AMS 274 Homework #3 **Mary Silva**
UCSC 16 November 2018

1. The results from a toxicological experiment which included the number of beetles killed after 5 hours of exposure to gaseous carbon disulphid at various concentration (log dose) are plotted in the figure 1.

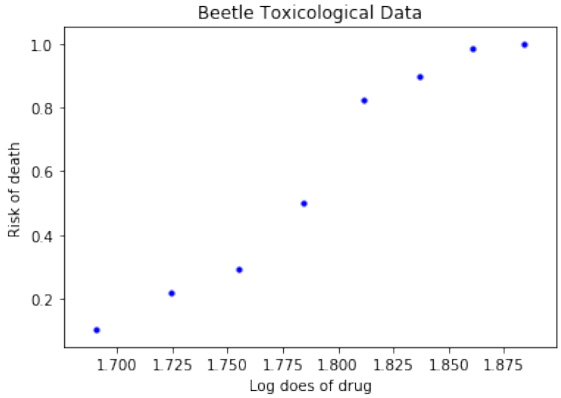

Figure 1: Risk of death (number of beetles dead/total number of beetles) plotted against log doses of carbon sulphide.

Based on the structure of the data, it seems reasonable to model the outcomes as independent with equal probabilities. In other words, the data points for beetles killed (y_i) are binomially distributed:

$$y_i|\theta_i \sim Bin(n_i, \pi_i)$$

where π_i is the probability of death for beetles given dose x_i. With this, we will study the effect of the choice of link function. We consider binomial GLMs with (i) logit, (ii) probit, (iii) complimentary log-log , and (iv) an alternate logit link function.

$$\pi_i = g^{-1}(\boldsymbol{X\beta}) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \quad \text{(i)}$$

$$\pi_i = g^{-1}(\boldsymbol{X\beta}) = \Phi(\beta_1 + \beta_2 x_i) \quad \text{(ii)}$$

$$\pi_i = g^{-1}(\boldsymbol{X\beta}) = 1 - \exp(-\exp(\beta_1 + \beta_2 x_i)) \quad \text{(iii)}$$

$$\pi_i = g^{-1}(\boldsymbol{X\beta}) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \quad \text{(iv)}$$

The R outputs of the fitted models are summarized in table 1.

Link	Coefficient	Estimate	Std. Error	Lower 95%	Upper 95%
logit	$\hat{\beta}_1$	-60.717	5.18	-71.44	-51.08
	$\hat{\beta}_2$	34.27	2.912	28.85	40.30
probit	$\hat{\beta}_1$	-34.934	2.648	-40.29	-29.93
	$\hat{\beta}_2$	19.798	1.487	16.91	22.74
cloglog	$\hat{\beta}_1$	-39.572	3.240	-46.204	-33.539
	$\hat{\beta}_2$	22.041	1.799	18.690	25.723

Table 1: Summary of the parameter estimates using the `glm()` function in R. For each of the models, β_1 is the intercept coefficient, and β_2 is the coefficient for log dose variables.

(a) Next, we perform residual analysis for each model. The results of the deviance residuals are summarized in Table 2.

Link	Deviance	p-value(sorry)
logit	11.126	0.085
probit	9.987	0.125
cloglog	3.514	0.742

Table 2: Based on the R summary of the `glm()` function, we obtain the deviance residuals for the 3 proposed models

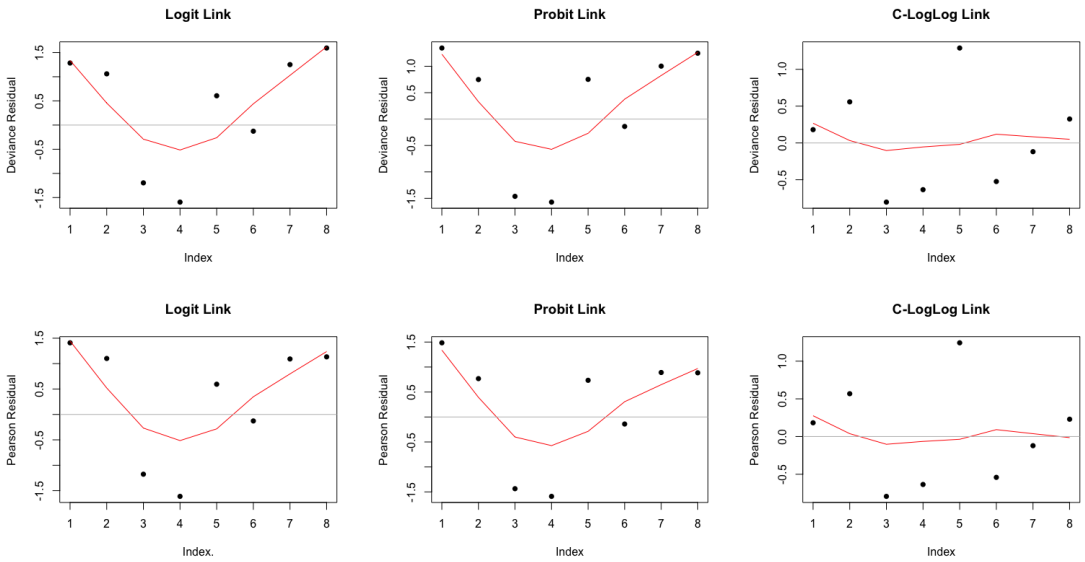

Figure 2: Plots of the deviance residuals and Pearson residuals.

