Stat 574B: Bayesian Methods, HW 3

- Make the presentation nice, following directions on HW 1. Please hand in one problem on each of the next 3 Wednesdays (Oct 18,25 and Nov 1st). I will leave it up to you which problem you decide to hand in each week.

Problem 1

Recall the Lyme disease study, where in HW 1 we made inferences about the specificity of the test θ. In the study, 25 patients with no history of Lyme disease were tested and 19 tested negative, for a sample proportion of $19/25 = 0.76$. Assume a Beta(2,2) prior on θ. Simulation is not needed for this problem.

(a) Compute and plot the posterior distribution of θ along with the normal approximations to the posterior based on the MLE and the Laplace approximation. Comment on the accuracy of the approximations.

(b) For each of these three posteriors (two are approximations), compute the mean, SD and 4 equal-tail probability intervals: 0.90, 0.95, 0.975 and 0.99. Compare the approximations to the summaries from the actual posterior. Are the results consistent with your assessments in (a)?

Remark: I mentioned in class the connection between the 95% probability interval based on the Laplace approximation using a Beta(2,2) prior and the Agresti-Coull 95% frequentist confidence interval.

Problem 2

This folder includes a file **peptide.csv** that contains data for 43 participants in a study of the factors affecting patterns of insulin-dependent diabetes in children. The

3 variables in the file are the age in years a, the base-deficit b and the logarithm of serum C-peptide (Y). Although the base deficit b is expected to impact Y because decreasing levels of b are reflective of insufficient insulin delivery, we will initially consider a regression model to predict the log serum C-peptide level from age.

Consider the function $\mu(a, \beta, \kappa) = \beta_1 + \beta_2 a + \beta_3 (a - \kappa)_+$ where v_+ is v if $v > 0$ and 0 otherwise. So $(a - \kappa)_+$ is $a - \kappa$ if $a > \kappa$ and 0 otherwise. The function $\mu(a, \beta, \kappa)$ is a linear spline with a knot at κ, and regression coefficients $\beta = (\beta_1, \beta_2, \beta_3)$. This spline joins two linear functions of a so that they meet at κ. Splines are a staple for non-parametric regression. Typically, practitioners use cubic splines with multiple knots, B-splines, or one of several other approaches.

(a) Suppose $0 < a < 15$ and $\kappa = 7.5$. Choose 4 sets of values for $\beta = (\beta_1, \beta_2, \beta_3)$ and plot $\mu(a, \beta, \kappa)$ against a. Describe what you see and is it consistent with my description of splines?

(b) Plot Y versus a for the data, describe what you see, and assess qualitatively whether a spline would seem to appear to be a sensible (empirically-motivated) model for the mean response.

(c) For the remainder of the problem, consider the spline model $Y_i = \mu(a_i, \beta, \kappa) + \epsilon_i$ for $i = 1, 2, ..., n$, where the errors are independent $N(0, \sigma^2)$. For now, we will not worry that the variability in the response seems to increase somewhat with the mean. Note that the spline model is a linear regression model when κ is fixed or specified, but is a non-linear model if κ is treated as a parameter. Suppose that we restrict attention to values of κ that are integers from 4 to 12 inclusive. For each fixed value of κ in this range, fit the linear regression model

$$Y = X(\kappa)\beta + e,$$

where the design matrix $X(\kappa)$ depends on κ, and compute the residual sums of

squares, calling it $RSS(\kappa)$. Classical inference often picks the value of κ that minimizes $RSS(\kappa)$, treats the selected κ as known, and estimates (β, σ^2) using the fit obtained from the selected κ. Implement this procedure here, summarizing the result by providing the optimal value of κ, a plot of $RSS(\kappa)$ versus κ, and a table of parameter estimates and related LS summaries for the selected model. Also include a plot of the estimated regression function (i.e. $\mu(a, \hat{\beta}, \hat{\kappa})$) superimposed on the data. Does the summary seem to provide a reasonable description of the trend in the data, and does the estimated value of κ agree with what you see in the data?

Remark: Mixing math and R: If a is the vector of ages with n elements, then I can compute the design matrix as $X(\kappa) = cbind(rep(1, n), a, (a - \kappa) * (a - \kappa > 0))$. The third column of $X(\kappa)$ is the spline effect.

