Bayesian Theory and Computation

Lecture 6: Linear and Generalized Linear Models

Cheng Zhang

School of Mathematical Sciences, Peking University

April 02, 2021

- Consider the following linear regression model:

$$y|X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

- y is a column vector of n observations for the outcome variable, X is an $n \times (p+1)$ matrix of observed predictors with its first column being all 1's.
- β is a column vector with $p+1$ elements $(\beta_0, \ldots, \beta_p)$ where β_0 is the intercept and β_j represents the effect of the j-th predictor x_j on y.

- To perform Bayesian analysis, we need to obtain the posterior distribution of parameters based on the model and the prior.
- A common prior for parameters are

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

$$\beta \sim \mathcal{N}_{p+1}(\mu_0, \Lambda_0)$$

where

$$\mu_0 = (\mu_{00}, \mu_{01}, \ldots, \mu_{0p}),\ \Lambda_0 = \text{diag}(\tau_0^2, \tau_1^2, \ldots, \tau_p^2)$$

- μ_0 is typically set to zero (unless we believe otherwise), Λ_0 should be sufficiently broad.

Posterior Distributions

- The posterior distribution of β has the following closed form:

$$\beta | X, y, \sigma^2 \sim \mathcal{N}(\mu_n, \Lambda_n)$$

where

$$\mu_n = \left(X_*' \Sigma_*^{-1} X_*\right)^{-1} X_*' \Sigma_*^{-1} y_*, \quad \Lambda_n = \left(X_*' \Sigma_*^{-1} X_*\right)^{-1}$$

$$X_* = \begin{bmatrix} x \\ I_{p+1} \end{bmatrix}, \quad y_* = \begin{bmatrix} y \\ \mu_0 \end{bmatrix}, \quad \Sigma_* = \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & \Lambda_0 \end{bmatrix}$$

- Looking at it this way, the prior plays the role of extra data with $x_\beta = I_{p+1}$, $y_\beta = \mu_0$ and the covariance Λ_0.
- That's why Bayesian models do not break down when $p > n$.

PEKING UNIVERSITY

- Now, we want to obtain the posterior distribution of σ^2
- Given β, again we have a simple normal model with observations y_i with known mean $X\beta$, unknown variance σ^2, and conditionally conjugate prior Inv-$\chi^2(\nu_0, \sigma^2)$
- As we saw before, the posterior distribution of $\sigma^2|X, y, \beta$ is also scaled Inv-χ^2

$$\sigma^2|X, y, \beta \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}\right)$$

$$\nu = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

- If we do not have an informative prior, we can instead use the following prior

$$p(\beta, \sigma^2 | X) \propto \sigma^{-2}$$

- For β this is equivalent to taking all $\tau_j^2 \to \infty$.
- The posterior distribution therefore becomes

$$\begin{aligned} \beta | y, \sigma^2 &\sim \mathcal{N}(\hat{\beta}, V_\beta \sigma^2) \\ \hat{\beta} &= (X'X)^{-1} X'y \\ V_\beta &= (X'X)^{-1} \end{aligned}$$

- The posterior distribution of σ^2 also has a closed form

$$\sigma^2|X, y, \hat{\beta} \sim \text{Inv-}\chi^2(n-p-1, s^2)$$
$$s^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2$$

- These close-form conditional posterior distributions allow efficient Gibbs sampling for Bayesian linear regression models.

Example: Children's Test Score

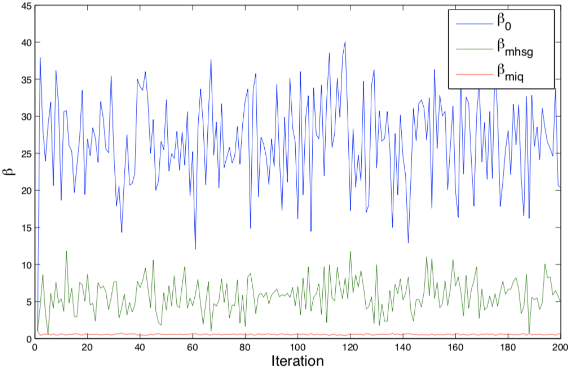
- Consider the children's test score example discussed by Gelman and Hill (2007).
- In this example, we are interested in the effect of mother's education (mhsg) and her IQ (miq) on the cognitive test score of 3 to 4 year old children.
- For our Bayesian model, we use the following broad priors

