Sareh Kouchaki

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

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

*** useful link for generalized linear model: http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

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`.

Solution:

Useful Link: https://rpubs.com/cakapourani/bayesian-binary-probit-model

$$\mathrm{p}(\beta|\mathrm{y},\mathrm{x}) \propto \mathrm{p}(\beta)\,\mathrm{p}(\mathrm{y}|\beta,\mathrm{x}) = \pi(\beta)\prod_{i=1}^{n} p(y_i|\beta,x_i) = \pi(\beta)\prod_{i=1}^{n}\Phi(x_i\beta)^{y_i}(1-\Phi(x_i\beta))^{1-y_i}$$

Performing inference for this model in the Bayesian framework is complicated by the fact that no conjugate prior π(β) exists. To overcome this problem, Albert and Chib (1993) augmented the original model with an additional auxiliary variable that renders the conditional distributions of the model parameters equivalent to those under a Bayesian normal linear regression model with Gaussian noise; and derived an efficient Gibbs sampling scheme for computing the posterior statistics.

Augmented Model:

$$y_i = \begin{cases} 1 & if\ z_i > 0 \\ 0 & if\ z_i \le 0 \end{cases}$$

$$z_i = x_i\beta + \epsilon_i$$

$$z_i \sim N(x_i^T\beta\,, \tau^2)$$

We are interested in computing the joint posterior distribution of the latent variables z and the model parameter β given the data y and x.

$$\mathrm{p}(\mathrm{z},\beta|y,x) \propto p(\beta)\,p(z|\beta,x)\,p(y|z) = \pi(\beta)\prod_{i=1}^{n} p(z_i|\beta,x_i)\,p(y_i|z_i)$$

where we have,

$p(z_i|\beta,x_i) = N(x_i^T\beta\,,\tau^2)$

$\mathrm{p}(y_i|z_i) = I(y_i = 1)I(z_i > 0) + I(y_i = 0)I(z_i \le 0)$

where I is the indicator function, equal to 1 if the quantities inside the function are satisfied, and 0 otherwise.

The joint posterior is difficult to normalize and sample from directly. However, computation of the marginal posterior of β and z using the Gibbs sampling requires only computing $\mathrm{p}(\beta|\mathrm{z},\mathrm{y},\mathrm{x})$ and $\mathrm{p}(\mathrm{z}|\beta,\mathrm{y},\mathrm{x})$.

$$\mathrm{p}(\beta|\mathrm{z},\mathrm{y},\mathrm{x}) = \mathrm{p}(\beta|\mathrm{z},\mathrm{x}) \propto \pi(\beta)\prod_{i=1}^{n} N(x_i\beta\,,\tau^2)$$

This is the posterior density for the normal linear regression. The posterior distribution would be normal.

If we assign a constant prior on β, $\mathrm{p}(\beta) \propto 1$, then the conditional posterior would be as the following:

$$\beta|\mathrm{z},\mathrm{x} \sim \mathrm{N}((x^Tx)^{-1}x^Tz, (x^T\mathrm{x})^{-1})$$

$$\mathrm{p}(\mathrm{z}|\beta,\mathrm{y},\mathrm{x}) \propto \mathrm{p}(\mathrm{z}|\beta,\mathrm{x})\,\mathrm{p}(\mathrm{y}|\mathrm{z}) = \begin{cases} N(x\beta,\tau^2)\,I(z_i > 0) & if\ y_i = 1 \\ N(x\beta,\tau^2)\,I(z_i \le 0) & if\ y_i = 0 \end{cases}$$

R code for Gibbs Sampling part is on the GitHub under the name of Exercise 3.1.R. Figure 3.1 shows the plots of coefficients.

