STAT 209 – Generalized Linear Models (Fall 2019)

Homework 4 (due by 3pm, Monday December 9)

1. Consider a binary response ($y = 0, 1$) that corresponds to the choice between two options (say, two product brands), a choice that can be explained by a covariate x. Let U_0 denote the utility of choice $y = 0$, and U_1 the utility of choice $y = 1$. For $y = 0, 1$, suppose that $U_y = a_y + b_y x + \varepsilon_y$, using a scale such that ε_y follows a distribution in standard form. (Assume that the coefficients a_0, a_1, b_0 and b_1 are all known.) A subject selects $y = 1$ if $U_1 > U_0$ for that subject.
   (a) If ε_0 and ε_1 are independent $\mathrm{N}(0, 1)$ random variables, show that $\Pr(Y = 1)$ satisfies the probit regression model structure, and write the regression coefficients in terms of a_0, a_1, b_0 and b_1.
   (b) If ε_0 and ε_1 are independent random variables with c.d.f. $F(\varepsilon) = \exp\{-\exp(-\varepsilon)\}$, show that $\Pr(Y = 1)$ satisfies the logistic regression model structure (again, express the regression coefficients in terms of a_0, a_1, b_0 and b_1).

2. Consider the "alligator food choice" data example, the full version of which is discussed in Section 7.1 of Agresti (2002), *Categorical Data Analysis*, Second Edition. Here, consider the subset of the data reported in Table 7.16 (page 304) of the above book. This data set involves observations on the primary food choice for $n = 63$ alligators caught in Lake George, Florida. The nominal response variable is the primary food type (in volume) found in each alligator's stomach, with three categories: "fish", "invertebrate", and "other". The invertebrates were mainly apple snails, aquatic insects, and crayfish. The "other" category included amphibian, mammal, bird, reptile, and plant material. Also available for each alligator is covariate information on its length (in meters) and gender.

   (a) Focus first on length as the single covariate to explain the response probabilities for the "fish", "invertebrate" and "other" food choice categories. Develop a Bayesian multinomial regression model, using the baseline-category logits formulation with "fish" as the baseline category, to estimate (with point and interval estimates) the response probabilities as a function of length. (Note that in this data example, we have $m_i = 1$, for $i = 1, ..., n$.) Discuss your prior choice and approach to MCMC posterior simulation.

   (b) Extend the model from part (a) to describe the effects of both length and gender on food choice. Based on your proposed model, provide point and interval estimates for the length-dependent response probabilities for male and female alligators.

3. The table below reports results from a developmental toxicity study involving ordinal categorical outcomes. This study administered diethylene glycol dimethyl ether (an industrial solvent used in the manufacture of protective coatings) to pregnant mice. Each mouse was exposed to one of five concentration levels for ten days early in the pregnancy (with concentration 0 corresponding to controls). Two days later, the uterine contents of the pregnant mice were examined for defects. One of three (ordered) outcomes ("Dead", "Malformation", "Normal") was recorded for each fetus.

Concentration	Response			Total number
(mg/kg per day)	Dead	Malformation	Normal	of subjects
(x_i)	(y_{i1})	(y_{i2})	(y_{i3})	(m_i)
0	15	1	281	297
62.5	17	0	225	242
125	22	7	283	312
250	38	59	202	299
500	144	132	9	285

Build a multinomial regression model for these data using continuation-ratio logits for the response probabilities $\pi_j(x)$, $j = 1, 2, 3$, as a function of concentration level, x. Specifically, consider the following model

$$L_1^{(\text{cr})} = \log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) = \alpha_1 + \beta_1 x; \quad L_2^{(\text{cr})} = \log\left(\frac{\pi_2}{\pi_3}\right) = \alpha_2 + \beta_2 x$$

for the multinomial response probabilities $\pi_j \equiv \pi_j(x)$, $j = 1, 2, 3$.

(a) Show that the model, involving the multinomial likelihood for the data $= \{(y_{i1}, y_{i2}, y_{i3}, x_i) : i = 1, ..., 5\}$, can be fitted by fitting separately two Binomial GLMs. Provide details for your argument, including the specific form of the Binomial GLMs.

(b) Use the result from part (a) to obtain the MLE estimates and corresponding standard errors for parameters $(\alpha_1, \alpha_2, \beta_1, \beta_2)$. Plot the estimated response curves $\hat{\pi}_j(x)$, for $j = 1, 2, 3$, and discuss the results.

(c) Develop and implement a Bayesian version of the model above. Discuss your prior choice, and provide details for the posterior simulation method. Provide point and interval estimates for the response curves $\pi_j(x)$, for $j = 1, 2, 3$.

4. Consider the data set from homework 2, problem 3 on the incidence of faults in the manufacturing of rolls of fabric:

`http://www.stat.columbia.edu/~gelman/book/data/fabric.asc`

where the first column contains the length of each roll (the covariate with values x_i), and the second contains the number of faults (the response with values y_i and means μ_i).

(a) Fit a Bayesian Poisson GLM with the logarithmic link, $\log(\mu_i) = \beta_1 + \beta_2 x_i$. Obtain the posterior distributions for β_1 and β_2 (under a flat prior for (β_1, β_2)), as well as point and interval estimates for the response mean as a function of the covariate. Obtain the distributions of the posterior predictive residuals, and use them for model checking.

(b) Develop a hierarchical extension of the Poisson GLM from part (a), using a gamma distribution for the response means across roll lengths. Specifically, for the second stage of the hierarchical model, assume that $\mu_i \mid \gamma_i, \lambda \overset{ind.}{\sim} \text{gamma}(\lambda, \lambda\gamma_i^{-1})$, a gamma distribution with mean $\mathrm{E}(\mu_i \mid \gamma_i, \lambda) = \gamma_i$ and variance $\mathrm{Var}(\mu_i \mid \gamma_i, \lambda) = \gamma_i^2/\lambda$, where $\log(\gamma_i) = \beta_1 + \beta_2 x_i$.

Derive the expressions for $\mathrm{E}(Y_i \mid \beta_1, \beta_2, \lambda)$ and $\mathrm{Var}(Y_i \mid \beta_1, \beta_2, \lambda)$, and compare them with the corresponding expressions under the non-hierarchical model from part (a). Develop an MCMC method for posterior simulation providing details for all its steps. Derive the expression for the posterior predictive distribution of a new (unobserved) response y_0 corresponding to a specified covariate value x_0, which is not included in the observed x_i. Implement the MCMC algorithm to obtain the posterior distributions for β_1, β_2 and λ, as well as point and interval estimates for the response mean as a function of the covariate. Discuss model checking results based on posterior predictive residuals.

Regarding the priors, you can use again the flat prior for (β_1, β_2), but perform prior sensitivity analysis for λ considering different proper priors, including $p(\lambda) = (\lambda + 1)^{-2}$.

(c) Based on your results from parts (a) and (b), provide discussion on empirical comparison between the two models. Moreover, use the *quadratic loss L measure* for formal comparison of the two models, in particular, to check if the hierarchical Poisson GLM offers an improvement to the fit of the non-hierarchical GLM. Provide details on the required expressions for computing the value of the model comparison criterion.