SDS 383D: Modeling II **Spring 2017**

Section 3: Bayesian GLMs

3.1 Modeling non-Gaussian observations

So far, we've assumed real-valued observations. In this setting, our likelihood model is a univariate normal, parametrized by a mean $x_i^T\beta$ and some precision that does not directly depend on the value of x_i. In general, $x_i^T\beta$ will take values in $\mathbb{R}$

If we don't want to use a Gaussian likelihood, we typically won't be able to parametrize our data using a real-valued parameter. Instead, we must transform it via an appropriate link function. This is, in essence, the generalized linear model.

As a first step into other types of data, let's consider binary valued observations. Here, the natural likelihood model is a Bernoulli random variable; however we cannot directly parametrize this by $x_i^T\beta$. Instead, we must transform $x_i^T\beta$ to lie between 0 and 1 via some function $g^{-1} : \mathbb{R} \to (0, 1)$. We can then write a linear model as

$$\begin{aligned} y_i|p_i \sim & \text{Bernoulli}(p_i) \\ p_i = & g^{-1}(x_i^T\beta) \\ \beta|\theta \sim & \pi_\theta(\beta) \end{aligned}$$

where $\pi_\theta(\beta)$ is our choice of prior on β. Unfortunately, there is no choice of prior here that makes the model conjugate.

Let's start off with a normal prior on β. One appropriate function for g^{-1} is the inverse CDF of the normal distribution – known as the probit function. This is equivalent to assuming our data are generated according to

$$\begin{aligned} y_i = & \begin{cases} 1 & if z > 0 \\ 0 & \text{otherwise} \end{cases} \\ z_i \sim & \text{N}(x_i^T\beta, \tau^2) \end{aligned}$$

If we put a normal-inverse gamma prior on β and τ, then we have a *latent* regression model on the (x_i, z_i) pairs, that is idential to what we had before! Conditioned on the z_i, we can easily sample values for β and τ.

Exercise 3.1 *To complete our Gibbs sampler, we must specify the conditional distribution $p(z_i|x_i, y_i, \beta, \tau)$. Write down the form of this conditional distribution, and write a Gibbs sampler to sample from the posterior distribution. Test it on the dataset* `pima.csv`*, which contains diabetes information for women of Pima indian heritage. The dataset is from National Institute of Diabetes and Digestive and Kidney Diseases, full information and explanation of variables is available at* `http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`.

Answer: The conditional distribution z is given by:

$$p(z_i|x_i, y_i, \beta, \tau) = \begin{cases} N_{(0,\infty)}(x_i^T\beta, \tau^2) & if y_i = 1 \\ N_{(-\infty,0)}(x_i^T\beta, \tau^2) & if y_i = 0 \end{cases} \tag{3.1}$$

It was implemented in MATLAB on the `pima.csv` dataset, and used to predict the value of y_i. The constant parameters were set at: $a_o = 0.1, b_o = 5, K_o = 10 * I_{dxd}, K_n = X'X + K_o, a_n = a_o + d/2, b_n = b_o$. The distribution was sampled $N = 1000$ times according to the equations below:

$$\tau = Gamma(a_n, b_n) \tag{3.2}$$
$$\beta = N(\beta_n, (K_n\tau)^{-1} \tag{3.3}$$
$$Z = N_{trunc}(\beta, \tau) \tag{3.4}$$
$$\beta_n = (K_n)^{-1}(X'Z + K_o * \beta_o^T) \tag{3.5}$$

Where N_{trunc} is given by (3.1), calculated using the `rmvnrnd.m` function. The y-values were predicted using:

$$y_i = \begin{cases} 1 & if x_i\bar{\beta} \leq 0 \\ 0 & if x_i\bar{\beta} > 0 \end{cases} \tag{3.6}$$

where $\bar{\beta}$ is the average beta over the iterations (excluding the first 185 iterations). The validity of the model was evaluated based on the fraction of correct values, C:

$$C = 1 - \frac{incorrect}{total} \tag{3.7}$$

The model was identified 5 times, with C values: $C = \{0.7604, 0.8268, 0.7214, 0.9323, 0.9010\}$.

Another choice for $g^{-1}(\theta)$ might be the logit function, $\frac{1}{1+e^{-x^T\beta}}$. In this case, it's less obvious to see how we can construct an auxilliary variable representation (it's not impossible! See **?**. But for now, we'll assume we haven't come up with something). So, we're stuck with working with the posterior distribution over β.