We use the residuals versus order plots to verify the assumption that the residuals are independent from one another. Independent residuals show no trends or patterns when displayed in time order. However, the logit and probit link models appear to have a "v" pattern which is undesirable.

	Logit	Probit	Cloglog	Alt Logit
AIC	41.43	40.32	33.64	35.55
BIC	41.59	40.48	33.80	35.78

Table 3: Model comparison using AIC and DIC

Lastly, table 3 shows the AIC and BIC for the models under each of the aforementioned link functions. Note that the AIC and BIC is lowest for the complimentary loglog link, meaning that it is the preferred model. The alternate logit link model is a close second, but as discussed, there are some drawbacks to this choice of link function.

Based on the results and analysis above, we can conclude that better model is the complementary log-log. But, as one final check, we plot the dose-response curves for each model and compare with the scatter plot of the observed x_i (Figure 4).

(b) The likelihood for the model under the link function in (iv) is

$$L(\beta, \alpha; x, y) = \prod_{i=1}^{n} \binom{m_i}{y_i} \left\{ \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right\}^{\alpha y_i} \left\{ 1 - \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{\alpha} \right\}^{m_i - y_i}$$

The parameter α appears to some kind of tuning parameter. Large α values (much greater than 1) will cause the π_i's to approach zero quickly; while α values between 0 and 1 will increase the value of each π_i. The logit link function is a special case of this link function when $\alpha = 1$. In Binomial regression, this link function is no longer the canonical link function. Fisher's scoring method cannot be used to estimate the model parameters. Newton-Raphson method may be used to estimate the parameters, but convergence will be slower. The residuals versus order plot (Figure 3) verify the assumption that the residuals are independent from one another. There does not appear to be any trends or patterns when displayed in time order.

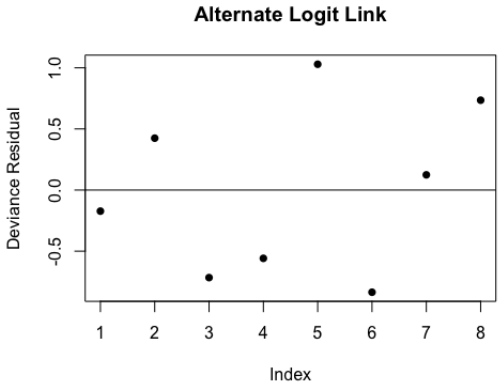

Figure 3: The obtained the deviance residuals from the model with link (iv)

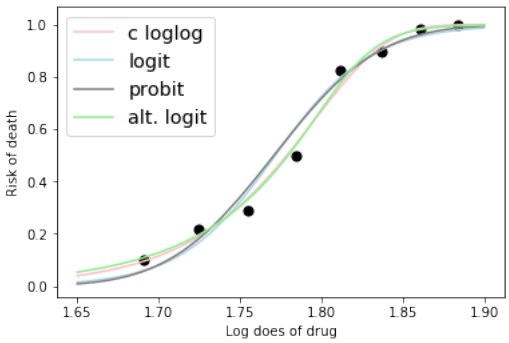

Figure 4: Estimated the dose-response curve $\pi(x_i)$ (over a grid of values x_i for log dose) for each model.

2. First, we plot estimated the dose-response curve $\pi(x_i)$ (over a grid of values x_i for log dose) for complimentary log-log (Model I), logit (Model II), and a modified logit (III) (Figure 4).

   From the plot, note that the dose response curves seem to be skewed for the beetle data. In parts a-c, we consider complimentary log-log, logit, and the alternate/modified logit as defined in problem 1.

   The goal is to design and implement an MCMC method to sample from the posterior distribution of (β_1, β_2). Originally, the goal was to generate posterior samples of $\boldsymbol{\beta}$ using a random walk Metropolis-Hastings algorithm. The gist of which is outlined here:

Algorithm 1: Metropolis-Hastings

Given β^0
for $m = 1$ to M:
    Sample $\beta' \sim$ some prior
    Generate $U \sim \mathcal{U}(0, 1)$
    if $U < min\left\{1, \frac{\pi(\beta'|\boldsymbol{y})}{\pi(\beta^{(k-1)}|\boldsymbol{y})}\right\}$, let $\beta^{(k)} = \beta'$
    else let $\beta^{(k)} = \beta^{(k-1)}$
end for

   However, as part (c) will show, the M-H algorithm performs poorly for the "modified" logit. An alternative proposal, was a Slice sampling algorithm (Neal, 1997), as well as a Hamiltonian MC algorithm (Hoffman & Gelman 2011).