$$\sigma^2 \sim \text{Inv-}\chi^2(1, 0.5)$$

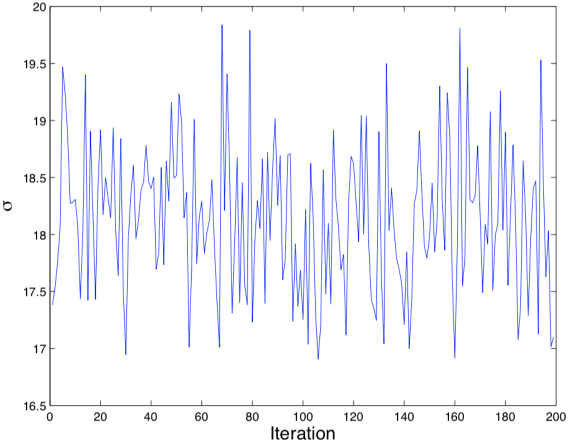
$$\beta \sim \mathcal{N}_{p+1}(0, 100^2 I)$$

- We use the Gibbs sampler to obtain 10000 samples and discarded the first 1000.

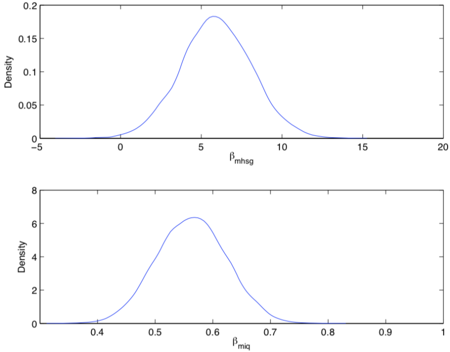
- The following plot shows the trace plot of posterior samples for β's



- The following plot is the trace plot of posterior samples for σ



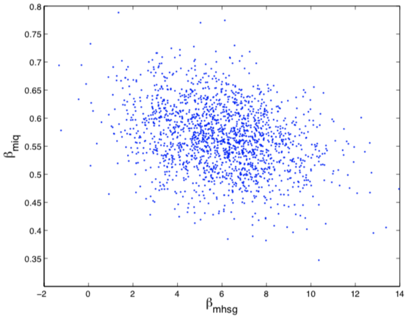
- Using the MCMC samples, we can also plot the posterior distribution of β's



- These are of course marginal distributions. We can plot the joint distribution of $(\beta_{\text{mhsg}}, \beta_{\text{miq}})$

Example: Children's Test Score

- The following plot shows the scatter plot of posterior samples for β_{mhsg} and β_{miq}



- Note that in general, β's are not independent in posterior although we might assume them independent in prior.

- We can also summarize the result of our analysis (i.e., posterior mean and 95% credible intervals) as follows

Parameter	Posterior expectation	95% Probability Interval
β_0	25.7939	[14.4, 37.2]
β_{mhsg}	5.9278	[1.6, 10.3]
β_{miq}	0.5633	[0.4, 0.7]
σ	18.2	[16.9, 19.4]

Example: Human, Age and Body Fat

- For the second example, we are interested in modeling body fat in terms of age and gender

$$\mathbb{E}(\text{bodyFat}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{gender}$$

- The above model, however, assumes that the effect of age on body fat is the same for Male (gender $= 0$) and Female (gender $= 1$)
- If we don't believe in that, we can include an interaction term ageGender $=$ age $\times$ gender into our model

$$\mathbb{E}(\text{bodyFat}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{gender} + \beta_{12} \text{ageGender}$$