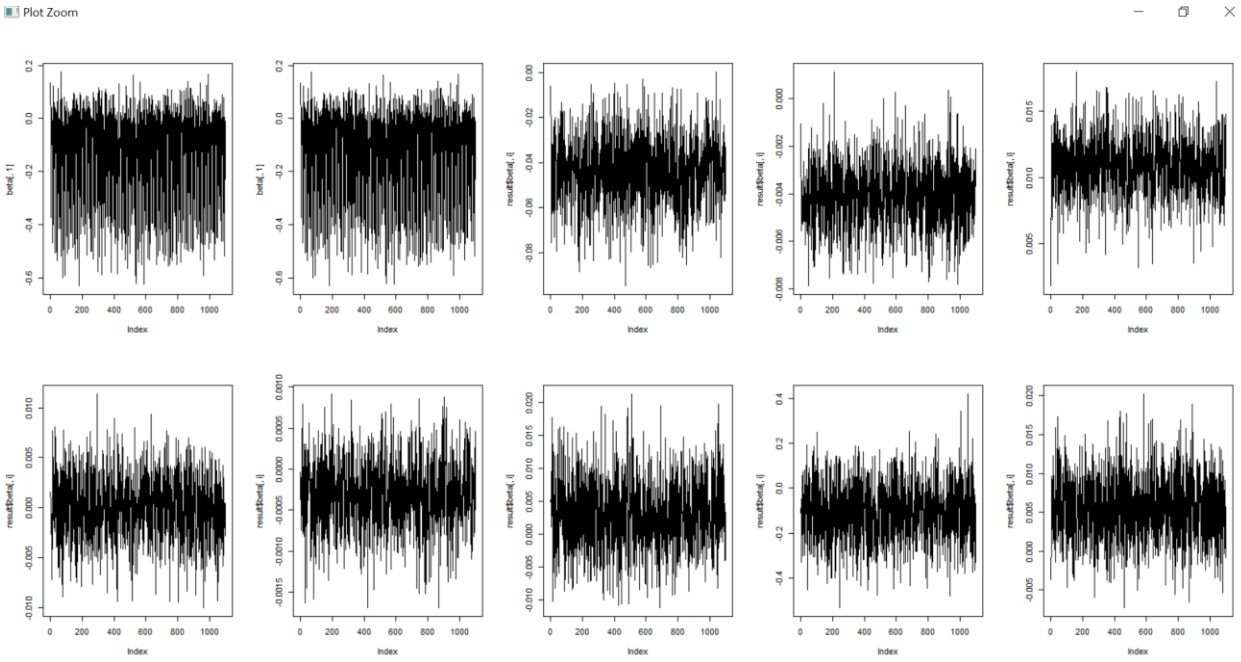


Figure 3.1

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 Polson et al. (2013). 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.*

Solution:

Wikipedia: In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

Now assume that a prior distribution g over θ exists. This allows us to treat θ as a random variable as in Bayesian statistics. We can calculate the posterior distribution of θ using Bayes' theorem:

$$\theta \mapsto f(\theta \mid x) = \frac{f(x \mid \theta)\, g(\theta)}{\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta}$$

where g is density function of θ, Θ is the domain of g.

The method of maximum a posteriori estimation then estimates θ as the mode of the posterior distribution of this random variable:

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} f(\theta \mid x) = \arg\max_{\theta} \frac{f(x \mid \theta)\, g(\theta)}{\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta} = \arg\max_{\theta} f(x \mid \theta)\, g(\theta).$$

Logit regression:

Useful link: https://www.cs.princeton.edu/~bee/courses/lec/lec_jan24.pdf

For a Bernoulli distribution, with y ϵ {0,1} and p representing the probability of success, $0 \leq p \leq 1$, we have:

$$p(y|p) = p^{y}(1-p)^{1-y}$$

Exponential family form of the Bernoulli distribution is as following:

$$\exp\{(\log(\frac{p}{1-p})y + log(1-p)\}$$

In the Bernoulli distribution, in the exponential family form, note that the logit function (i.e., log odds function) maps the mean parameter vector, p, to the natural parameter, η is shown below:

$$\eta = \log(\frac{p}{1-p}) \quad \rightarrow \quad p = \frac{1}{1+exp\{-\eta\}}$$

logistic regression model:

As in linear regression, we have pairs of observed variables $D = \{(x_1, y_1), ..., (x_n, y_n)\}$.

- Observed input x is assumed to enter the model via a linear combination, $x^T\beta$.
- The conditional mean p is represented as a function of $x^T\beta$.
- The response y is characterized by an exponential family distribution with conditional mean p.

For logistic regression, we set our natural parameter $\eta = x^T\beta$. Therefore, for our regression model where the conditional probability is modeled as a Bernoulli distribution, the parameter $p = E[Y|X, \beta]$ can be obtained from the logistic function,

$$p = \frac{1}{1+exp\{-\eta\}} = \frac{1}{1+exp\{-x^T\beta\}}$$

Likelihood function: $p(y_1, y_2, \ldots, y_n|x_1, x_2, \ldots, \beta) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$

If we change p to $\frac{1}{1+exp\{-x^T\beta\}}$, the likelihood function would be:

$$\prod_{i=1}^{n}\left(\frac{1}{1+exp\{-x^T\beta\}}\right)^{y_i}\left(1-\frac{1}{1+exp\{-x^T\beta\}}\right)^{1-y_i}$$