Algorithm 1 Hamiltonian Monte Carlo

Given θ^0, ϵ, L, $\mathcal{L}$, M:
for $m = 1$ to M **do**
  Sample $r^0 \sim \mathcal{N}(0, I)$.
  Set $\theta^m \leftarrow \theta^{m-1}, \tilde{\theta} \leftarrow \theta^{m-1}, \tilde{r} \leftarrow r^0$.
  for $i = 1$ to L **do**
    Set $\tilde{\theta}, \tilde{r} \leftarrow \text{Leapfrog}(\tilde{\theta}, \tilde{r}, \epsilon)$.
  end for
  With probability $\alpha = \min\left\{1, \frac{\exp\{\mathcal{L}(\tilde{\theta}) - \frac{1}{2}\tilde{r}\cdot\tilde{r}\}}{\exp\{\mathcal{L}(\theta^{m-1}) - \frac{1}{2}r^0\cdot r^0\}}\right\}$, set $\theta^m \leftarrow \tilde{\theta}$, $r^m \leftarrow -\tilde{r}$.
end for

function Leapfrog(θ, r, ϵ)
Set $\tilde{r} \leftarrow r + (\epsilon/2)\nabla_\theta \mathcal{L}(\theta)$.
Set $\tilde{\theta} \leftarrow \theta + \epsilon\tilde{r}$.
Set $\tilde{r} \leftarrow \tilde{r} + (\epsilon/2)\nabla_\theta \mathcal{L}(\tilde{\theta})$.
return $\tilde{\theta}, \tilde{r}$.

An appendix of all of the trace and posterior density plots are attached to this report.

a) First, we consider a Bayesian binomial GLM with a complementary log-log link.

We will model the number of deaths as a random sample from a binomial distribution, where n is the number of beetles and π the probability of a beetle dying. As x increases the number of beetles dying seems to increase, and since $\pi(x_i)$ is a probability, we use the following model:

$$\begin{aligned} y_i &\sim \text{Bin}(m_i, \pi(x_i)) \\ \pi(x_i) &= 1 - \exp\{\exp(\beta_1 + \beta_2 x_i)\}. \end{aligned} \qquad \text{(Model I)}$$

We apply a Slice sampling algorithm to sample from the posterior distribution of (β_1, β_2). Using various prior specifications for β_1 and β_2, we analyze the posterior distributions(Table 4).

The idea behind slice sampling is to simulate from a distribution by sampling uniformly from a region underneaeth the graph of the density function.

As expected, if the prior information is strong, the estimate for $\boldsymbol{\beta}$ will get closer to the prior mean for β. (The way I think of this is: if we use a strong prior, it will "drag down" the mean). When we use a flat prior, we get the posterior mean for $\boldsymbol{\beta}$ to be the same as the MLE's for $\boldsymbol{\beta}$ in the Frequentist approach (Figures 5 - 7).

Prior	Mean	SD	hpd 2.5%	hpd 97.5%	Rhat
$\beta_1 \sim$ Flat Prior	-39.892852	3.247513	-46.128413	-33.310861	1.000135
$\beta_2 \sim$ Flat Prior	22.218685	1.802725	18.562274	25.679998	1.000127
$\beta_1 \sim N(1, 100)$	-39.790227	-39.790227	-46.088329	-33.608019	1.000636
$\beta_2 \sim N(1, 100)$	22.161363	1.778240	18.753223	25.688533	1.000631
$\beta_1 \sim N(0, 1)$	-5.407784	0.778827	-6.969091	-3.926571	1.000038
$\beta_2 \sim N(0, 1)$	3.014853	0.437669	2.202568	3.908197	1.000025

Table 4: Summary of the posterior distribution of the coefficients in Model I (cloglog) with various prior specifications.

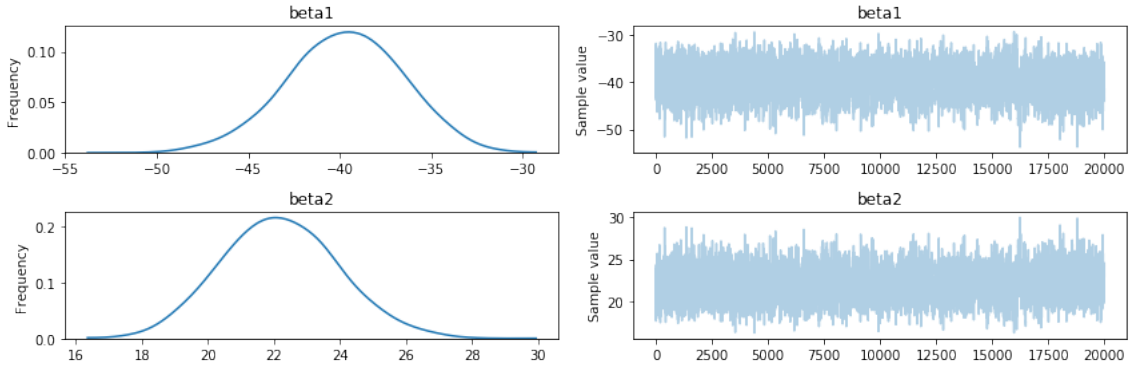

Figure 5: Posterior distribution of (β_1, β_2 cloglog link model using Normal priors on β). After 20000 iterations, the parameters converge properly to the global maxima

