University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Assignment cover sheet

All work handed in for assessment must have a completed copy of this form.
In completing the following declaration and signing below, you are also considered to have signed the following University of Edinburgh Own work declaration cover sheet:
http://www.maths.ed.ac.uk/hall/ORMSc/OwnWorkDeclaration.pdf

Name: Salvador Garcia Gonzalez

Matriculation number: s1655274

Assignment 4: Assgn4: Salmonella Data

I certify that (delete as applicable):

(a) ~~I did not discuss this work with other students~~

(b) ~~I wrote this account independently after having discussed it with the students named below~~

(c) ~~I received help from another source (please specify below)~~

Assgn4: Salmonella Data

Salvador Garcia, s1655274

17 February 2017

1 Description of the problem

Breslow (1984) analyses mutagenicity assay data on salmonella in which three plates have each been processed at various doses of quinoline (0,10,33,100,333,1000), and the number of colonies of TA98 salmonella subsequently measured.

Denote the dose by x_i, $i = 1, \ldots, 6$, and the number of colonies observed on plate j at dose x_i by $y_{i,j}$, $j = 1, 2, 3$.

The theory suggests the following model for $\mu_i = Ey_{i,j}$:

$$\log(\mu_i) = \alpha + \beta \log(x_i + 10) + \gamma x_i, \qquad \text{with } \alpha, \beta, \gamma \in \mathbb{R}.$$

2 Likelihood

The proposed model by Breslow considers the following likelihood for the $y_{i,j}$:

$$y_{i,j} \mid \mu_i \sim Pois(\mu_i) \quad \text{independently (given) } \mu_i).$$

3 Prior

For the prior distribution in the model, three noninformative priors will be used for $\alpha, \beta, \gamma \in \mathbb{R}$ ($N(0, 100^2)$). One for each parameter. In the first part of this report, the data for dose $x_i = 100$ will be excluded from the analysis. In the second part, the predictive distribution will be compared with the observed data for the dose $x_i = 100$.

4 Posterior inference

4.1 Converge analysis

When running the first model (with $X_i = (0, 10, 33, 100, 333, 1000)$), the autocorrelation for each of the parameters α, β and γ is very strong and the chains does not look good.

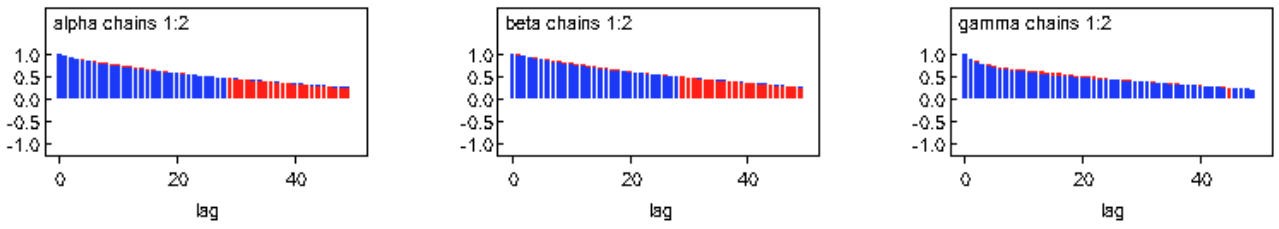

Figure 1: Autocorrelation for the parameters α, β and γ.

This problem can be solved when thinning the chain. Due to the strong correlation, a thinning of 50 will be used. With this modification the chains look much better and the autocorrelation for the three parameters is almost gone.

