STAT 209 – Generalized Linear Models (Fall 2019)

Homework 3 (due Tuesday November 19)

1. Consider the inverse Gaussian distribution with density function

$$f(y \mid \mu, \phi) = (2\pi\phi y^3)^{-1/2} \exp\left\{-\frac{(y-\mu)^2}{2\phi\mu^2 y}\right\}, \quad y > 0; \quad \mu > 0, \phi > 0.$$

Denote the inverse Gaussian distribution with parameters μ and ϕ by $IG(\mu, \phi)$.

(a) Show that the inverse Gaussian distribution is a member of the exponential dispersion family. Show that μ is the mean of the distribution and obtain the expression for the variance function.

(b) Consider a GLM with random component defined by the inverse Gaussian distribution. That is, assume that y_i are realizations of independent random variables Y_i with $IG(\mu_i, \phi)$ distributions, for $i = 1,...,n$. Here, $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ $(p < n)$ is the vector of regression coefficients, and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ is the covariate vector for the ith response, $i = 1,...,n$. Define the full model so that the y_i are realizations of independent $IG(\mu_i, \phi)$ distributed random variables Y_i, with a distinct μ_i for each y_i. Obtain the scaled deviance for the comparison of the full model with the inverse Gaussian GLM.

2. Consider the data set reported in Table 1 of Merrick, J.R.W., Soyer, R. and Mazzuchi, T.A. (2003), "A Bayesian semiparametric analysis of the reliability and maintenance of machine tools," *Technometrics*, vol. 45, pp. 58-69. For this problem, focus on cutting speed as the single covariate to explain/predict machine tool failure time. Cutting speed is recorded in feet per minute (fpm), and time to failure in minutes.

Develop and implement two Bayesian GLMs for these data, one based on the gamma distribution and one on the inverse Gaussian distribution for the random component. Include model assessment and model comparison in your analysis. Report point and interval estimates for the regression function under your models. Moreover, under each of your two models, obtain the posterior distribution for the mean failure time at cutting speed 400 fpm and 650 fpm, as well as the posterior predictive distribution for machine tool failure time at cutting speed 400 fpm and 650 fpm.

3. The following table reports results from a toxicological experiment, including the number of beetles killed (y_i) after 5 hours exposure to gaseous carbon disulphide at various concentrations. Concentration (log dose, x_i) is given on the $\log_{10}$ scale.

Log Dose, x_i	Number of beetles, m_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Consider a binomial response distribution, and assume that the y_i are independent realizations from $\text{Bin}(m_i, \pi_i)$, $i = 1,...,n$. The objective is to study the effect of the choice of link function $g(\cdot)$, where $\pi_i = g^{-1}(\eta_i) = g^{-1}(\beta_1 + \beta_2 x_i)$.

(a) Using R, fit three binomial GLMs to these data corresponding to link functions: logit, probit, and complementary log-log. Perform residual analysis for each model, using the deviance residuals. Obtain fitted values, $\hat{\pi}_i$, under each model and compare with observed proportions, y_i/m_i. Obtain the estimated dose-response curve under each model by evaluating $\hat{\pi}(x) = g^{-1}(\hat{\beta}_1 + \hat{\beta}_2 x)$ over a grid of values x for log dose in the interval $(1.65, 1.9)$. Plot these curves and compare with the scatter plot of the observed x_i plotted against the observed proportions. Based on all the results above, discuss the fit of the different models.

(b) One of the more general (parametric) link functions for binomial GLMs that has been suggested in the literature is defined through

$$g_\alpha^{-1}(\eta_i) = \frac{\exp(\alpha\eta_i)}{\{1 + \exp(\eta_i)\}^\alpha} \quad \text{for } \alpha > 0. \tag{3.1}$$

Note that the logit link arises as a special case of (3.1), when $\alpha = 1$. Discuss the effect of the additional model parameter α, in particular, for values $0 < \alpha < 1$ and $\alpha > 1$. Provide the expression for the log-likelihood for β_1, β_2 and α under the link in (3.1), and discuss the complications that arise for maximum likelihood estimation under this more general model compared with the logit GLM. (You do not need to fit the model, estimates are given below.)

(c) The MLEs under the model with link given in (3.1) are $\hat{\beta}_1 = -113.625$, $\hat{\beta}_2 = 62.5$ and $\hat{\alpha} = 0.279$. (The MLEs can be obtained using the Newton-Raphson method.) Using these estimates, obtain the fitted values $\hat{\pi}_i$ and the estimated dose-response curve under the link (3.1). Compare with the corresponding results under the three models in part (a). Obtain the deviance residuals from the model with link (3.1) and analyze them graphically.

(d) Although the four models considered in parts (a) to (c) are not nested, they involve the same (discrete) random component. Therefore, AIC and BIC values may be useful for comparing model performance. Compute the AIC and BIC for the four models to check if their values support your results regarding the fit of the different models.

4. This problem involves Bayesian analysis of the beetle mortality data from the previous problem.

(a) Consider a Bayesian binomial GLM with a complementary log-log link, i.e., assume that, given β_1 and β_2, the y_i are independent $\text{Bin}(m_i, \pi(x_i))$, $i = 1,...,8$, where

$$\pi(x) \equiv \pi(x; \beta_1, \beta_2) = 1 - \exp\{-\exp(\beta_1 + \beta_2 x)\}.$$

Design and implement an MCMC method to sample from the posterior distribution of (β_1, β_2). Study the effect of the prior for (β_1, β_2), for example, consider a flat prior as well as (independent) normal priors. Under the flat prior, obtain the posterior distribution for the median lethal dose, LD_{50}, that is, the dose level at which the probability of response is 0.5. Finally, plot point and interval estimates for the dose-response curve $\pi(x)$ (over a grid of values x for log dose).

(b) Next, consider a binomial GLM with a logit link, i.e., now the y_i are assumed independent, given β_1 and β_2, from $\text{Bin}(m_i, \pi(x_i))$, $i = 1,...,8$, where

$$\pi(x) \equiv \pi(x; \beta_1, \beta_2) = \exp(\beta_1 + \beta_2 x)/\{1 + \exp(\beta_1 + \beta_2 x)\}.$$

Working with a flat prior for (β_1, β_2), obtain MCMC samples from the posterior distributions for β_1, β_2, and for LD_{50}, along with point and interval estimates for the dose-response curve $\pi(x)$.

(c) As a third model, consider the binomial GLM with the parametric link given in (3.1). Develop an MCMC method to sample from the posterior distribution of $(\beta_1, \beta_2, \alpha)$, and obtain the posterior distribution for LD_{50}, and point and interval estimates for $\pi(x)$.

(d) Use the results from parts (a), (b) and (c) for an empirical comparison of the three Bayesian binomial GLMs. Moreover, perform residual analysis for each model using the *Bayesian residuals*: $(y_i/m_i) - \pi(x_i; \beta_1, \beta_2)$ for the first two models, and $(y_i/m_i) - \pi(x_i; \beta_1, \beta_2, \alpha)$ for the third. Finally, compare the three models using the *quadratic loss L measure*.