Example: Human, Age and Body Fat

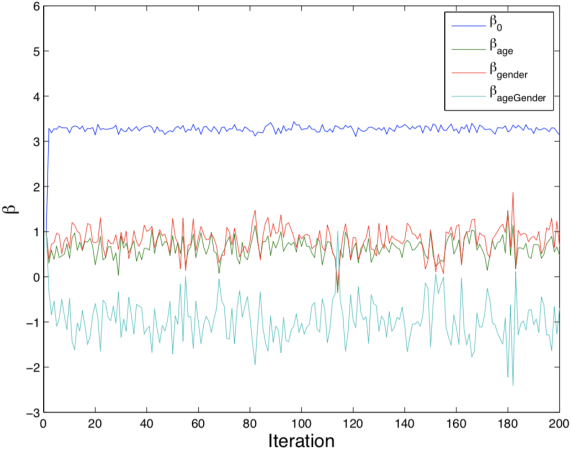
- Before analyzing the data, we first center and standardize predictors so they have mean zero and standard deviation 1.
- This type of transformation (centering predictors and maybe the outcome variable too) is usually (not always) appropriate and makes setting up the priors easier.
- Moreover, we use the log(fat) as the outcome.
- We use the following priors for model parameters:

$$\sigma^2 \sim \text{Inv-}\chi^2(1, 0.5)$$

$$\beta_j \sim \mathcal{N}(0, 10^2)$$

Example: Human, Age and Body Fat

- The following plot shows the trace plot of posterior samples for β's



- As before, we can also summarize the result of our analysis in the following table

Parameter	Posterior expectation	95% CR
β_0	3.28	[3.14, 3.40]
β_1	0.63	[0.19, 1.07]
β_2	0.82	[0.25, 1.37]
β_{12}	- 0.94	[-1.80, -0.08]
σ	0.28	[0.19, 0.41]

- Once we develop a model and perform Bayesian inference to obtain posterior estimation, we need to evaluate the adequacy of our model and assumption.
- This is done mainly based on how well it agrees with the data we have already observed, or we observe in future.
- Note that this is not the question of whether the model is true of false ("all models are wrong, but some are useful" – George Box), rather, how much our inference is affected by our simplifications.
- One good approach for evaluating models is using future observations assuming they are generated based on the same process as the observed data.
- Since this is not always possible, sometimes we hold out a part of the data, and treat them as future observations.

- An alternative approach for model checking is to replicate data (denoted as y^{rep}) using the posterior distribution and make sure there is no substantial and systematic difference between the replicated data and observed data.
- To replicate data, we can sample from the posterior distribution, and use each sample to generate a set of data. For example, if we are assuming a normal model $y \sim \mathcal{N}(\mu, \sigma^2)$. We first obtain the joint posterior distribution of (μ, σ^2), generate L samples from this distribution, and for each $\ell = 1, \dots, L$, generate $y^{\text{rep}} \sim \mathcal{N}(\mu^\ell, (\sigma^2)^\ell)$.
- If we have a hierarchical model, we have to first start with hyperparameters, given their sampled values, we sample from the parameters of the model, replicate new data as before.

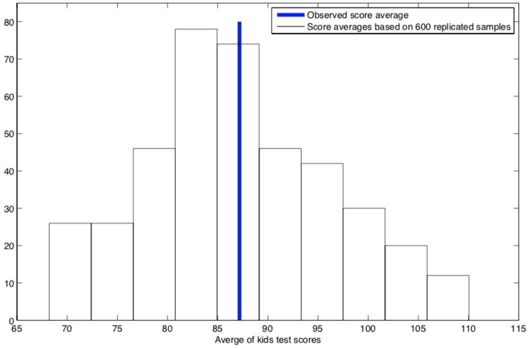
- For linear regression models, we generate samples $(\beta^\ell, (\sigma^2)^\ell)$ from the posterior distribution of (β, σ^2), and then generate n samples $y^{\text{rep}} \sim \mathcal{N}(X\beta^\ell, (\sigma^2)^\ell)$.
- Note that y^{rep} is different from $\tilde{y}$ (i.e., future observations) since it has the same X as the observed data.
- In practice, we already have samples from the posterior distribution when we use MCMC simulation. Therefore, we can directly use these samples to replicate data.
- As mentioned above, we perform model checking by comparing the observed data y and replicated datasets y^{rep}.
- We can do this comparison based on some appropriate *test quantity*, $T(y, \theta)$, where $\theta = (\beta, \sigma^2)$ in regression models.
- Note that in the Bayesian framework, test quantities could be a function of both data and unknown parameters θ.