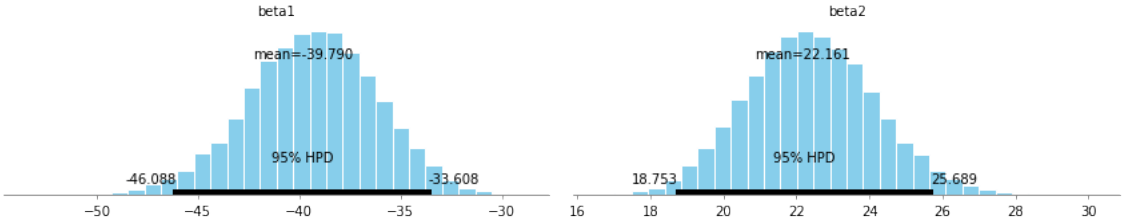

Figure 6: Posterior distribution of (β_1, β_2 for Model I using N(1,100) priors. After 20000 iterations, the parameters converge properly to the global maxima

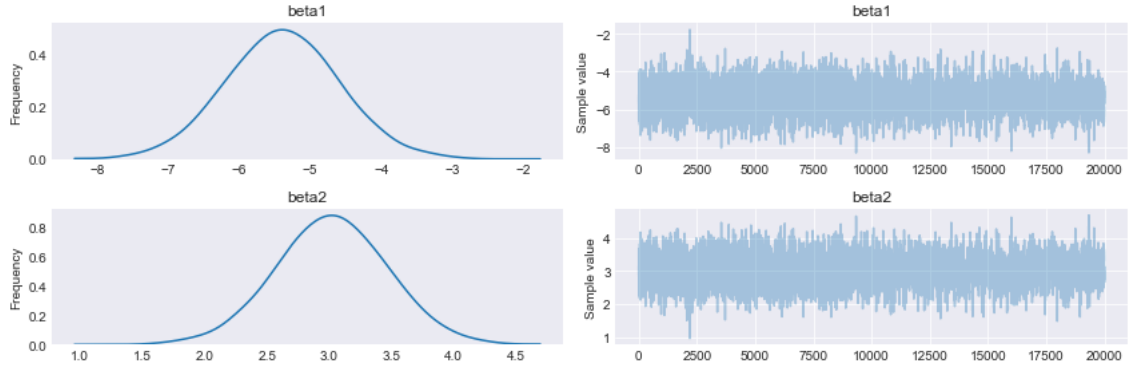

Figure 7: Posterior distribution of (β_1, β_2 for the model with cloglog link using N(0,1) priors). After 20000 iterations, the parameters converge properly to the global maxima

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

$$\pi(x_i) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \qquad \text{(Model II)}$$

Again using a Slice sampling algorithm and specifying flat (table 5 and figure 8) and Normal priors on $\boldsymbol{\beta}$, we analyze the results.

	mean	sd	hpd 2.5%	hpd 97.5%
β_1	-61.320536	5.177014	-71.881831	-51.684644
β_2	34.610170	2.910742	29.067387	40.425223

Table 5: Posterior summary for β in Model II (logit link)