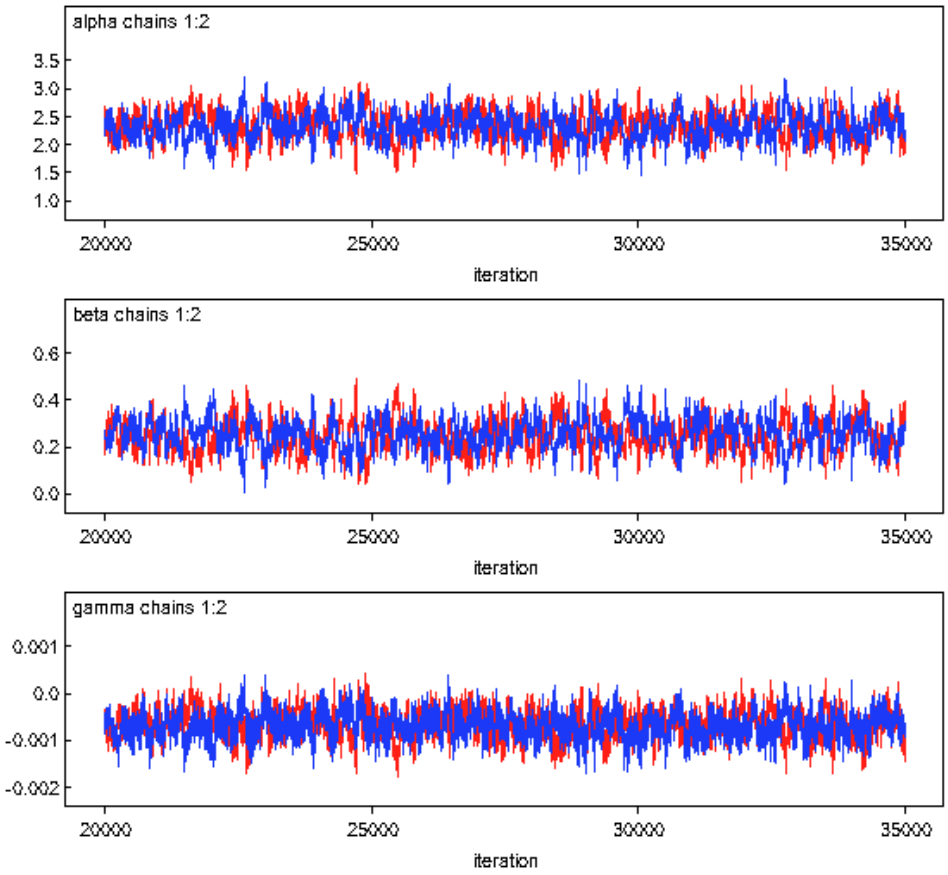

Figure 2: Chain for the parameters α, β and γ.

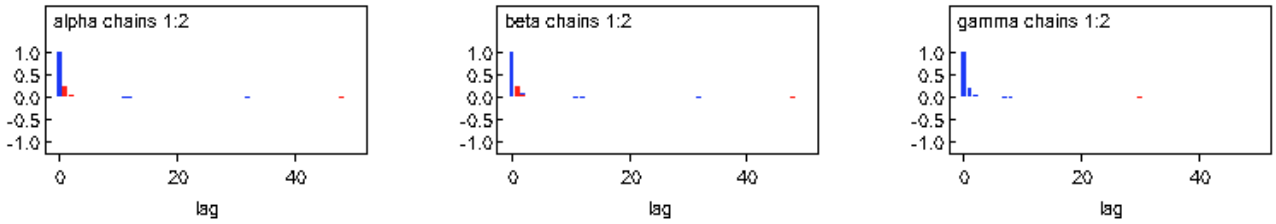

Figure 3: Autocorrelation for parameters α, β and γ with thinning of 50.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	2.346	0.2322	0.001591	1.886	2.347	2.797	20000	30002
beta	0.247	0.06488	4.544E-4	0.1196	0.2468	0.3743	20000	30002
gamma	-6.538E-4	2.851E-4	1.977E-6	-0.001218	-6.54E-4	-9.821E-5	20000	30002

Figure 4: Statistics for the parameters α, β and γ with thinning of 50.

4.2 Residuals analysis

The Poisson distribution only uses one parameter λ (in this example, μ_i for $i = 1, ..., 6$). And also, it have the property that:

$$\begin{aligned} E(y_{i,j}|\mu_i) &= \mu_i \\ V(y_{i,j}|\mu_i) &= \mu_i \end{aligned} \tag{1}$$

with μ_i the fitted parameter from our model. Then the standarised residuals can be expressed as $\frac{y_{i,j}-\mu_i}{\sqrt{mu_i}}$ with i the dose and y the number of colonies.

