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 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 $p(y_i|x_i, z_i, \beta, \tau)$. 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`*.*

Begin Solution: The posterior distribution of the latent variable z_i is

$$z_i|x_i, y_i, \beta, \omega \sim \begin{cases} N_{(0,+\infty)}\left(x_i^T\beta, \omega^2\right) \text{ if } y_i = 1 \\ N_{(-\infty,0)}\left(x_i^T\beta, \omega^2\right) \text{ if } y_i = 0 \end{cases}$$

To get samples from the posterior $p\left(\beta|x, y, z, \omega\right)$, choose the following as prior:
$a = 0.1, b = 5, k = I_n$ so that $a_n = a + \frac{n}{2}, K_n = X^TX + K, \mu_n = K_n^{-1}\left(X^Tz\right), b_n = b + \frac{1}{2}\left(z^Tz - mu_n^TK_n\mu_n\right)$
The MCMC samples from the following distribution:
$\omega \sim Ga\left(\omega; a_n, b_n\right)$
$\beta \sim N\left(\beta; \mu_n, \left(\omega K_n\right)^{-1}\right)$
The resulting correct rate is 0.65
End Solution

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.*

Begin Solution: To maximize the log likelihood function

$$\log\left(P\left(\beta\right)\right) = -\frac{1}{2\sigma^2} - \sum_i y_i \log\left(1 + \exp\left(-x_i^T\beta\right)\right) - \sum_i \left(1 - y_i\right)\log\left(1 + \exp\left(x_i^T\beta\right)\right) + \lambda\beta^2 - t$$

With initial guess of $\beta = 0.5$, the optimum is found at $\beta^* = -0.011$
End Solution

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).*

Begin Solution: The following plot shows the log-likelihood

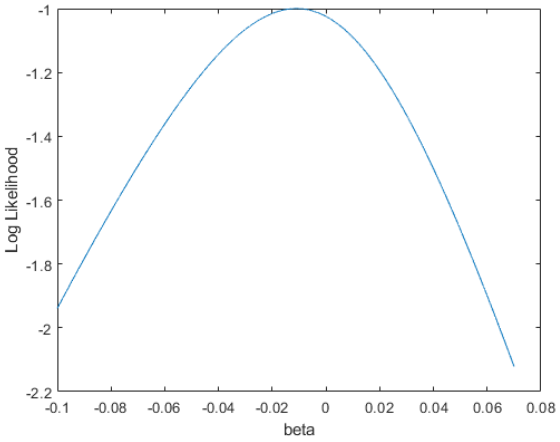

Figure 3.1: Exercise 3.3

End Solution

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.*

Begin Solution:

$$\frac{\partial^2 \log\left(p(\beta|x,y)\right)}{\partial \beta_j^2} \propto -\frac{1}{\sigma^2} - \sum_{i=1}^{n} \frac{x_{ij}^2 \exp\left(x_i^T \beta\right)}{\left(1 + \exp\left(x_i^T \beta\right)\right)^2}$$

For a Gaussian distribution $N(\mu, \omega)$,$\frac{\partial^2 \log(P)}{\partial x^2} = -\omega^2$,$\mu = \beta^*$
End Solution

[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.

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.

Begin Solution:

$$\begin{aligned}
p(\beta|x,y) &\propto p(y|x,\beta)p(\beta) \\
&\propto N(\beta;0,\sigma^2 I)\prod_{i=1}^{n} Bern(y_i;\frac{1}{1+\exp(-x_i^T\beta)}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\beta^T\beta\right)\prod_{i=1}^{n}\left\{\left[\frac{1}{1+\exp\left(-x_i^T\beta\right)}\right]^{y_i}\left[1-\frac{1}{1+\exp\left(-x_i^T\beta\right)}\right]^{1-y_i}\right\} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\beta^T\beta\right)\prod_{i=1}^{n}\left\{\exp\left(x_i^T\beta y_i\right)\cdot\frac{1}{\exp\left(x_i^T\beta\right)+1}\right\}
\end{aligned}$$

Take log transformation

$$\begin{aligned}
\log\left(p\left(\beta\right)\right) &\propto -\frac{1}{2\sigma^2}\beta^T\beta + \sum_{i=1}^{n}\log\left\{\exp\left(x_i^T\beta y_i\right)\cdot\frac{1}{\exp\left(x_i^T\beta\right)+1}\right\} \\
&= -\frac{1}{2\sigma^2}\beta^T\beta + \sum_{i=1}^{n}\left\{x_i^T\beta y_i - \log\left(\exp\left(x_i^T\beta\right)+1\right)\right\}
\end{aligned}$$

