Homework 6 STATS 449 (Due 11-29)

Instructions: You can work together on homework, but your final answers must be your own. Late homework cannot be accepted. **Attach** and annotate output where relevant.

1) This problem demonstrates the usefulness of centering the explanatory variable x when fitting regression models.

a) Using the WinBUGS code on Slide 19 of Lecture 17 or in the document `Pump_example.pdf` exactly as it appears, conduct a Bayesian analysis of the pump data. We are only interested in the parameters α and β. Run 20,000 iterations of a Markov chain and discard the first 5,000 iterations for burn-in (in the *beg* dialogue box type in 5001). Look at the time series plots (iterations 5001-20000) for α and β by clicking on the *history* button. Attach these plots to your homework. (Note: Plots can be highlighted by clicking on them, and copied by right-clicking in the plot area). Also attach a plot of the posterior density of β and report the posterior mean and standard error.

b) Click on `Inference -> Correlations` to bring up the Correlation Tool. In the "nodes" dialog boxes, type in "alpha" and "beta" and begin the iterations at 5001 (type in 5001 in the *beg* dialogue box). Click on the "scatter" button to look at a scatter plot of the simulated values of α and β generated from their respective posterior distributions, and click on "print" to calculate the correlation. Attach the scatter plot to your homework and report the correlation.

c) Repeat parts a) and b), but code the x variable to equal -1 for pump type 1, and 1 for pump type 2. That is, let $x = c(-1, -1, -1, -1, 1, 1, 1, 1, 1, 1)$. What happened to the correlation between the estimated regression parameters? A Markov chain simulation is more efficient when the magnitude of the correlation between simulated values decreases. That is, when the correlation is closer to zero. Which coding of the x variable is more efficient in terms of the simulation? For this new coding of x, attach a plot of the posterior density of β and report the posterior mean and standard error. We know that for the coding $x = c(0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$, β has the interpretation as a $\log RR$. What is the relationship between the $\log RR$ and β for the new coding $x = c(-1, -1, -1, -1, 1, 1, 1, 1, 1, 1)$?

2) The data below are reproduced from a bioassay experiment where x_i represents the ith of $N = 6$ dose levels (measured on the log scale) given to n_i animals, of which y_i subsequently respond with positive outcome. A total of 480 animals were tested, 80 at each of the six dose levels.

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.553	80	2
-0.113	80	19
0.059	80	24
0.185	80	49
0.446	80	69
0.753	80	78

a) Use R to fit the standard logistic regression model

$$y_i|\,p_i \sim \text{Binomial}(n_i, p_i)\,, \qquad y_i|\,p_i \text{ independent} \qquad i = 1, \ldots, 6$$

$$\text{logit}(p_i) = \alpha + \beta\, x_i$$

and fill in the following table. Report G^2 and the model degrees of freedom. Test goodness-of-fit for the model.

Parameter	Estimate	Std Error
α		
β		

b) Based on part a) calculate the LD50 and use the bootstrap to estimate its standard error. Make a histogram of the LD50 bootstrap sample and include the R code and results.

c) Design a Bayesian model with a noninformative prior and use WinBUGS to fill in the same table as in part a). Run a chain with 20,000 iterations and discard the initial 5,000 for burn-in. How do the estimates and standard errors for α and β compare to those found in part a)?

d) Include a kernel density plot of the posterior distribution of the LD50 and report the posterior mean and the posterior standard error? How do these results compare with the bootstrap results from part b)?