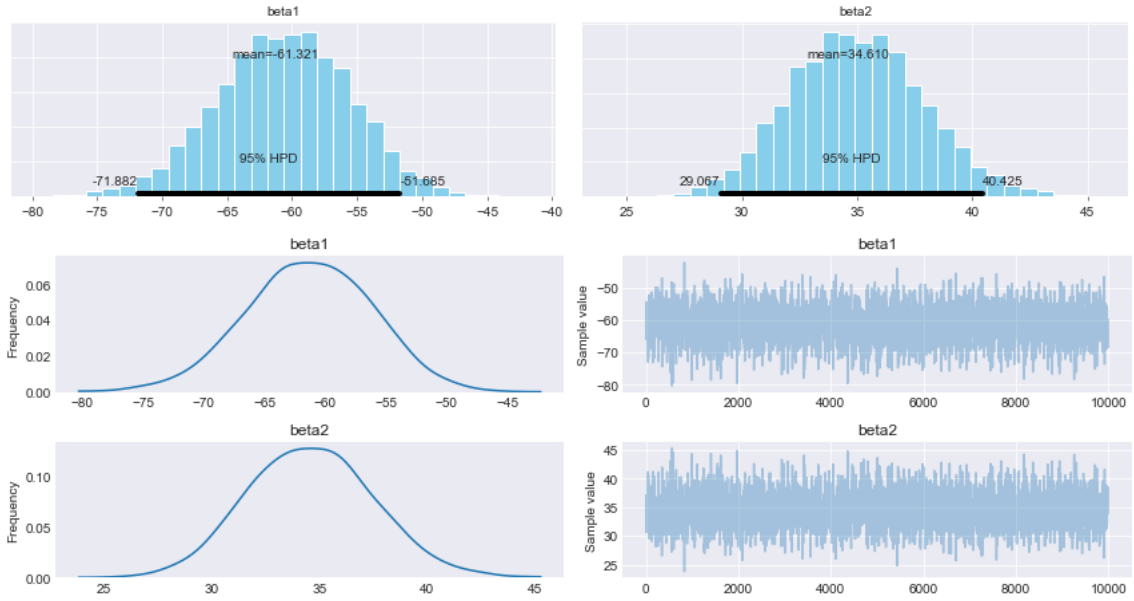

Figure 8: Posterior distributions of (β_1, β_2) for logit link using flat priors. After 10000 iterations, the parameters converge properly to the global maxima.

c) As a third model, we consider the binomial GLM with the parametric link given by

$$\pi(x) = \frac{\exp(\alpha\,(\beta_1 + \beta_2 x))}{(1 + \exp(\beta_1 + \beta_2 x))^{\alpha}} \qquad \text{(Model III)}$$

For this model, we again specify flat and Normal priors for $\boldsymbol{\beta}$ parameters. Furthermore, we consider a Gamma prior on α with various combinations of rate and shape parameters(Table 6).

As stated before, with the Metropolis-Hastings algorithm, convergence was poor (Appendix Figure 11). Thus, a Hamiltonian MC algorithm was used. Ultimately, each of the initial rate and shapes values for α yield very similar results. These results are summarized in table 6 with a traceplot included in figure 10.

Table 6: The posterior distributions of $(\beta_1, \beta_2, \alpha)$ for Model III using various priors.

Prior	Mean	SD	Lower 95%	Upper 95%
$\beta_1 \sim$ Flat	-138.461623	49.544304	-237.242697	-69.796848
$\beta_2 \sim$ Flat	75.938081	26.722983	38.961303	129.024600
$\alpha \sim Gamma(1,1)$	0.264578	0.115601	0.066012	0.493207
$\beta_1 \sim N(0,100)$	-112.371063	27.024615	-166.196757	-63.439118
$\beta_2 \sim N(0,100)$	61.868392	14.575325	35.437595	90.795099
$\alpha \sim Gamma(2,1)$	0.335498	0.136850	0.124579	0.614809

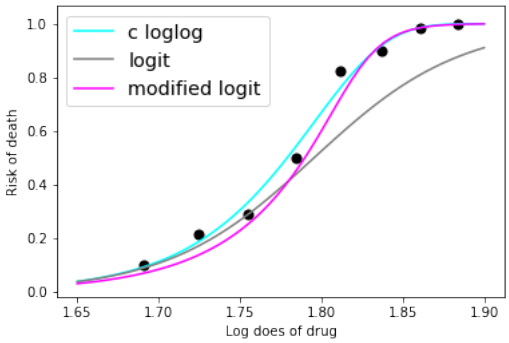

Figure 9: Dose response curves for each model. The modified appears to best fit the data, with the cloglog a close second

d) We now focus on the following models: $(\mathcal{M}_1)$ Cloglog link model having $\boldsymbol{\beta} \sim N(\mathbf{0}, 100\mathbf{I}_2)$, where $\mathbf{I}_2$ is a 2×2 identity matrix, the $(\mathcal{M}_2)$ Logit link model having flat priors, and the $(\mathcal{M}_3)$ alternate logit link having $N(\mathbf{0}, 100\mathbf{I}_2)$ priors on $\boldsymbol{\beta}$ and $Gamma(1,1)$ prior on α.

Figure 9 shows the dose-response curves for these models $\mathcal{M}_1$ - $\mathcal{M}_3$ (with response being the probability of death) for each model. By examining how closely the curves pass through the actual observed data, we can get a rough idea of how good the model fits. Based on this, we determine that the curves for the alternative logit and complementary log-log fits the data best. We run an additional tests to be certain.