Compute partial derivatives

$$\frac{\partial\log\left(p(\beta|x,y)\right)}{\partial\beta_j} \propto \frac{1}{\sigma^2}\beta_j + \sum_{i=1}^{n}\left[x_{i_j}y_i - \frac{x_{i_j}}{1+\exp\left(-x_i^T\beta\right)}\right]$$

$$\frac{\partial^2\log\left(p(\beta|x,y)\right)}{\partial\beta_j^2} \propto -\frac{1}{\sigma^2} - \sum_{i=1}^{n}\frac{x_{i_j}^2\exp\left(x_i^T\beta\right)}{\left(1+\exp\left(x_i^T\beta\right)\right)^2}$$

$$\frac{\partial^2\log\left(p(\beta|x,y)\right)}{\partial\beta_j\beta_k} \propto -\sum_{i=1}^{n}\frac{x_{i_j}x_{i_k}\exp\left(x_i^T\beta\right)}{\left(1+\exp\left(x_i^T\beta\right)\right)^2}$$

The values to maximize the posterior distribution is $\beta_{MAP} = [-0.1989, -0.0083]$,plug into the Hessian matrix and the results are

$$H = \begin{bmatrix} -49.67 & -2749 \\ -2749 & -20200 \end{bmatrix}$$

The posterior distribution takes the form

$$P(\beta) = \left(\frac{-H}{2\pi}\right)\exp\left\{-\frac{1}{2}\left(\beta-\beta_{MAP}\right)^T\left(-H\right)\left(\beta-\beta_{MAP}\right)\right\}$$

End Solution

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.*

Begin Solution: 1. The data is heavily skewed to the left, as can be seen from the histogram. If the censored values are available, the left-skewedness will be even more obvious.
2. For a Poisson distribution, the mean, mode and variance are equal. However in this dataset, the mean is 15.92, variance is 460.91, and mode is 5.
End Solution

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.*

Begin Solution:

$$\begin{aligned} \lambda_i &= \exp(X\beta) \\ p(\beta|X,Y) &\propto p(\beta)\prod_{i=1}^{N} \frac{\lambda_i^{y_i}\exp(-\lambda)}{y_i!} \end{aligned}$$

Assign a Gaussian prior to β: $\beta \sim N(0, \sigma^2 I)$

$$\begin{aligned} \log P(\beta|X,Y) &\propto -\frac{1}{2\sigma^2}\beta^T\beta + \sum_{i=1}^{N}[y_i x_i \beta - \exp(x_i\beta)] \\ \frac{\partial^2 \log P(\beta|X,Y)}{\partial \beta_j^2} &\propto -\frac{1}{\sigma^2} - \sum_{i=1}^{N}\exp(x_i\beta)x_{i_j}^2 \\ \frac{\partial^2 \log P(\beta|X,Y)}{\partial \beta_j \partial \beta_k} &\propto -\sum_{i=1}^{N}\exp(x_i\beta)x_{i_j}x_{i_k} \end{aligned}$$

The value that maximize the posterior distribution of β is $\hat{beta} = [0.3894, 0.0503]$. The covariance matrix is estimated by $(-H)^{-1}$.

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

The 95% marginal credible interval for the entries are estimated to be

$$\beta_{int} : [2.3794, 2.3995], \beta_{grade} : [0.0490, 0.0515]$$

End Solution

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_rstan.htm` 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.*

Begin Solution: With 3 chains, 3000 iterations and 1000 burn-in iterations, the trace plots are shown in Figure 3.2.
The 95% confidence interval for the β are

$$\hat{\beta}_{int} = [2.375362, 2.397973], \hat{\beta}_{grade} = [0.04920777, 0.05165387]$$

