Mice Radiation Experiment

Disclaimer: I don't speak English very well, so there are some grammatical mistakes in this report, thanks to being understandable.

Summary

In this project I used the data base about the effective of Streptomycin Therapy on mice irradiated with fast neutrons, to create a model to predict if the difference of therapy on treatment (Streptomycin or saline control) and the neutron dose effects the of number of deaths

To do this I used a logistic regression to predict if the mouse died or not (1 or 0), using 2 different approaches, and comparing the results.

We concluded that these 3 models didn't affect the results of accuracy, and both predicted the same answers using the same threshold.

Introduction

In this dataset we have 2 treatment groups, so it's natural to do an ANOVA analysis to distinguish the results to see if a group tends to lead higher survival rate, in this case the special therapy is the Streptomycin, and the control group is the saline control. So, to do that the model used 2 different interceptions to each group. To compare the results, I used a normal logistic regression with same intercept to theses 2 groups. With this 2 models I added one more, using the lm() function in R

Data

The data has 3 columns, the first one is the neutrons dose received, the second is if the mouse is in treatment 1 (saline control) or 2 (Streptomycin), and the third column is if the mouse died or not (1 or 0). The head of the data in the Figure below.

	neutrons	treatment	died
1	201	2	1
2	201	2	1
3	201	2	1
4	201	2	0
5	201	2	0
6	201	2	0

Figure: First 6 samples of the data base

To see the relation between the deaths and the treatment we can use the table in the Figure below, the columns are the treatment 1 and 2, so if the new therapy 2 is better so the rate of deaths using this is lower, looking to the table we can presume the therapy control (0) is related with more dead mice, and the therapy with Streptomycin can be associated with less dead rate, but we need construct a model to validate that hypothesis.

Therapy		
Died	1	2
0	67	157
1	194	104

Neutrons dose				
Died	201	220	243	260
0	57	107	51	9
1	13	133	131	21

Figure: Table between Therapy and death, and between Neutrons dose and death

Another table in the Figure above is the relationship between Neutrons dose and deaths, here the 2 largest doses are related to a higher death rate, it makes sense, so we expect a positive relation between these 2 variables.

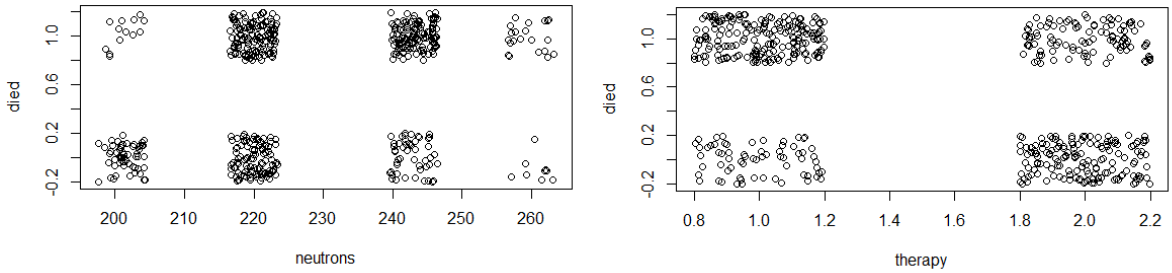

Figure: The same relation in scatterplot

Model

The first model is a logistic regression, with one intercept and 2 parameters β_1 and β_2 associated with neutrons dose and therapy. So, the likelihood of dying is a Bernoulli distribution with probability p, and p is calculated using link function $logit[p[i]] = b_0 + b_1 neutrons[i] + b_2 therapy[i]$, that's a linear regression.

The priors chosen was $b_0 \sim N(\mu = 0, \sigma^2 = 10^4)$, the large variance is to have an informative prior to intercept. $\beta_1 \sim doubleExp(\mu = 0, \sigma^2 = 1)$ and $\beta_2 \sim doubleExp(\mu = 0, \sigma^2 = 1)$, these two was chosen this distribution to trying minimize β and assuming low variance.

The second model is also a logistic regression but with the intercept depending on the group of therapy and one parameter β_1 associated with neutrons dose. In this case the probability p is calculated using the link function $logit[p[i]] = int[therapy[i]] + b_1 neutrons[i]$.