Model	$L_q(0.5)$
$\mathcal{M}_1$	146.187
$\mathcal{M}_2$	230.373
$\mathcal{M}_3$	133.969

Table 7: Quadratic Loss L measure

Let $\boldsymbol{z} = (z_1, \ldots z_n)'$ denote the future response vector with the same sampling density $(y|\beta)$. Gelfand & Ghosh (1998) and Ibraham et. al (2001) consider the quadratic loss L measure

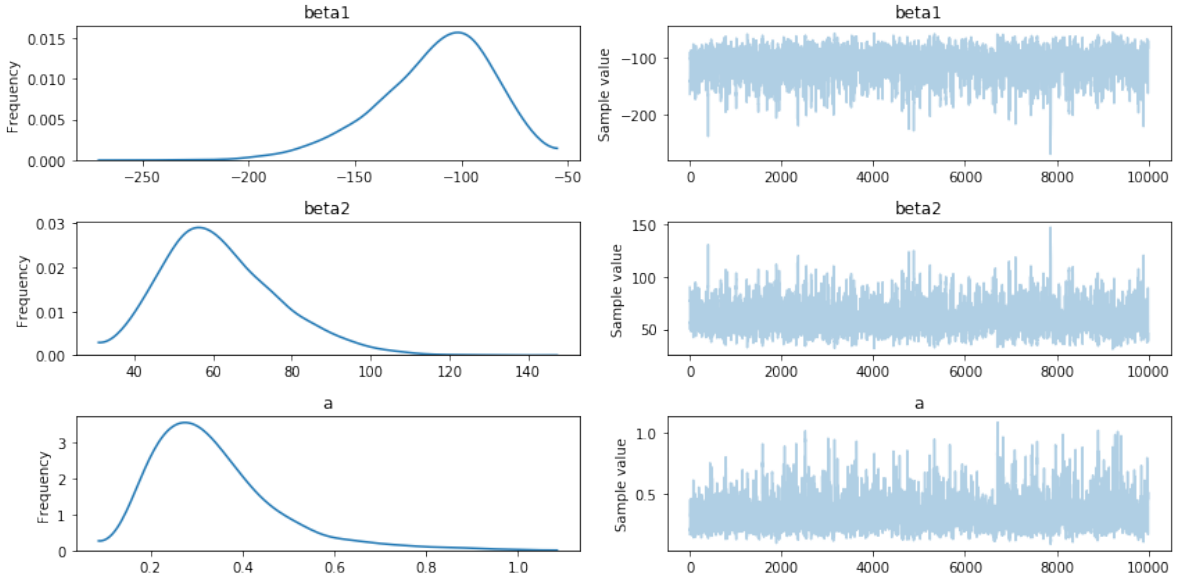

Figure 10: Posterior distributions of $(\beta_1, \beta_2, \alpha)$ for $\mathcal{M}_3$. After 10000 iterations, the parameters converge properly to the global maxima. It was, however, ever so slightly more computationally exhaustive, taking a total of 8 minutes and 25 seconds to sample as opposed to < 1 minute for $\mathcal{M}_1$ and $\mathcal{M}_2$.

$$L_q(v) = \sum_{i=1}^{n} Var(z_i|x_i, D) + v \sum_{i=1}^{n} (\mu_i - y_i)^2$$

Where $\mu_i = E(z_i|x_i, D)$. Essentially, this expression is the summation of some penalty term for model complexity and a goodness-of-fit term. A larger choice of v penalizes models that are more complex. Thus, we choose $v = 0.5$. For each of the models $\mathcal{M}_1 - \mathcal{M}_3$, the quadratic loss L measure is computed in the table 7.

3 a) We Consider the inverse Gaussian distribution with density function

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

to show that the inverse Gaussian distribution is a member of the exponential dispersion family.

$$\begin{aligned}
f(y|\mu,\phi) &= (2\pi\phi y^3)^{-1/2}\exp\left\{-\frac{(y-\mu)^2}{2\phi\mu^2 y}\right\} \\
&= \exp\left\{-\frac{1}{2}\log(2\pi\phi y^3) + \frac{-y^2+2\mu y-\mu^2}{2\phi\mu^2 y}\right\} \\
&= \exp\left\{\frac{y(-1/2\mu^2)-(-1/\mu)}{\phi} + \left[-\frac{1}{2}\left(\frac{1}{\phi y}+\log(2\pi\phi y^3)\right)\right]\right\} \\
&= \exp\left\{\frac{y\theta-b(\theta)}{\phi}+c(y,\phi)\right\}
\end{aligned}$$

Therefore, the inverse gaussian distribution is a member of the EDF family, with

$$\begin{aligned}
\phi &= \phi \\
\theta &= -1/2\mu^2 \\
b(\theta) &= -\sqrt{-2\theta} \\
c(y,\phi) &= -\frac{1}{2}\left(\frac{1}{\phi y}+\log(2\pi\phi y^3)\right)
\end{aligned}$$

Next, we show that μ is the mean of the distribution and obtain the variance function.