By considering Normal(0, I) as a prior on β, the posterior distribution would be:

$$p(\beta|y,x) \propto \exp\left(-\frac{\beta^2}{2}\right)\prod_{i=1}^{n}\left(\frac{1}{1+exp\{-x^T\beta\}}\right)^{y_i}\left(1-\frac{1}{1+exp\{-x^T\beta\}}\right)^{1-y_i}$$

$$\log\left(p(\beta|y,x)\right) \propto -\frac{\beta^2}{2} + \sum_i \left(y_i log\left(\frac{1}{1+exp\{-x^T\beta\}}\right)\right) + (1-y_i)\left[log\left(1-\frac{1}{1+exp\{-x^T\beta\}}\right)\right]$$

To find MAP, we need to maximize the above function. The code is available on GitHub under the name Exercise3.2.

$$\beta_{MAP} = -0.011$$

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).*

Solution: Code is available on the GitHub. Figure 3.2 shows the plot $p(\beta|x,y,\mu,\sigma^2)$.

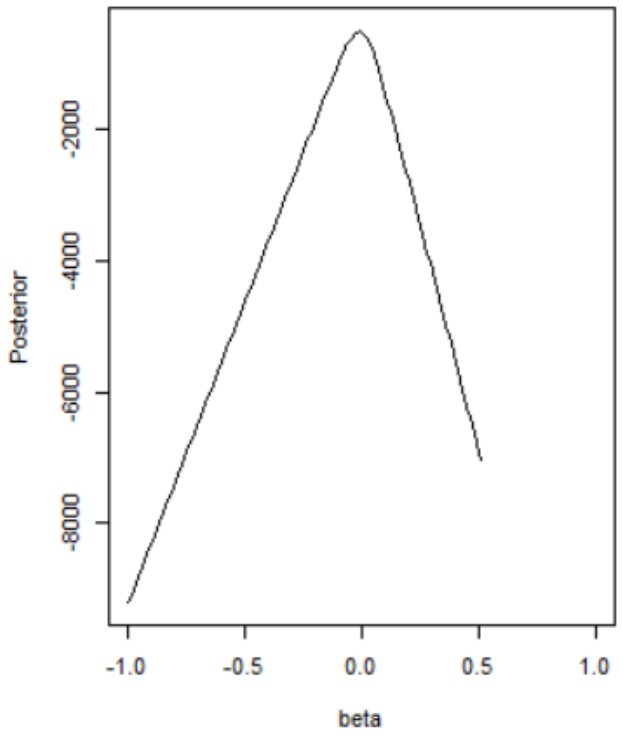


Figure 3.2

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:

$$\log(\mathrm{p}(\beta|\mathrm{y},\mathrm{x})) \propto -\frac{\beta^2}{2} + \sum_i \left(y_i log\left(\frac{1}{1+exp\{-x^T\beta\}}\right)\right) + (1-y_i)\left[log\left(1 - \frac{1}{1+exp\{-x^T\beta\}}\right)\right]$$

$$mean = \beta_{MAP} = -0.011$$

$$\mathrm{c} = -\frac{\partial^2}{\partial\beta^2} log(\mathrm{p}(\beta|\mathrm{y},\mathrm{x}))|_{\beta=\beta_{MAP}}$$

$$\frac{\partial}{\partial\beta} log(\mathrm{p}(\beta|\mathrm{y},\mathrm{x})) = -\beta + \sum_i \left(y_i \frac{\frac{\partial}{\partial\beta}\left(\frac{1}{1+exp\{-x^T\beta\}}\right)}{\frac{1}{1+exp\{-x^T\beta\}}} + (1-y_i)\frac{\frac{\partial}{\partial\beta}\left(1-\frac{1}{1+exp\{-x^T\beta\}}\right)}{1-\frac{1}{1+exp\{-x^T\beta\}}}\right)$$

$$= -\beta + \sum_i \left(y_i \frac{x^T exp(-x^T\beta)}{(exp(-x^T\beta)+1)} + (1-y_i)\frac{-x^T}{(exp(-x^T\beta)+1)}\right)$$

$$= -\beta + \sum_i \left(\frac{y_i x^T}{(exp(-x^T\beta)+1)}(\exp(-x^T\beta)+1) - \frac{x^T}{(exp(-x^T\beta)+1)}\right)$$