(d) Consider a Bayesian analysis for this spline model, assuming a prior for κ that is uniform on the integers 4-12, and a prior of your choice for (β, σ^2) given κ. Devise a sampling procedure to approximate the posterior of $(\beta, \sigma^2, \kappa)$ and summarize this posterior appropriately (i.e. means, SD, probability intervals, plots).

(e) Describe the posterior of the knot κ. Does the posterior suggest that the data strongly inform the location of the knot, or knot?

(f) For each a over a fine grid of points spanning the range of ages, compute and plot the posterior mean and a 95% posterior interval for $\mu(a, \beta, \kappa) = \beta_1 + \beta_2 a + \beta_3 (a - \kappa)_+$. You can view this as a Bayesian analog of the LS regression function with a frequentist pointwise CI for the regression function. Plot the results as a function of a, and include the observed data on the plot. You might think of using the same posterior samples for each value of a which should make the posterior mean of $\mu(a, \beta, \kappa)$ a smooth function of a. Comment on what you see - does the Bayesian summary make sense given the trend in the data?

Problem 3

The R datasets library has a data set called InsectSprays, which gives the counts of insects in identical agricultural experimental units treated with one of six insecticides. There are 72 observations on 2 variables - the count, and the insecticide used (A,B,C,D,E,F).

(a) Access the data, and make side-by-side boxplots of the distribution of the counts for sprays C, D and E (i.e. ignore A,B and F in this problem). Do the distributions appear to be roughly normal with similar spreads (as assumed by standard ANOVA), or not? Describe.

(b) Continuing with an analysis of the counts, consider the 1-way ANOVA model $Y_{ij} = \mu_i + \epsilon_{ij}$ for sprays $i = 1, 2, 3$ (1=C; 2=D, 3=E), where μ_i is the mean number of bugs per unit with spray i, and the ϵ_{ij} are independent $N(0, \sigma^2)$ with σ^2 unknown. Assuming a reference prior for $(\mu_1, \mu_2, \mu_3, \sigma^2)$, provide a summary table for each μ_i and all unique pairwise differences $\mu_i - \mu_j$. Include the posterior mean, SD, and a 95% posterior interval (which are also HPDs), as well as the posterior probability that the effect is < 0 or > 0, as described in the notes. You can this without simulation, by basing calculations on selected summaries from fitting the one-way ANOVA as a regression or ANOVA model in R, or using simple matrix calculations. Interpret the pairwise differences - and do any of the sprays appear to differ?

(c) Scientists in many disciplines (even some in statistics/biostatistics) like to report effect sizes. Here we may quantify the difference among means with

$$MAD = \sum_{i=1}^{3} |\mu_i - \bar{\mu}|/3 \quad \text{or} \quad NCP = \sum_{i=1}^{3} (\mu_i - \bar{\mu})^2/\sigma^2$$

(or anything else that might make sense because approximating posteriors is simple!) where $\bar{\mu} = \sum_{i=1}^{3} \mu_i/3$ is the average of the 3 groups means. Devise a simulation method to approximate the posterior distributions of MAD (the mean absolute deviation

in the μ_is) and NCP (proportional to the Non-Centrality Parameter in the F-test of no difference in means when group sizes are identical - a standard frequentist summary). Provide a table with estimated posterior means, SDs, and 95% probability intervals for MAD and NCP, along with plots of the posterior densities. Try your best to interpret the posterior summaries for MAD and NCP. As part of your analysis, carefully describe your simulation strategy.

(d) A standard interest in one-way ANOVA is whether the group means are identical: $\mu_1 = \mu_2 = \mu_3 = \mu$, say. If this hypothesis holds then the ANOVA model is a one-sample normal model: $Y_{ij}|\mu, \sigma^2 \sim$ independent $N(\mu, \sigma^2)$, which can also be formulated as a linear regression model with just an intercept. With this in mind, compute the BIC and DIC for the ANOVA model and the one-sample model. Which model is preferred by each criterion? Using BIC, compute an approximate Bayes Factor for comparing the one-sample model and the ANOVA model. Interpret the results using Jeffrey's scale. Do the results seem congruent with earlier analyses or not?

(e) Do any further analysis that you deem warranted and write a short conclusion.

Extra Credit: I would be inclined to do the 3 group comparison assuming the count distributions are Poisson. Devise and implement a method for deciding statistically whether the Poisson model or the standard normal theory ANOVA model is to be preferred, and summarize your findings. If you decide the Poisson model is better, don't feel the need to redo the original analysis.