Stat 474/574 Biostatistics. Survival Analysis and Logistic Regression. Final Take Home. Fall 2013. Due on December 12th at 5.00pm. (prefer a hard copy but and e-mail version in PDF format could be submitted to ghuerta@stat.unm.edu by the due date/time)

Directions: *The answer to these problems should be presented as a summary. It should be word processed and containing the most relevant information in your analysis following the format of the Homeworks. An appendix is allowed but that will be examined only to my discretion. The better organized your appendix is, the more likely it is to get examined. You may not consult when any other person when working on this exam or discuss your exam with anyone else in or outside the class. Of course, you may use you course notes, books or web resources to answer*

Question 1. (40 points)- The following data are the survival times (weeks in remission) of 42 leukemia patients in a clinical trial to compare treatment with a placebo. The data are taken from Freireich et al. ``The effect of 6-mercaptopurine on the the duration of steroid-induced remissions in acute leukemia,'' Blood 21, 699-716, 1963.

The columns correspond to the following variables:

1. survival time in weeks
2. status (0 = censored, 1 =relapse)
3. Sex (0= Male, 1=female)
4. log (WBC) log of the white cell blood count
5. treatment group (1=placebo, 0=treatment)

(a) The researchers are interested in the potential effect of the treatment on the survival time. Do a broad exploratory analysis of the data that will help the researchers to assess the treatment of effect. Also build a model or models (parametric or non-parametric) that will allow to assess the impact and direction of the treatment variable by taking into account other factors. Try to assess your models using some form of residual analysis if possible. Summarize your findings in a way that could be accessible for the researchers.

(b) Following the description in Section 14.5 of Dobson and Barnett's textbook, fit a Weibull survival model using Openbugs/Winbugs where the only covariate that is considered is 'treatment group' (so ignore Sex and Log(WBC) here, but of course if you consider models with these covariates, that is a bonus). In particular, produce 3000 MCMC samples for the 3 model parameters and show trace plots, ACF plots and summary statistics for these parameters. Do you think the MCMC works well for this model? Also produce the posterior distributions of median survival times for both placebo and treatment groups. Furthermore, consider this test statistic of difference of median survival times and produce posterior inference for this difference. What do you conclude from these analyses? (Note: handling censoring can be done with Openbugs/Winbugs, but for partial credit, I will also consider an analyses where censoring is not incorporated.)

```
35 0 1 1.45 0
34 0 1 1.47 0
```

32	0	1	2.20	0
32	0	1	2.53	0
25	0	1	1.78	0
23	1	1	2.57	0
22	1	1	2.32	0
20	0	1	2.01	0
19	0	0	2.05	0
17	0	0	2.16	0
16	1	1	3.60	0
13	1	0	2.88	0
11	0	0	2.60	0
10	0	0	2.70	0
10	1	0	2.96	0
9	0	0	2.80	0
7	1	0	4.43	0
6	0	0	3.20	0
6	1	0	2.31	0
6	1	1	4.06	0
6	1	0	3.28	0
23	1	1	1.97	1
22	1	0	2.73	1
17	1	0	2.95	1
15	1	0	2.30	1
12	1	0	1.50	1
12	1	0	3.06	1
11	1	0	3.49	1
11	1	0	2.12	1
8	1	0	3.52	1
8	1	0	3.05	1
8	1	0	2.32	1
8	1	1	3.26	1
5	1	1	3.49	1
5	1	0	3.97	1
4	1	1	4.36	1
4	1	1	2.42	1
3	1	1	4.01	1
2	1	1	4.91	1
2	1	1	4.48	1
1	1	1	2.80	1
1	1	1	5.00	1

Question 2 (40 points)- Consider the Armadillo hunting with age trends: Poisson regression for longitudinal count data example that is available from our class website in

the file 'Examples-Winbugs.doc'. For this question consider models that do not include the extra random effects term. Consider that,
Model 1: Is a model that represents the log intensity rate exclusively as a linear function of age (i.e. age[i]-50).
Model 2: Is the model where the log intensity is represented with a linear and quadratic function of age (i.e. model that includes both term (age[i]-50) and pow((age[i]-50),2)).
(a) Produce, say 5000 MCMC samples for each model and summarize the samples using the 'Stats' tool in Openbugs/Winbugs. Monitor the quantities that you consider are the most important for the analysis of these two models. Finally run additional samples and compute the DIC statistic for both models and report the values of DIC for Model 1 and 2. Also report the values of 'p_D', known as the effective number of model parameter and the model deviance based on DIC. (deviance = DIC- 2p_D). Which of the two models is preferred according to this criteria? Summarize your findings presenting relevant graphs or table.
(b) Now using R and with the usual maximum likelihood approach, compute AIC values for Models 1 and 2 and report the corresponding values along
with the DIC calculation of part (a). Do you see any notable changes between using DIC or AIC in this case? Comment on your results.

Question 3 (20 points). From the Dobson and Barnett textbook, Exercise 13.2 parts (a) and (b). I recommend using R for this exercise and follow closely the suggestions made in the book.
(*For students enrolled in Stat 474, part (b) is an extra credit question. For those enrolled in Stat 574 part(c) will be considered for additional credit*).