The RMSE of this model is 21.39579

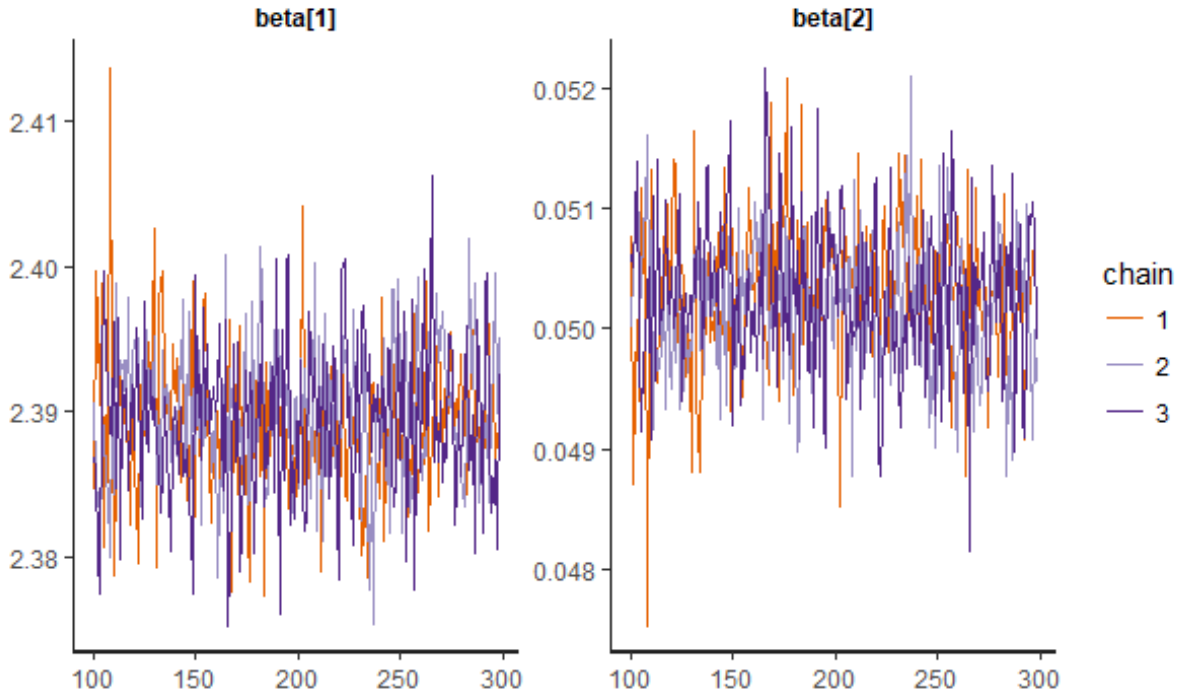

Figure 3.2: Exercise 3.9

After adding SEATTEND and SEXX as features,

$$\hat{\beta}_{int} = [2.350340, 2.378363], \hat{\beta}_{grade} = [0.05149095, 0.05434953]$$

$$\hat{\beta}_{SEATTEND} = [0.1533364, 0.1704776], \hat{\beta}_{gender} = [-0.5997926, -0.5742068]$$