Exercise 3.2 *Sadly, the posterior isn't in a "known" form. As a starting point, let's find the maximum a posteriori estimator (MAP). The dataset "titantic.csv" contains survival data from the Titanic; we're going to look at probability of survival as a function of age. For now, we're going to assume the intercept of our regression is zero – i.e. that β is a scalar. Write a function (that can use a black-box optimizer! No need to reinvent the wheel. It shouldn't be a long function) to estimate the MAP of β. Note that the MAP corresponds to the frequentist estimator using a ridge regularization penalty.*

Answer: The MAP was estimated by minimizing the function:

$$\max_{\beta} J = -(-\frac{\beta^2}{2\sigma} - \sum_i [y_i log(1 + exp[-\beta X]) + (1 - y_i) log(1 + exp[\beta X])] + \lambda\beta^2 - t \tag{3.8}$$

Where $y_i = 0$ if the person did not survive and $y_i = 1$ if the person survived. The resulting optimal value of β is $\beta^* = -0.011$.

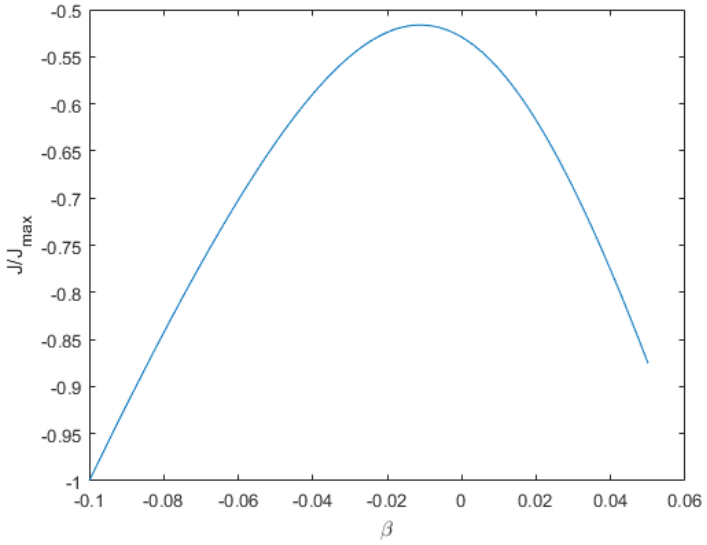

Figure 3.1: Plot showing the posterior PDF

Exercise 3.3 *OK, we don't know how to sample from the posterior, but we can at least look at it. Write a function to calculate the posterior pdf $p(\beta|\mathbf{x}, \mathbf{y}, \mu, \sigma^2)$, for some reasonable hyperparameter values μ and θ (up to a normalizing constant is fine!). Plot over a reasonable range of β (your MAP from the last question should give you a hint of a reasonable range).*

Answer: The function is plotted in Figure 3.1.

The Laplace approximation is a method for approximating a distribution with a Gaussian, by matching the mean and variance at the mode.[1] Let P^* be the (unnormalized) PDF of a distribution we wish to approximate. We start by taking a Taylor expansion of the log (unnormalized) PDF at the global maximizing value x^*

$$\log P^*(x) \approx \log P^*(x^*) - \frac{c}{2}(x - x^*)^2$$

where $c = -\frac{\delta^2}{\delta x^2} \log P^*(x)\Big|_{x=x^*}$.

We approximate P^* with an unnormalized Gaussian, with the same mean and variance as P^*:

$$Q^*(x) = P^*(x^*) \exp\left\{-\frac{c}{2}(x - x^*)^2\right\}$$

Exercise 3.4 *Find the mean and precision of a Gaussian that can be used in a Laplace approximation to the posterior distribution over β.*

Answer: The steps below show how to solve for c.