$$= -\beta + \sum_i \left(y_i x^T - \frac{x^T}{(exp(-x^T\beta)+1)}\right)$$

$$\frac{\partial^2}{\partial^2\beta} log(\mathrm{p}(\beta|\mathrm{y},\mathrm{x})) = \frac{\partial}{\partial\beta}\left(-\beta + \sum_i \left(y_i x^T - \frac{x^T}{(exp(-x^T\beta)+1)}\right)\right)$$

$$= -1 + \sum_i \left(\frac{-{x_i}^2 exp(-{x_i}^T\beta)}{(exp(-{x_i}^T\beta)+1)^2}\right)|_{\beta=\beta_{MAP}=-0.011}$$

$$\mathrm{c} = -\frac{\partial^2}{\partial\beta^2} log(\mathrm{p}(\beta|\mathrm{y},\mathrm{x}))|_{\beta=\beta_{MAP}} = 1 + \sum_i \left(\frac{{x_i}^2 exp(-{x_i}^T\beta)}{(exp(-{x_i}^T\beta)+1)^2}\right)|_{\beta=\beta_{MAP}=-0.011}$$

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.

Solution:

Multivariate distribution: β is a vector.

Let's consider:

$$\sigma(\beta^T x_i) = \sigma(z_i) = \frac{1}{1 + \exp(-z_i)}$$

$$L(\beta) = -\beta^T\beta + \sum_i y_i \log\sigma(z_i) + (1-y_i)\log(1-\sigma(z_i))$$

Note that:

$$\frac{\partial}{\partial\beta}\sigma(z) = \frac{\partial}{\partial\beta}(1+e^{-z})^{-1} = \sigma(z)(1-\sigma(z))$$

$$\frac{\partial \log\sigma(z_i)}{\partial\beta^T} = \frac{1}{\sigma(z_i)}\frac{\partial\sigma(z_i)}{\partial\beta^T} = \frac{1}{\sigma(z_i)}\frac{\partial\sigma(z_i)}{\partial z_i}\frac{\partial z_i}{\partial\beta^T} = (1-\sigma(z_i))x_i$$

$$\frac{\partial \log(1-\sigma(z_i))}{\partial\beta^T} = \frac{1}{1-\sigma(z_i)}\frac{\partial(1-\sigma(z_i))}{\partial z_i} = -\sigma(z_i)x_i$$

$$\frac{\partial}{\partial\beta^T}L(\beta) = -\lambda\beta + \sum_i y_i(1-\sigma(z_i))x_i - (1-y_i)\sigma(z_i)x_i = -\beta - \sum_i (y_i - \sigma(z_i))\, x_i$$

$$\frac{\partial^2}{\partial\beta^T\partial\beta}L(\beta) = -I - \sum_i x_i x_i^T \sigma(z_i)(1-\sigma(z_i)) = -I - \sum_i \frac{\exp(-x_i^T\beta)}{(1+\exp(-x_i^T\beta))^2}\, x_i x_i^T$$

$$H(\beta) = -I - \sum_i \frac{\exp(-x_i^T\beta)}{(1+\exp(-x_i^T\beta))^2}\, x_i x_i^T = -I - X\sigma(X^T\beta)(1-\sigma(X^T\beta))X^T$$

The approximation of the posterior distribution over β is:

$$\beta \sim N(\beta_{MAP}, \text{precision} = \text{inverse of negative of Hessian Matrix})$$

Code is provided on GitHub.

$$\beta_{MAP} = [-0.07905, -0.00886]$$

$$\text{cov} = [0.0293402164, -0.0007990073; -0.0007990073, 0.0000267148]$$

Credible Interval = [(-0.0415, 0.25668), (-0.019, 0.00127)]

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

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

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.*

Code:

```
data = read.csv("C:/Users/Sareh/Documents/GitHub/sds383d/data/tea_discipline_oss.csv")

newdata <- subset(data, ACTIONS !=-99)

mean(newdata$ACTIONS)

var(newdata$ACTIONS)
```

Solution:

One reason is that the mean and variance of Poisson distribution are equal, but here variance is greater than the mean.

mean(newdata$ACTIONS) = 15. 9256

var(newdata$ACTIONS) = 460.9088

Second, the data is skewed to the right.

Third, this could be multimodal distribution.