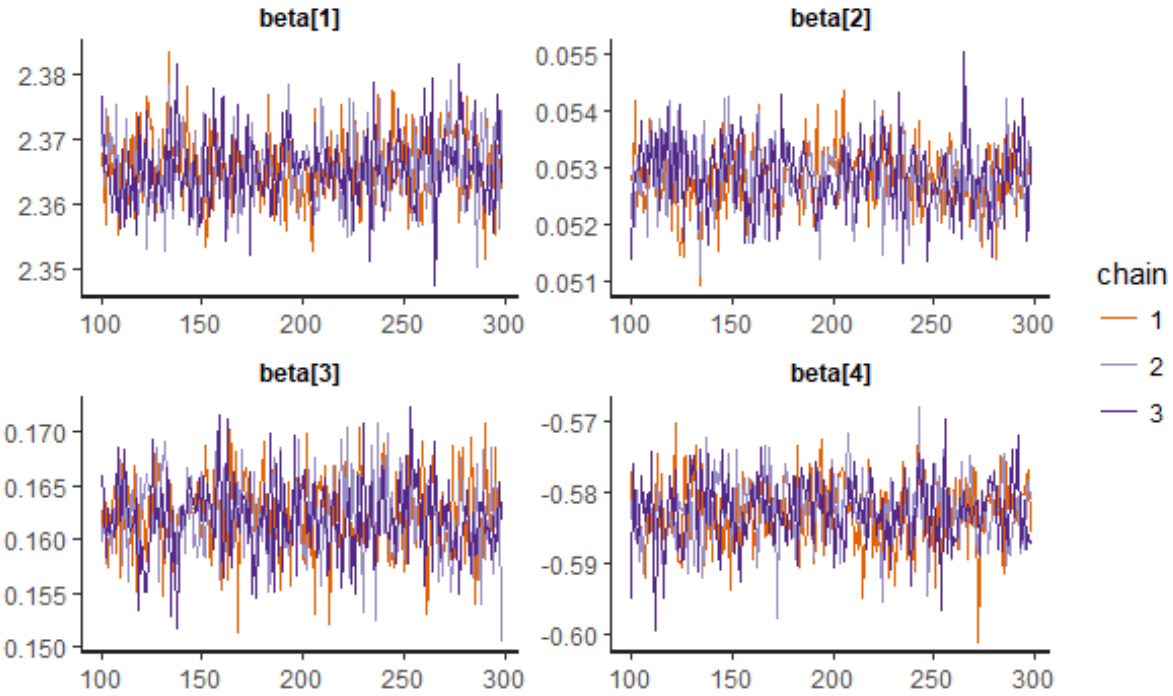

Figure 3.3: Exercise 3.9 - Added variables

End Solution

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.*

Begin Solution: Consider the interaction terms between gender and grade, as well as gender and attendance. The trace plot results are as follows.

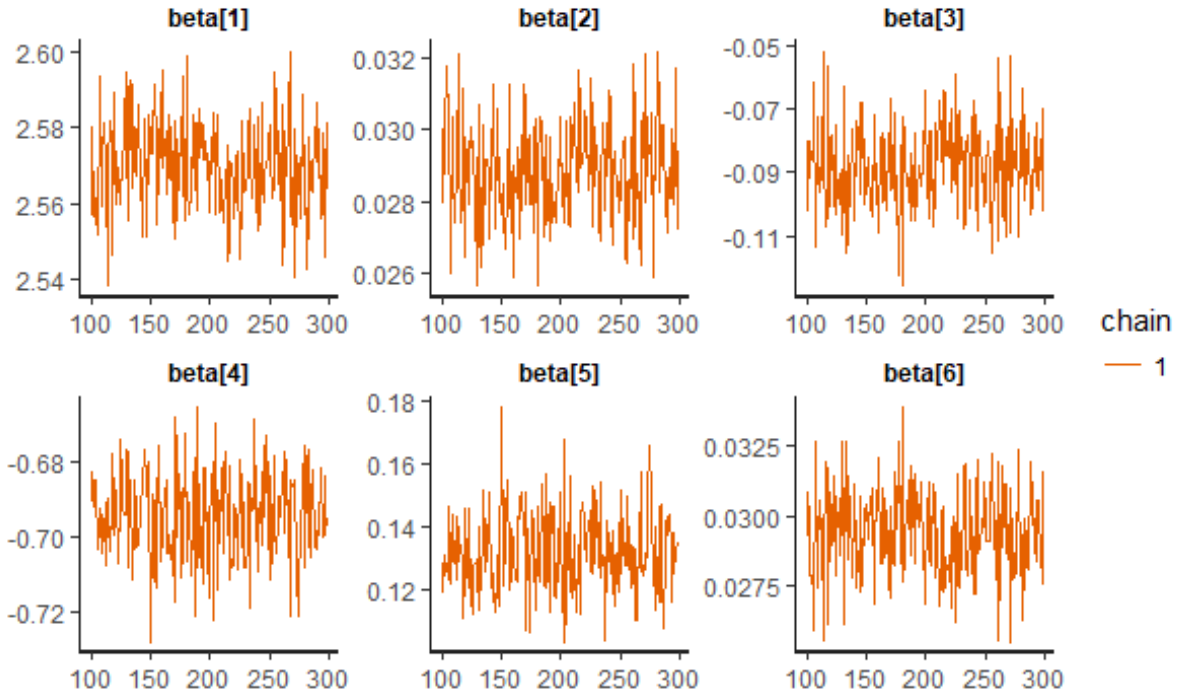

Figure 3.4: Exercise 3.10

The RMSE is now 21.11479, which is only slightly lower than the original model with RMSE 21.39579.
End Solution

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).*

Begin Solution: I will predict the censored values using the available features. One way to perform the prediction is to calculate the l2 norms of the feature vector that does not include the censored variable. Then evenly split up the censored data into 4 portions according to the calculated l2 norms. With the portion with lowest potion corresponds to 1 student disciplined and the highest portion with 4 students disciplined.
End Solution