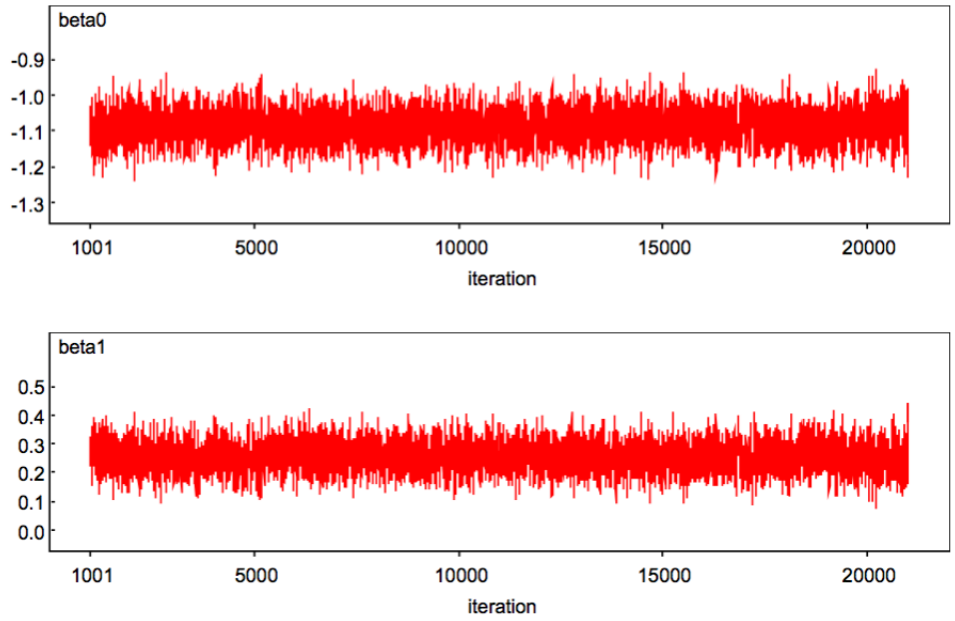


Figure 24: Αλυσίδες β_0, β_1

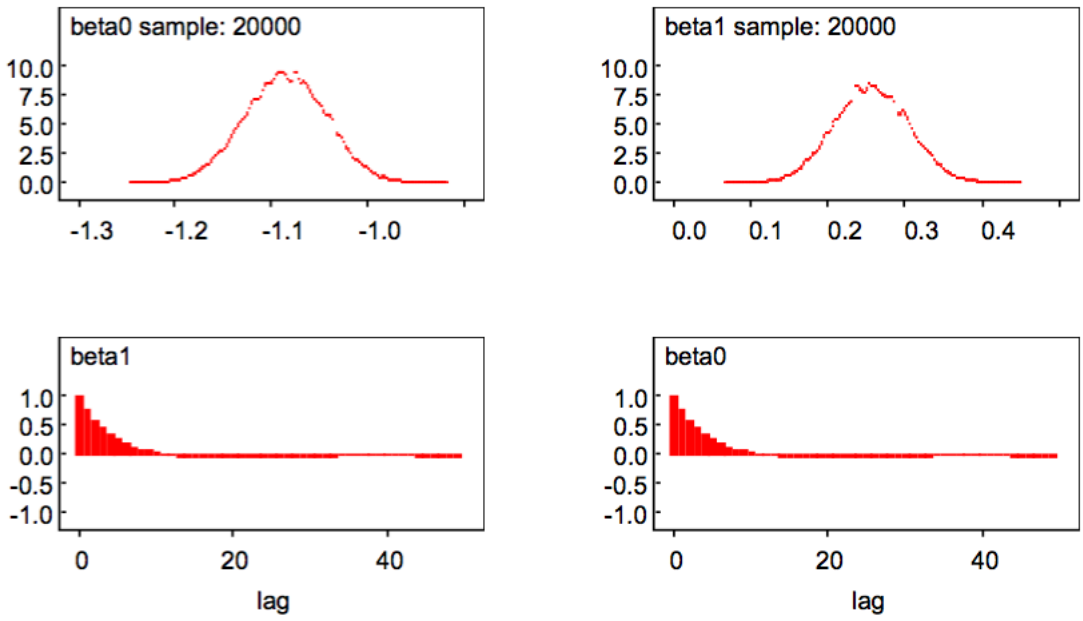


Figure 25: Κατανομές β_0, β_1 και αυτοσυσχέτιση

Το μοντελο Probit είναι το εξής : $\Phi^{-1}(p_\iota) = \sum_0^{k=n} \beta_k x_{ik}$. Για να καταφέρουμε να ερμηνεύσουμε λοιπόν το μοντέλο πρέπει να προχωρήσουμε σε έναν μετασχηματισμο τον παραμέτρων του μοντέλου. Αυτός μας δίνεται από την εκφώνηση. Για να βρούμε το Odds ratio έχουμε $\frac{\Phi^{-1}(p_\iota|X=1)}{\Phi^{-1}(p_\iota|X=0)} = e^{1.6*\beta_1} = 1.506065$. Χρησιμοποιούμε την τιμή or που έχει παρουσιαστεί στον πίνακα των τιμών του μοντέλου και είναι ο μετασχηματισμός που μας δίνεται. Το αποτέλεσμα μας λέει πως ένας φοιτητής αν έχει γονιό καπνιστή η πιθανότητα να καπνίσει αυξάνεται κατα 1.506065φορές.

Και στα τρία μοντέλα (Κλασικής στατιστικής , Probit, Logit) έχουμε αρκετά κοινά αποτελέσματα. Αυτό όπως έχει επισημανθεί συμβαίνει γιατί στον υπολογισμό των ύστερων μοντέλων έχουν χρησιμοποιηθεί αρκετά ασαφής πρότερες κατανομές και τα δεδομένα του προβλήματος έχουν πολύ μεγάλο αριθμό. Άρα είναι λογικό να είναι παρεμφερή τα αποτελέσματα και με τα τρία μοντέλα.

Ακόμα και το μοντέλο της κλασικής στατιστικής μας δίνει Odds ratio 1.582807. Η πιθανότητα δηλαδή να καπνίσει κάποιος φοιτητής αν έχει κάποιο γονιό καπνιστή αυξάνεται κατά 1.58 φορές. Λίγο μεγαλύτερη από το Probit μοντελο και σχεδόν ίδια με του Logit μοντέλου 1.583282.

DIC

Για να αποφανθούμε ποιό μοντέλο είναι καλύτερο υπάρχουν πολλά εργαλεία. Στην συγκεκριμένη περίπτωση θα χρησιμοποιήσουμε το DIC (Deviance information criteria).

Το μοντέλο που θα επιλεγεί είναι αυτό με την μικρότερη τιμή DIC. Παρακάτω παρουσιάζονται οι τιμές του και για τα δύο μοντέλα:

	Dbar	Dhat	pD	DIC
y	17.206	15.242	1.960	19.166
total	17.206	15.246	1.960	19.166

Table 7: DIC τιμές για το Logit μοντέλο

	Dbar	Dhat	pD	DIC
y	17.254	15.242	2.012	19.266
total	17.254	15.242	2.012	19.266

Table 8: DIC τιμές για το Probit μοντέλο

Οι τιμές και για τα δύο μοντέλα είναι πάρα πολύ κοντά. Για μια πάρα πολύ μικρή διαφορά θα επιλέγαμε το Logit μοντέλο για να περιγράψουμε τα αποτελέσματα της έρευνας.