[1] More generally, the Laplace approximation is used to approximate integrands of the form $\int_A e^{Nf(x)}dx...$ but for our purposes we will always be working with PDFs.

$$c = -\frac{\partial^2}{\partial \beta^2} \log P^*(\beta)\Big|_{\beta=\beta^*} \tag{3.9}$$

$$\frac{\partial}{\partial \beta_k} = \frac{-\beta_k}{\sigma^2} + \frac{\sum_i y_i x_i}{exp[\sum_j x_{ij}\beta_j]} - \frac{\sum_i (1-y_i) x_i exp[\sum_j x_{ij}\beta_j]}{1 + exp[\sum_j x_{ij}\beta_j]} \tag{3.10}$$

$$\frac{\partial^2}{\partial \beta_k^2} = -\frac{1}{\sigma^2} - \frac{\sum_i x_{ik}^2 exp[\sum_j x_{ij}\beta_j]}{1 + exp[\sum_j x_{ij}\beta_j]} \tag{3.11}$$

Exercise 3.5 *That's all well and good... but we probably have a non-zero intercept. We can extend the Laplace approximation to multivariate PDFs. This amounts to estimating the precision matrix of the approximating Gaussian using the negative of the Hessian – the matrix of second derivatives*

$$H_{ij} = \frac{\delta^2}{\delta x_i \delta x_j} \log P^*(x)\Big|_{x=x^*}$$

Use this to approximate the posterior distribution over β. Give the form of the approximating distribution, plus 95% marginal credible intervals for its elements.

Answer: Putting the equation in vector form, where β is a 1x2 system, and $\sigma_{X\beta} = \frac{1}{1+exp(\beta X^T)}$.

$$y^T log(\sigma_{X\beta}) + (1-y)^T(1-\sigma_{X\beta}) - \frac{\lambda \beta^T \beta}{2} \tag{3.12}$$

$$\frac{\partial}{\partial \beta} = X^T y^T - X^T \sigma_{x\beta} - \lambda\beta \tag{3.13}$$

$$\frac{\partial}{\partial \beta \beta^T} = -X^T(\sigma_{x\beta}(1-\sigma_{x\beta})X - \lambda = H(\beta) \tag{3.14}$$

The function to maximize is the negative of equation (3.12).

With $H = [45.6, 2742; 2742, 201370]$, $\beta_2^* = [-0.1519 - 0.0090]$, and $CI = [(-0.8302, 0.5264), (-0.0192, 0.0012)]$

The posterior is then approximated by the following equation:

$$\left(\frac{-H}{2\pi}\right)^{1/2} exp\Big[-\frac{1}{2}(\beta-\beta^*)^T(-H)(\beta-\beta^*)\Big] \tag{3.15}$$

Let's try the same thing with a Poisson likelihood. Here, the obvious transformation is to let $g^{-1}(\theta) = e^\theta$, i.e.

$$\begin{aligned} y_i|p_i \sim & \text{Poisson}(\lambda_i) \\ \lambda_i = & e^{x_i^T \beta} \end{aligned}$$

We're going to work with the dataset `tea_discipline_oss.csv`, a dataset gathered by Texas Appleseed, looking at the number of out of school suspensions (ACTIONS) accross schools in Texas. The data is censored for privacy reasons – data points with fewer than 5 actions are given the code "-99". For now, we're going to exclude these data points.

Exercise 3.6 *We're going to use a Poisson model on the counts. Ignoring the fact that the data is censored, why is this not quite the right model? Hint: there are several answers to this – the most fundamental involve considering the support of the Poisson.*

Answer: The poisson model isn't quite the correct choice because the data is sparse and over-dispersed, i.e. the data may be multimodal. In addition, the data is skewed to the right.

Exercise 3.7 *Let's assume our only covariate of interest is GRADE*[2] *and put a normal prior on β. Using a Laplace approximation and an appropriately vague prior, find 95% marginal credible intervals for the entries of β. You'll probably want to use an intercept.*

Answer: In order to find the MAP, the log-likelihood function for the poisson distribution was maximized.

$$logP(\beta|x,y) = \sum_i (y_i \beta^T x_i - e^{\beta^T x_i} - log(y_i!)) \tag{3.16}$$

Since the function will be maximized with respect to β, any terms not dependent on β were dropped. The equation was then put into vector form and a penalty term, $\lambda = 1$, was added to penalize the size of the coefficients, $\beta = \{\beta_{int}, \beta_{grade}\}$:

$$\max_\beta logP = Y^T X\beta - \sum_i e^{X_i\beta} - \frac{\lambda}{2}\beta^T\beta \tag{3.17}$$

The optimal values of β are $\beta = \{2.3894, 0.0503\}$, with 95% credible intervals of (2.3794,2.3994) and (0.0490,0.0515), respectively.