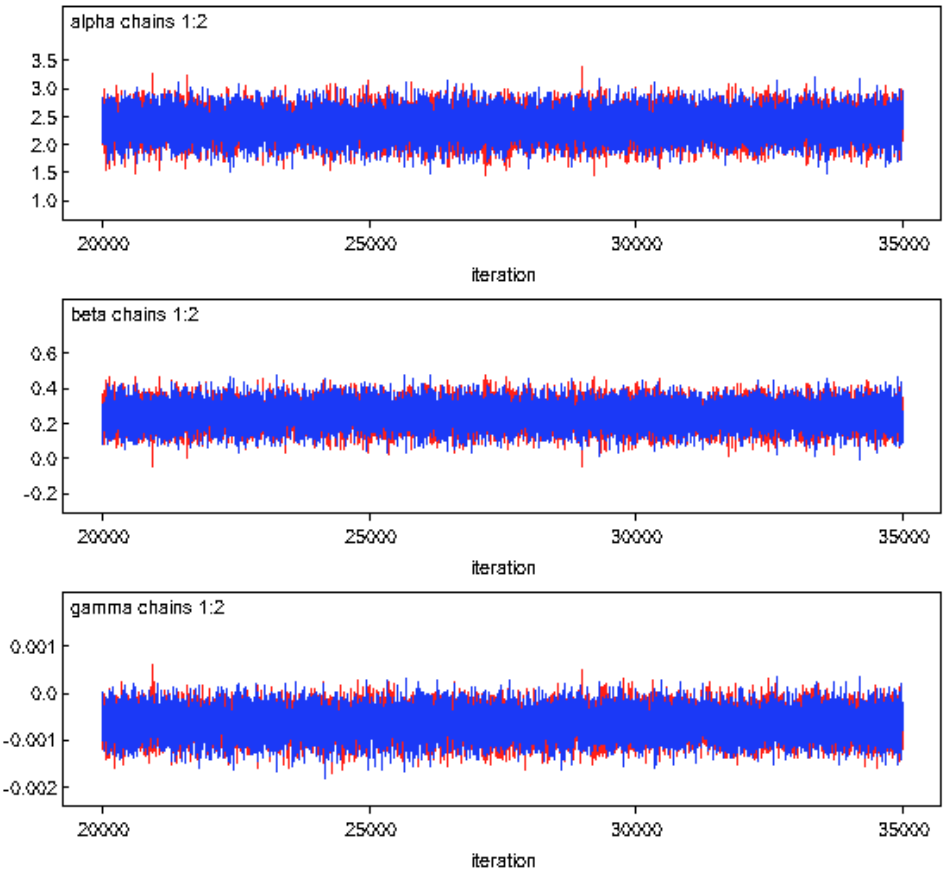

Figure 5: Chain for the parameters α, β and γ with thinning of 50.

```
model {
  for (i in 1:5) {
    for (j in 1:3) {
      y[i,j]     ~ dpois(mu[i])}
    log(mu[i]) <- alpha + beta*log(x[i] + 10) + gamma*x[i]}
# priors
  alpha          ~ dnorm(0, 0.0001)
  beta           ~ dnorm(0, 0.0001)
  gamma          ~ dnorm(0, 0.0001)
# residuals

  for (i in 1:5) {
    for (j in 1:3) {
  # add pearson (standardised) residuals for each observation y[i,j]
         stres[i,j]<-(y[i,j]-mu[i])/sqrt(mu[i])
   #    p[i,j] <- phi(stres[i,j])

  # deviance for the ith element
     DS[i,j] <- 2*(mu[i]-y[i,j] - y[i,j]*(log(mu[i]) -log(y[i,j]) ) )
     sign[i,j] <- 2*step(y[i,j]-mu[i])-1
  # deviance residual
    res.dev[i,j] <- sign[i,j]*sqrt(DS[i,j])
    }
  }
}
```

4.3 Residual analysis

In this section the boxplots of the residuals will be made. Each box in the plot represent the distribution of the residuals corresponding to (i, j) (again, i is the dose, and j is the plate). From a quick analysis of the fig. 6 and fig. 7 is important to say that the residuals corresponding to the plate 1: (1,1), (2,1), (3,1), (4,1) and (5,1) almost all are negative. On the other hand, the residuals corresponding to the plate 3: (1,3), (2,3), (3,3), (4,3) and (5,3) are almost always positive. This could mean some kind of preordering (for example, the plate 3 is the plate with more number the colonies). Now, analysing (i,1), (i,2) and (i,3) for each i, it is common that in the mean the residuals are centered in zero. This is because in general, the boxplot (i,3) is the positive residuals, (i,1) in the negative and (i,2) with center in 0. The same idea applies to the deviance residuals. About the outliers, almost all the residuals are between -2 and 2, but there are some (for example in the plate (5,3) and (5,1)) where the whiskers are in values near 4.