$$\begin{aligned}
\mathrm{E}\left[y|\mu,\phi\right] &= b'(\theta) \\
&= -\frac{1}{2}(-2\theta)^{-1/2}(-2) \\
&= (-2\theta)^{-1/2} \\
&= \mu \\
\mathrm{Var}\left[y|\mu,\phi\right] &= \phi\ b''(\theta) \\
&= -\frac{\phi}{2}(-2\theta)^{-3/2}(-2) \\
&= \phi(-2\theta)^{-3/2} \\
&= \phi(\mu^{-1/2})^{-3/2} \\
&= \phi\mu^3
\end{aligned}$$

b) The scaled deviance is

$$
\begin{aligned}
D^* &= -2\log(\lambda) \\
&= \frac{2}{\phi}\sum_{i=1}^{n} w_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right\} \\
&= \frac{2}{\phi}\sum_{i=1}^{n} \left\{ y_i \left[\left(\frac{-1}{2\tilde{\mu}_i^2}\right) - \left(\frac{-1}{2\hat{\mu}_i^2}\right) \right] - \left(\frac{-1}{\tilde{\mu}_i}\right) + \left(\frac{-1}{\hat{\mu}_i}\right) \right\} \\
&= \frac{2}{\phi}\sum_{i=1}^{n} \left\{ y_i \left[\left(\frac{-1}{2y_i^2}\right) - \left(\frac{-1}{2g^{-1}(x_i'\hat{\beta})^2}\right) \right] - \left(\frac{-1}{y_i}\right) + \left(\frac{-1}{g^{-1}(x_i'\hat{\beta})}\right) \right\} \\
&= \frac{2}{\phi}\sum_{i=1}^{n} \left\{ \frac{-1}{2y_i} + \frac{y_i}{2g^{-1}(x_i'\hat{\beta})^2} + \frac{1}{y_i} - \frac{1}{g^{-1}(x_i'\hat{\beta})} \right\} \\
&= \frac{2}{\phi}\sum_{i=1}^{n} \left\{ \frac{1}{2y_i} + \frac{y_i}{2g^{-1}(x_i'\hat{\beta})^2} - \frac{1}{g^{-1}(x_i'\hat{\beta})} \right\} \\
&= \frac{2}{\phi}\sum_{i=1}^{n} \left\{ \frac{g^{-1}(x_i'\hat{\beta})^2 + y_i^2 - 2y_i g^{-1}(x_i'\hat{\beta})}{2y_i g^{-1}(x_i'\hat{\beta})^2} \right\} \\
&= \frac{2}{\phi}\sum_{i=1}^{n} \left\{ \frac{\left[y_i - g^{-1}(x_i'\hat{\beta}) \right]^2}{2y_i g^{-1}(x_i'\hat{\beta})^2} \right\} \\
&= \frac{1}{\phi}\sum_{i=1}^{n} \frac{\left[y_i - g^{-1}(x_i'\hat{\beta}) \right]^2}{y_i \; g^{-1}(x_i'\hat{\beta})^2}
\end{aligned}
$$

1 Appendix

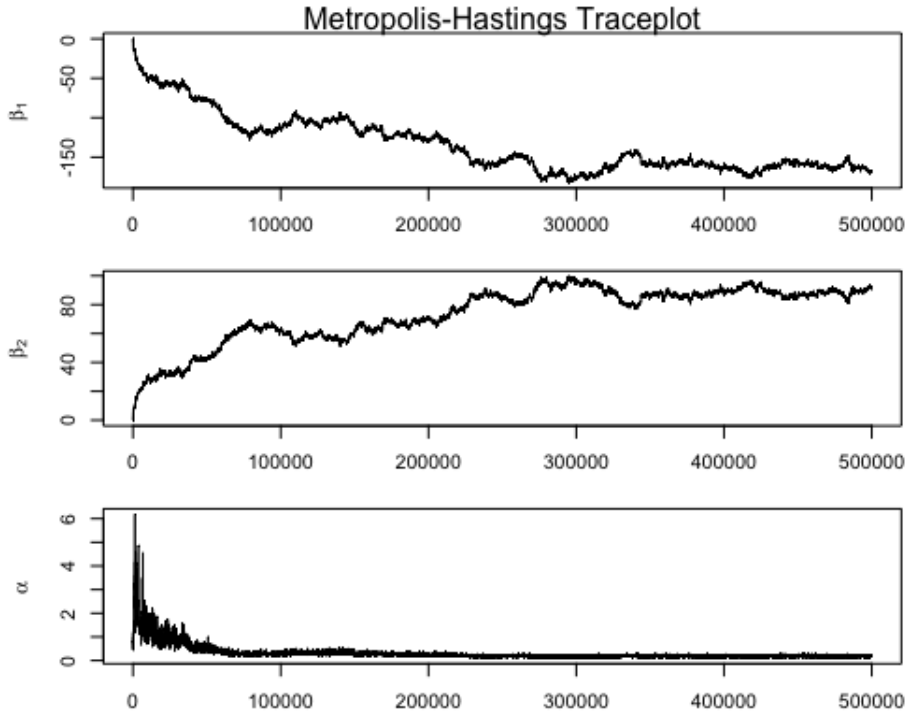

Figure 11: Trace plot of the posterior samples for Model III using the Metropolis-Hastings algorithm does not converge. The chain appears to be converging, but only after 500,000 iterations, which is computationally exhaustive compared to the Slice and Hamiltonian MC algorithms.