- Typical test quantities are mean, median, variance, min, and max. We can use multiple of these tests to evaluate different aspect of the model.
- We can calculate the tail probability

$$p_B = p(T(y^{\text{rep}}, \theta) \geq T(y, \theta))$$

  which is the probability that the replicated data could be more extreme than the observed data, and use it as a measure of the discrepancy between the observed data and what we would expect according to the model.
- We can obtain this by simply estimating the proportion of replicated samples for which $T(y^{\text{rep}_\ell}, \theta^\ell) \geq T(y, \theta^\ell)$, where $\ell = 1, \ldots, L$.
- The model is suspected if p_B is close to 0 or 1.

- The following plot shows the observed average of y in the children's test score example compared to the averages obtained from the replicated samples. The estimated p_B is 0.53.



Prediction

- A main objective of regression analysis is to predict future observations for which we would known the value of their predictors $\tilde{x}$, and we are interested in predicting their unknown outcome $\tilde{y}$.
- In order to predict $\tilde{y}$ when we know $\tilde{x}$, we use the posterior predictive probability $p(\tilde{y}|y)$.
- To sample from $p(\tilde{y}|y)$, we could use its closed form (which is a multivariate t distribution). However, we could simply sample (β, σ^2) from their joint posterior distribution, and then sample $\tilde{y} \sim \mathcal{N}(\tilde{x}\beta, \sigma^2)$.
- Since we used MCMC simulation, we already have samples from the posterior distribution, which we can use directly (after discarding the burn-in samples) to generate $\tilde{y}$.
- Finally, we can use the posterior predictive mean of $\tilde{y}|y$ to predict the outcome for future observation.

- To get the posterior predictive mean, instead of sampling $\tilde{y}$'s and averaging them, we can simply do as follows

$$\mathbb{E}(\tilde{y}|y) = \frac{1}{L}\sum_{\ell=1}^{L} \tilde{x}\beta^{\ell}$$

  where L is the number of posterior samples β^{ℓ} after convergence.
- Although for the above model, we could use $\tilde{x}\hat{\beta}$ (where $\hat{\beta}$ is the posterior mean of β). This is not the case in general. Always find the value of the function (in this case $\tilde{x}\beta$) over the posterior samples and then average.

- We now show how we can simulate data, build a linear regression model and predict future observations (in this case, simulated after building the model, or before but not used in the model).
- For simulations, because this is an imaginary situation, we can choose any arbitrary prior. Let's assume the following priors:

$$\beta_j \sim \mathcal{N}(0, 2^2), \quad j = 1, \ldots, p$$
$$\sigma^2 \sim \text{Inv-}\chi^2(5, 1)$$

- Let's set the number of predictors to 3 and sample one set of $\beta^* = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\sigma^2)^*$ from the above priors. These should be regarded as the true values of the parameters.

- Next, we create n samples of predictors x. Since in our model we assume x are independent, we sample them independently from some distribution. Here, we set $n = 150$ and generate $x_{ij} \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$ and $j = 1, 2, 3$. Note that for $j = 0$, we use a column of 1's.
- Now, we can simulate y_i using the assumed linear model with sampled predictors and true parameters

$$y_i | x_i, \beta, \sigma^2 \sim \mathcal{N}(x\beta^*, (\sigma^2)^*)$$

- We regard the first 100 samples as observed data to build our model (training set), and use the remaining 50 data (test set) as future observation pretending we do not know their outcome.
- Our objective is to predict outcome for the test set and compare our answers to their true values.

- We build a linear regression model as before, using the prior we assumed and data we simulated.
- Using MCMC simulation, we obtain posterior samples for β and σ^2.
- We use these samples to obtain the posterior predictive distribution and posterior predictive expectation (i.e., our prediction) of y for the test set. These would be regarded as our prediction.
- We can then use some common summary measure (e.g., MSE) to evaluate our model.

Generalized Linear Model

- In general, our data might not conform with the assumptions of linear models.
- For such situation, we need a more flexible family of models.
- The class of generalized linear models (GLM), that includes linear models as a special case, provides such flexibility while it is still easy to use.
- Generalized linear models have three components:
  - A random component
  - A systematic component
  - A link function

Random Component

- The random component identifies the response variable and its probability distribution.
- In most situations, we assume some sort of exchangeablility for the set of observed outcome values $y_1, \ldots, y_n$, and regard them as iid given a parametric model $p(y|\theta)$ from the exponential family, such as normal, binomial, multinomial and Poisson.
- In general, if the outcome variable is continuous and real-valued, we use the normal distribution.
- If the outcome is binary, we use the Bernoulli/binomial distribution. For outcome variables with multiple categories, we use the multinomial instead. If the outcome variable represent counts data, we use the poisson distribution.