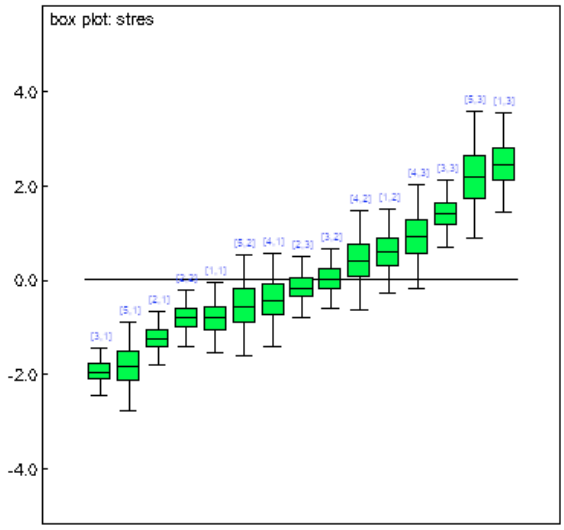

Figure 6: Boxplot of the standarised residuals ordered by rank.

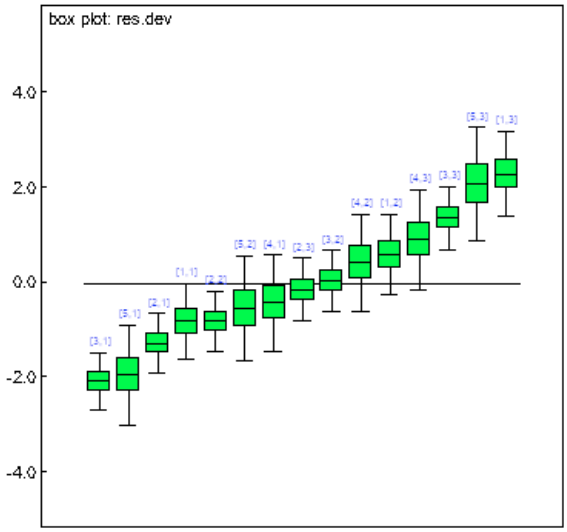

Figure 7: Boxplot of the deviance residuals ordered by rank.

4.4 Model Fit

For this part of the report, the model fit for each one of the plates will be analyzed fig. 8, fig. 9, fig. 10:

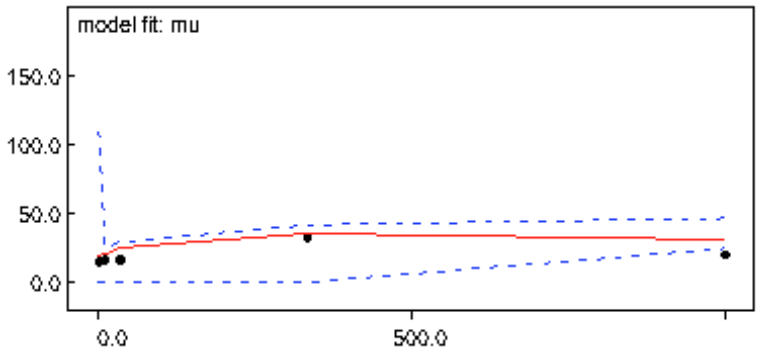

Figure 8: Model fit for plate 1

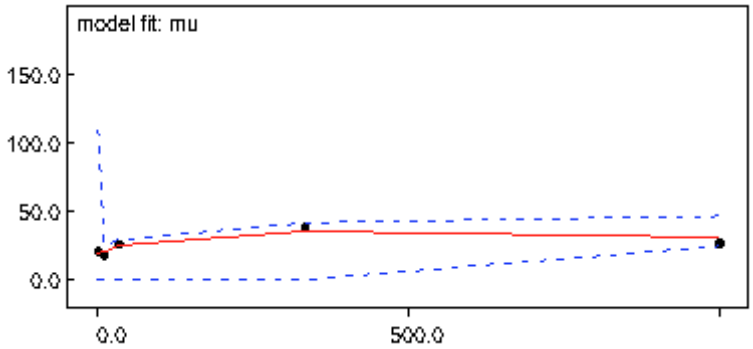

Figure 9: Model fit for plate 2

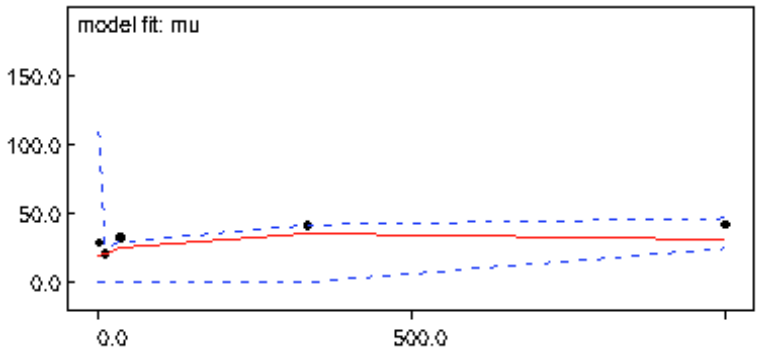

Figure 10: Model fit for plate 3

With these three plots, the idea presented in the last section (residuals) is corrobored. In the plate 1, the model prediction is larger than the original y(i,1); in the plate 2, the model prediction is close to the original y(i,2); and in the last plate (3) the prediction is smaller than the original y(i,3). About the outliers, similar conclusions from the residuals analysis will be given: For the plate 1 when x = 1000 (y(5,1)) it seems that the number of colonies observed is lower than the 95% intervals. In addition to this outlier, the plate 3 when x = 33 (y(3,3)) seems to be outside the 95% interval.