The intercept depending on therapy group was chosen assess whether this factor interferes in the mean value of each group, in this case it is expected that the control group will have the mean value of interception greater than the intercept of the Streptomycin group, because the hypothesis here is the control group have the mortality rate greater than the Streptomycin group.

The priors in this second models were $int[1] \sim N(\mu, \sigma^2)$, $int[2] \sim N(\mu, \sigma^2)$ with $\mu \sim N(0, 10^4)$ and $\sigma^2 \sim IG(1, 1)$, and $\beta_1 \sim doubleExp(\mu = 0, \sigma^2 = 1)$.

In the both models the neutrons doses was normalized and scaled, because their values were in the range between 201 and 260, so after scaling we got values between -1.64 and 1.98.

Results

To the model 1 I used 1000 samples to burn-in to exclude convergence period of the data samples, and sampled 3 chains of MCMC for 10^4 iterations, it was enough to convergence of these 3 chains. The results are in Table below, the Upper CI of Gelman test, effective size of the chains and Lag 5 of autocorrelation. Here we have a high autocorrelation between the samples, that impacts the effective size of our samples, mainly in β_0 and β_1, but in β_2 we had a good independence between samples.

	Upper CI	**Effective size**	**Lag 5**	**Mean**	**STD**	**2.5%**	**97.5%**
$\boldsymbol{\beta_0}$	1.0	821.33	0.76	2.79	0.35	2.11	3.45
$\boldsymbol{\beta_1}$	1.01	9113.94	0.75	0.76	0.11	0.56	0.98
$\boldsymbol{\beta_2}$	1.01	811.27	0.07	-1.61	0.21	-2.04	-1.20

Table: Summary Table of the first model.

In the same Table contain the summary of our model, like mean, standard deviation, and the interval of 95% of density probability. In this case we have all these 3 parameters different from zero, so we don't need discard or remodel any of them. Here we can infer there is a positive relation between Neutrons dose and mortality, and a negative relation between treatment and mortality, what was expected.

Using a decision threshold equals 0.5 (If the output of logistic regression is > 0.5 we assume death 1, in another case the mouse is alive), the Table is below, where we had 71.6% of accuracy.

	Estimation 0	Estimation 1
Died 0	133	91
Died 1	57	241

Table: Table of decisions.

Also is important calculate the **DIC**, in this case we got: Mean deviance: 592.6, penalty 2.97, Penalized deviance: 595.6.

In the second model where we defined the intercept depending on the type o therapy also using the same Burn-in and iterations of MCMC, getting the results below.

	Upper CI	**Effective size**	**Lag 5**	**Mean**	**STD**	**2.5%**	**97.5%**
$\boldsymbol{\beta_0}$	1.0	16662.33	0.001	0.77	0.11	0.56	0.98
$\boldsymbol{INT_1}$	1.0	15594.42	0.001	1.19	0.15	0.89	1.49
$\boldsymbol{INT_2}$	1.0	17715.43	0.003	-0.46	0.14	-0.72	-0.19

Table: Summary Table of second model.

In this case we had a good convergence and a good independence samples, the Lag 5 is near to 0 and our effective sample size is larger than the model 1. And the intercept 1 have a positive mean, because the treatment 1 (control group) tends to have a greater mortality rate, and the intercept 2 have a negative mean, because their relationship with the mortality rate is negative.

Also, the **DIC** in this second model is better, but not so much, in this case we got: Mean deviance: 592.6, penalty 3.103, Penalized deviance: 595.7. So, the Penalized deviance is 0.9 greater than the original model.

To verify our hypothesis of the different means to each group we can just test the samples of each parameter, so if we try to computer mean(mod2_csim[,"int[1]"] > mod2_csim[,"int[2]"]), we get a probability if 1, so, the group control 1 is mor likely to die in this experiment.

But comparing the Table estimation using threshold equals 0.5, like the first model, we get the same answers.

Conclusion

In this report we concluded that the second model, using an intercept different to each group improve our model comparing the logistic regression using the same intercept. We get a better effective size of our sample because in this second model the autocorrelations are lower, and we showed that each intercept has influence in the result of our model.

But the improvement was only in the model structure, because their results of accuracy was the same. Otherwise, the hypothesis tested was confirmed using the second model, so we can conclude that the therapy using Streptomycin decrease the chances of death in mice.