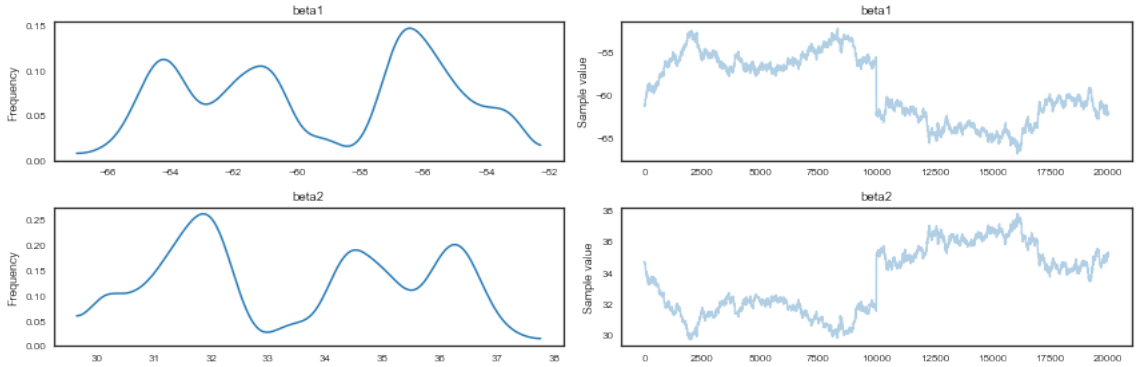

Figure 12: Metropolis-Hastings Algorithm applied to glm with logit link and $\boldsymbol{\beta} \sim N_2(\mathbf{0}, 100\boldsymbol{I}_2)$ prior. After 10,000 iterations, convergence and posterior densities are poor

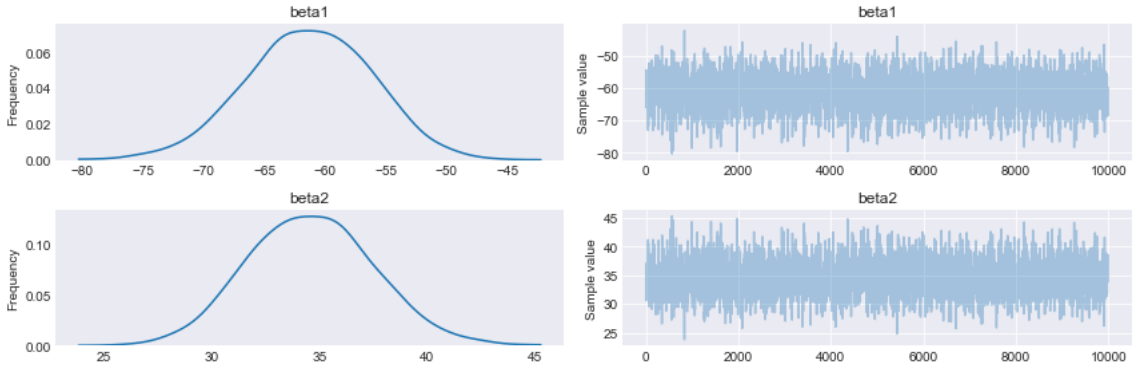

Figure 13: Flat priors for cloglog using Slice sampling algorithm

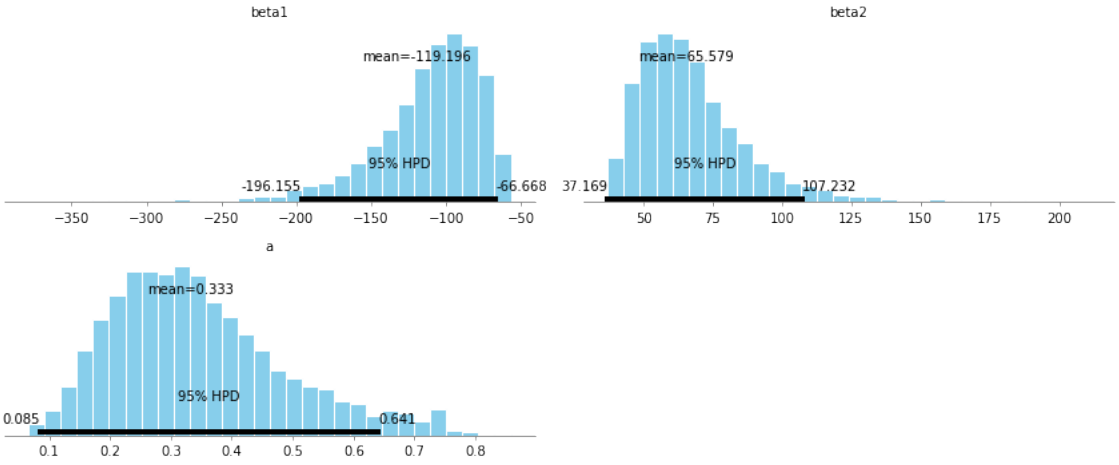

Figure 14: Flat priors for alternate logit link using Slice sampling