5 Predictions

Now, we are going to obtain the predictions when x = 100. The density of the predictive distribution of *y.pred* is presented in the fig. 11. The stats for this variable are in fig. 12

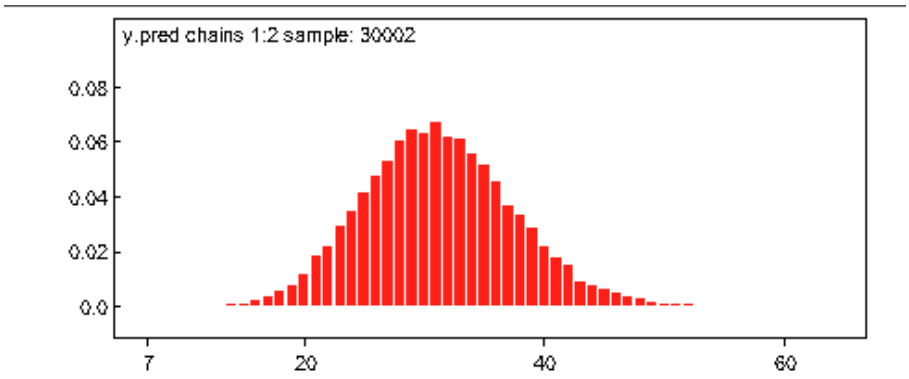

Figure 11: Density of the predictive distribution of *y.pred*

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
y.pred	31.24	6.181	0.03608	20.0	31.0	44.0	20000	30002

Figure 12: Statistics for the predictive distribution of *y.pred*

The observed number of colonies for this $x = 100$ is (27, 41, 60). Then, from the fig. 11 can be see that the values of 27 and 41 are common values in this distribution, but the value 60 seems to be not very common. In fact, the 95% credibility interval is from 20 to 44. Now, lets use the p-values p[1], p[2] and p[3] to adress the same question.

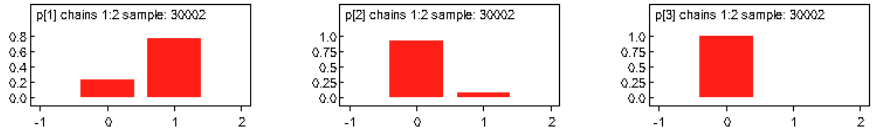

Figure 13: Density of p[1], p[2] and p[3]

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
p[1]	0.7742	0.4181	0.00246	0.0	1.0	1.0	20000	30002
p[2]	0.07183	0.2582	0.00144	0.0	0.0	1.0	20000	30002
p[3]	3.333E-5	0.005773	3.326E-5	0.0	0.0	0.0	20000	30002

Figure 14: Statistics for p[1], p[2] and p[3]

As thinked, the p-value of p[3] is very close to 0, then for this value the prediction it is not close. p[2] and p[1] are not too close to 1 or 0: .071 and .77. In fact, they lay in the interval (.05, .95).

As an overall conclusion, for this experiment two of the three predictions are inside the confidence intervals. Although the last prediction is not in this interval, it has been shown that the plate 3 contains the original values are much bigger than in the other two plattes.

A possible improvement to this model is to fit one model (or a variable that thakes into account this difference) for each type of plate. Another option is to try to find the variable that is making this difference. For example, maybe all the plates 1 are in total obscurity and the plates 3 recieve more sun ligth. One possible modification can be to add the total number of hours of sunlight received by the plate. I think that in general for the data that is included in this problem is a good model. But it can be improved.

If we consider the hypothesis $\gamma = 0$, taking into account the MC-error for this parameter, we can see that the MC error is approx 1.5% of the posterior sd of the parameter gamma. So this a good number of MC-error. On the other hand, the credibility interval of this parameter is (-.001218, -.00009821) that is very close to zero, but does NOT include the 0. With a 95% test, the hypothesis $\gamma = 0$ will be rejected.