It is important to note that the value of λ affects the optimal β values. For example, if the same problem is solved with $\lambda = 10,000$, the optimal β values are $\beta = \{1.8378, 0.1124\}$ with credible intervals of (1.8276,1.8480) and (0.112,0.1136), respectively.

Exercise 3.8 (Optional) *Repeat the analysis using a set of variables that interest you.*

Even though we don't have conjugacy, we can still use MCMC methods – we just can't use our old friend the Gibbs sampler. Since this isn't an MCMC course, let's use STAN, a probabilistic programming language available for R, python and Matlab. I'm going to assume herein that we're using RStan, and give appropriate scripts; it should be fairly straightforward to use if you're an R novice, or if you want to use a different language, there are hints on translating to PyStan at `http://pystan.readthedocs.io/en/latest/differences_pystan_rs` and info on MatlabStan (which seems much less popular) at `http://mc-stan.org/users/interfaces/matlab-stan`.

Exercise 3.9 *Download the sample STAN script* ***`poisson.stan`*** *and corresponding R script* ***`run_poisson_stan.R`****. The R script should run the regression vs GRADE from earlier (feel free to change the prior parameters). Run it and see how the results differ from the Laplace approximation. Modify the scripy to include more variables, and present your results.*

Answer:

[2] I have manually replaced Kindergarten and Pre-K with Grades 0 and -1, respectively.

Stan was used to develop a model for the Texas Appleseed School Suspension data. In the first model identification, just grade was used, along with an intercept. The results are given in the following tables and plots:

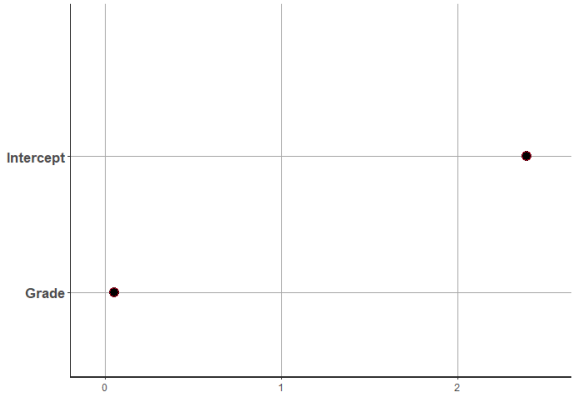

Figure 3.2: Plot showing the mean calculated by Stan, where actual values are 2.39 and 0.05 for the intercept and the grade, respectively.

The traceplot includes the sampled values of the parameters over time for each chain:

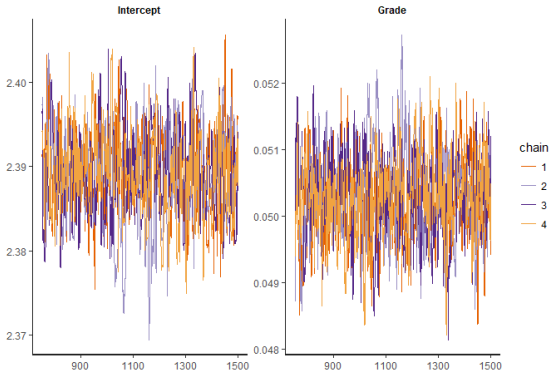

Figure 3.3: Traceplot of the Stan model.

The script was modified to include more variables in the set, with the results below.

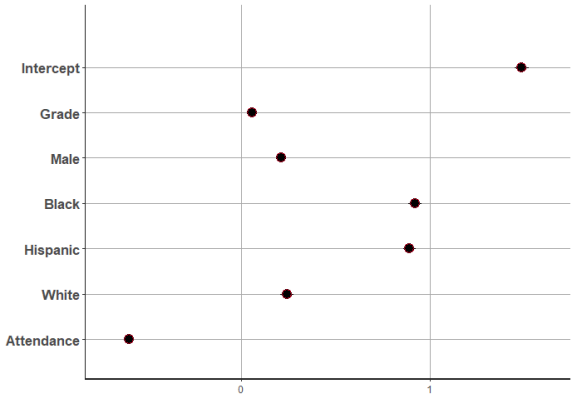

Figure 3.4: Plot showing the mean calculated by Stan, where actual values are 1.49, 0.06, 0.21, 0.92, 0.89, 0.24, -0.59, for the intercept, grade, male, black, hispanic, white, and attendance, respectively.