- The systematic component specifies the set of predictors (i.e., explanatory variables) $x = (x_1, \ldots, x_p)$ used in a *linear predictor* function.
- As before, we also append a vector of ones at the beginning of x.
- In the matrix form, the linear predictor function $\eta = x\beta$, where $\beta = (\beta_0, \ldots, \beta_p)$.
- Alternatively, for each observation i, where $i = 1, \ldots, n$, the linear predictor function is $\eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$.
- Also, as before, some of predictors could be a transformation (e.g., x^2) of original predictors.

- The link function is a monotonic differentiable function that connects the random and systematic components.
- More specifically, if $\mu = \mathbb{E}(y|x)$, the link function g connects μ to η such that $g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$ for each observation i.
- For the ordinary linear model we discussed before, the link function is identity: $g(\mu_i) = \mu_i$. That is $\mu_i = \eta_i = x_i\beta$.

- As mentioned before, for binary outcome variables, we use the Binomial distribution

$$y_i|n_i, \mu_i \sim \text{Binomial}(n_i, \mu_i)$$

  Bernoulli distribution is a special case when $n_i = 1$.
- As usual, we define the systematic part of the model $\eta_i = x_i\beta$.
- A common link function for this model is the logit function defined as follows

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i\beta$$

  where μ_i is the probability of success (i.e., $y_i = 1$).
- Therefore,

$$\mu_i = \frac{1}{1+\exp(-x_i\beta)}$$

- The likelihood is therefore defined in terms of β as follows

$$p(y|\mu) \propto \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{n_i-y_i}$$

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

- Note that in this model the variance of $y|x$ depends on the mean and therefore will not be constant

$$\mathbb{V}\mathrm{ar}(y_i|x_i) = n_i\mu_i(1-\mu_i)$$

- This is a generalization of logistic regression when the outcome could have multiple values (i.e., could belong to one of K classes).

$$y_i|n_i, \mu_{i1}, \ldots, \mu_{iK} \sim \text{Multinomial}(n_i, \mu_{i1}, \ldots, \mu_{iK})$$

  where μ_{ik} is the probability of class k for observation i such that $\sum_{k=1}^{K} \mu_{ik} = 1$.
- y_i is also a vector K elements with $\sum_{k=1}^{K} y_{ik} = n_i$.
- The systematic part is now a vector $\eta_{ik} = x_i\beta$, where β is a matrix of size $(p+1) \times K$.

- Each column k ($k = 1, 2, \dots, K$) corresponds to a set of $p + 1$ parameters associated with class K.
- This representation is redundant and results in nonidentifiability, since one of the β_k's can be set to zero without changing the set of relationship expressible with the model.
- Usually, either the first of the last column would be set to zero.
- In Bayesian models, removing this redundancy would make it difficult to specify a prior that treats all classes symmetrically. Therefore, we do not remove redundancy. In this case, what matters is the difference between the parameters of different classes.

Multinomial Logistic Model

- For the multinomial logistic model, we use a generalization of the link function we used for the binary logistic regression

$$\mu_{ik} = \frac{\exp(x_i\beta_k)}{\sum_{k'=1}^{K}\exp(x_i\beta_{k'})}$$

- The likelihood in terms of β is as follows

$$p(y|\mu) \propto \prod_{i=1}^{n}\prod_{k=1}^{K}\mu_{ik}^{y_{ik}} = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{\exp(x_i\beta_k)}{\sum_{k'=1}^{K}\exp(x_i\beta_{k'})}\right)^{y_{ik}}$$

- Here β_k is the column vector of $p+1$ parameters corresponding to class k.

Prior

- So far, we discussed the likelihood function for some common GLMs.
- Within the Bayesian framework, we also need to specify priors on model parameters.
- A common prior for β is normal prior: $\mathcal{N}(\mu_{0j}, \tau_{0j}^2)$.
- Usually we set $\mu_0 = 0$ unless we have good reasons to believe otherwise.
- After we specify the priors, the posterior sampling for β's can be performed using Metropolis algorithm with Gaussian jumps, or more advanced method such as the slice sampler.