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.*

$$p(y_1, y_2, \ldots, y_n|\lambda) = \prod_i \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}$$

$$\text{log.likelihood} = \sum_i -\lambda + y_i \log(\lambda) - \log(y_i!) \text{ , where } \lambda = \exp(x_i^T \beta)$$

$$\text{log.likelihood} = \sum_i -\exp(x_i^T \beta) + y_i x_i^T \beta - \log(y_i!)$$

$$p(\beta) = N(\text{mean}, \sigma^2)$$

$$\log(p(\beta|Y,X)) \sim = (\sum_i - \exp(x_i^T \beta) + y_i x_i^T \beta) - \lambda \beta\beta^T$$

$\frac{\partial}{\partial \beta}$: $\sum_i(-X^T \exp(X^T\beta) + YX^T) - \lambda\beta^T$

$\frac{\partial^2}{\partial\beta\partial\beta^T}$: $\sum_i(-X^T\exp(X^T\beta)\,X) - \lambda I$

We want to find β_{MAP} using the optim function in R. Code is available on GitHub.

$\beta_{MAP} = [\widehat{\beta_1}, \widehat{\beta_2}] = [2.22, 0.055]$

cov of the posterior distribution is the negative of Hessian which is as following:

cov = $\begin{bmatrix} 2.846114e-05 & -2.465906e-06 \\ -2.465906e-06 & 2.388357e-07 \end{bmatrix}$

95% Credible Interval for β_1 and β_2 : [(2.2095, 2.23); (0.0541, 0.056)]

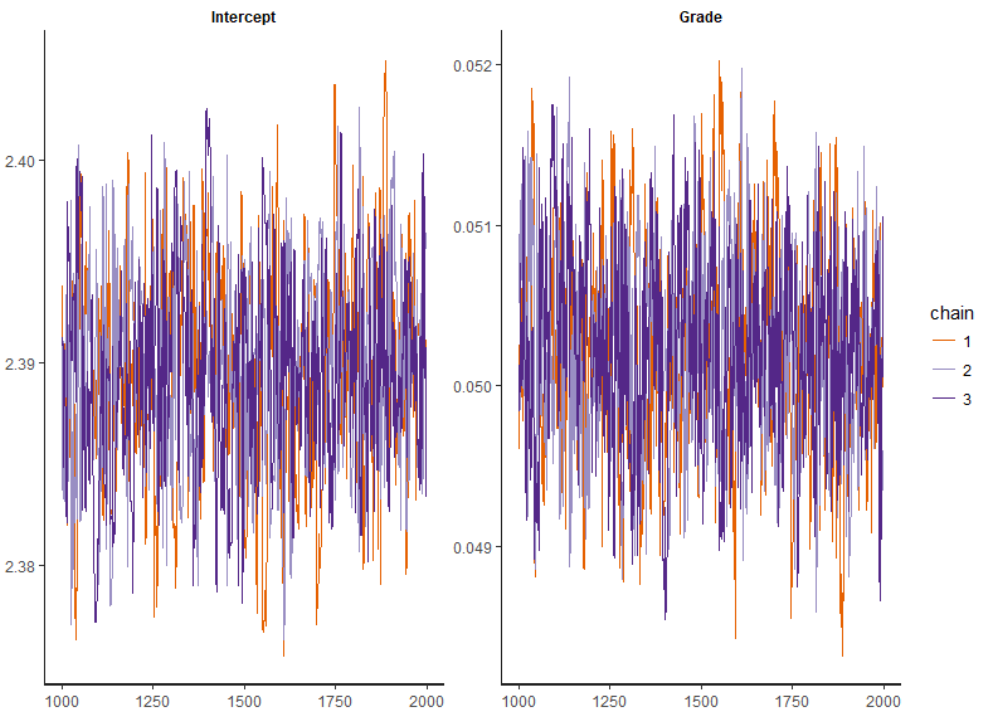
Exercise 3.8 (Optional) *Repeat the analysis using a set of variables that interest you.*

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_r` 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.*

Solution: some modifications were made to poisson.stan script and also the R code. Both files are available on GitHub.

According to the result obtained from Stan, $[\widehat{\beta_1}, \widehat{\beta_2}] = [2.30, 0.05]$. The trace plots of betas are provided in the following:



So far, we used Grade as a covariate in the model. I used other covariates like sex and ethnic. The code and stan script are available on GitHub.

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.*

Solutions:

In this question, I incorporated the interaction term, SEXX*GRADE, in the model to see how this modification affect the model. Code is available on GitHub.

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).*

Solution:

We could add the truncated part by including them using a probability distribution. For example:

$$y \sim Poisson\left(e^{-x^T\beta}\right)$$

$$y_{censored} \sim P_{\leq 4}(e^{-x^T\beta})$$