The traceplot includes the sampled values of the parameters over time for each chain:

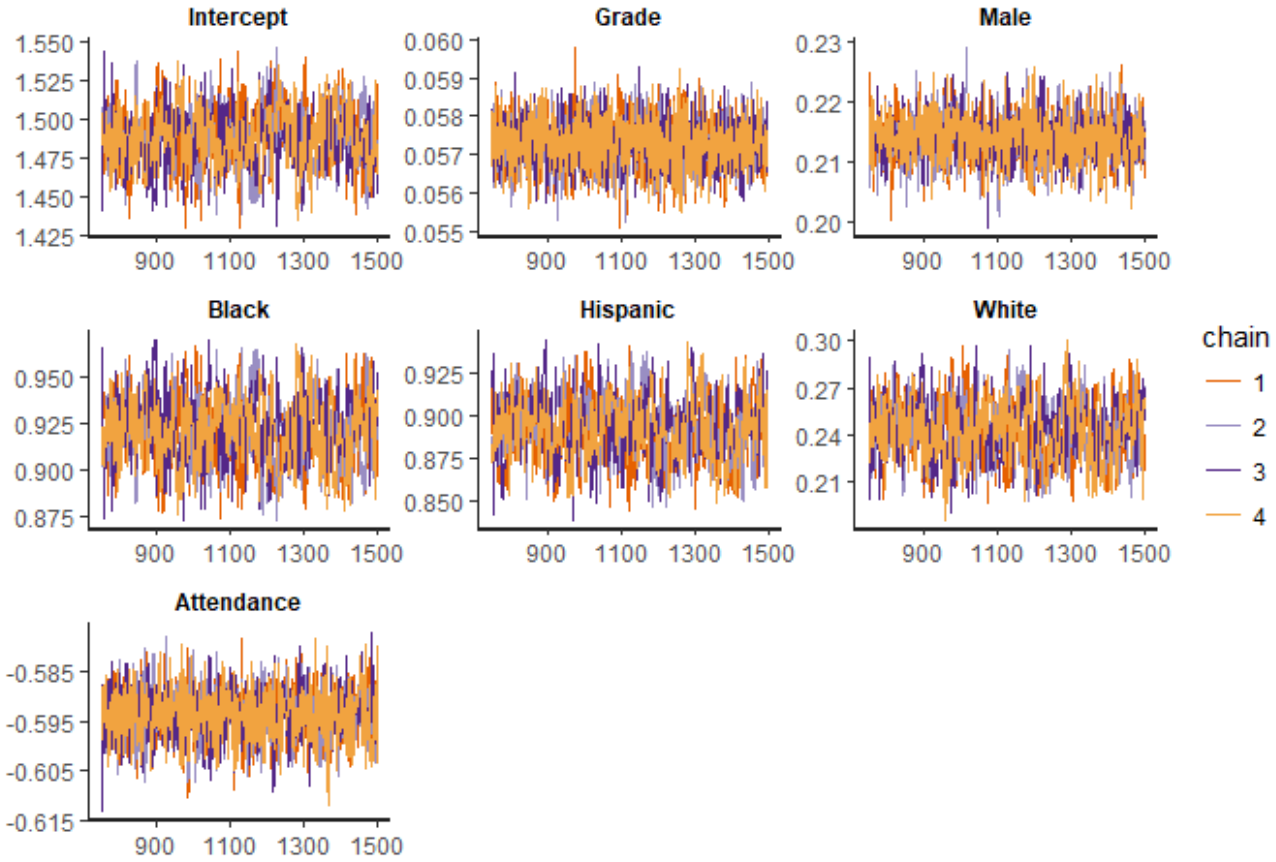

Figure 3.5: Traceplot of the Stan model.

The model above has an NMSE of 1.8%.

Exercise 3.10 *Consider ways you might improve your regression (still, using the censored data) - while staying in the GLM framework. Ideas might include hierarchical error modeling (as we looked at in the last set of exercises), interaction terms... or something else! Looking at the data may give you inspiration. Implement this in STAN.*

Answer: In order to improve the model, I added an error term to each output point in the model, ϵ_i.

$$Y_i \sim Poisson(e^{X_i\beta} + e^{\epsilon_i}) \tag{3.18}$$

Plots are shown below for the β values and the first three ϵ values.

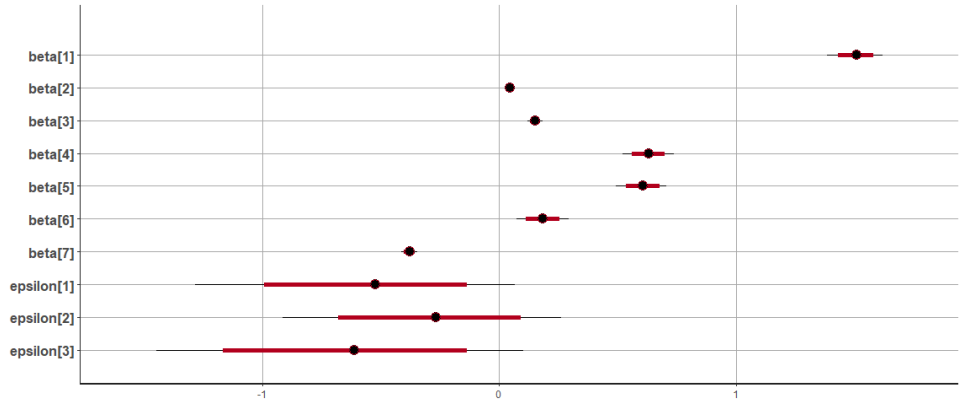

Figure 3.6: Plot showing the mean calculated by Stan.

The traceplot includes the sampled values of the parameters over time for each chain:

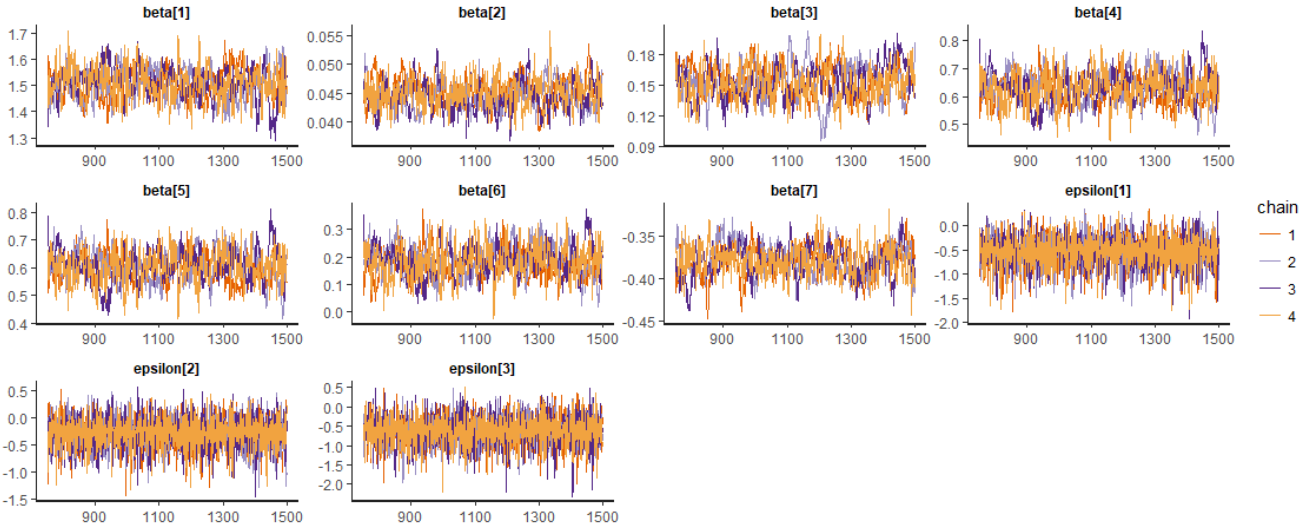

Figure 3.7: Traceplot of the Stan model.

Surprisingly, this method did not improve the model, where the NMSE was 71%.

Exercise 3.11 *We are throwing away a lot of information by not using the censored data. Come up with a strategy, and write down how you would alter your model/sampler. Bonus points for actually implementing it in STAN (hint: look up the section on censored data in the STAN manual).*

Answer: In order to deal with the censored data, I would split the data up into two sections, censored and un-censored and identify a stan model based on these two subsets, where the upper-bounded censored data would be treated by stan as data which has an upper limit but is unknown. The data inputted in the

stan model would be the difference between the censored output data points and the censored threshold, thereby counting the censored data as "observed" at zero, and the uncensored data as offset by the censored threshold.