- Here, we discuss a logistic regression model with normal priors for β.
- Recall that for logistic model, log-likelihood is obtained as follows

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

$$\log p(y|\beta) = \sum_{i=1}^{n} (y_i x_i \beta - n_i \log(1+\exp(x_i\beta))) + \text{Const}$$

- If we use a $\mathcal{N}(0, \tau_0^2 I)$ prior for β_j, the log-prior probability given τ_0^2 is simply

$$\log p(\beta|\tau_0^2) = -\frac{\|\beta\|^2}{2\tau_0^2} + \text{Const}$$

- The log-posterior is therefore

$$\log p(\beta|y) = -\frac{\|\beta\|^2}{2\tau_0^2} + \sum_{i=1}^{n} (y_i x_i \beta - n_i \log(1 + \exp(x_i \beta))) + \text{Const}$$

- Sometimes we may want to sample one parameter at a time. In that case, we can treat other parameters as constant (i.e., we don't need to calculate them if they can be absorbed into the constant part).

- The objective of this study (Norton and Dunn, 1985; Agresti, 2002) is to investigate whether there is a relationship between snoring and heart disease.
- We have the following data based on 2484 subjects (the snoring level is reported by spouses).

Snoring level	Number of people with heart disease: y_i	Total number of people surveyed: n_i
0	24	1355
2	35	603
4	21	192
5	30	224

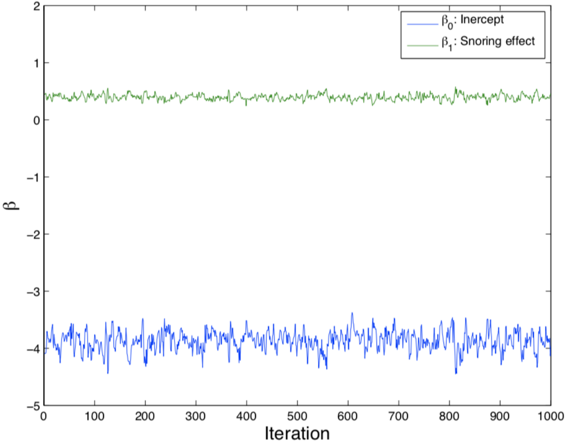
- Here, the snoring level (5 is the most severe) is the predictor or explanatory variable.
- The outcome variable is binary (i.e., heart disease = 1, no heart disease = 0).

- We assume y_i has a binomial distribution, and we model the relationship between snoring and heart disease using the logistic model.
- As before, we use a relatively broad prior for β

$$\beta_j \sim \mathcal{N}(0, 100^2) \quad j = 0, 1$$

- The role of prior here is mainly to provide a reasonable range for possible values of β (even if it is very broad). This helps us to avoid pitfalls associated with maximum likelihood estimates when the sample size is small or the data is sparse.
- Also, in general, we might want to use different priors for the intercept and coefficients.

- The following graph shows the trace plots of 1000 posterior samples after discarding the initial 500 samples



- We can use the posterior samples to obtain the posterior expectation of regression parameters as well as their 95% interval

	Posterior expectation	95% Interval
β_0	-3.87	[-4.24, -3.53]
β_1	0.4	[0.29, 0.51]

- As we can see, snoring is positively related to the increase in probability of heart disease.
- We can also talk about what is the posterior tail probability $p(\beta_1 < 0|y)$, and use it as a measure of our confidence when we make comments such as "snoring results in the increase risk of heart disease".
- Since this tail probability is zero (alternatively, we notice that 95% interval does not include 0), we believe the observed effect is statistically significant.

Priors for The Multinomial Logistic Model

- As before, we use normal priors for β's. But there is an issue we need to address.
- The above representation of multinomial logistic model is redundant since we only need $K-1$ parameters (say, $\mu_2, \ldots, \mu_K$). The first one would be determined based on these $K-1$ parameters since $\sum_{k=1}^{K} \mu_{ik} = 1$.
- Without this constraints, different set of parameter values can give the same probability. For example,

$$p(y_i = k|\eta + C) = \frac{\exp(\eta_{ik} + C)}{\sum_k' \exp(\eta_{ik'} + C)}$$
$$= \frac{\exp(\eta_{ik})}{\sum_k' \exp(\eta_{ik'})} = p(y_i = k|\eta)$$

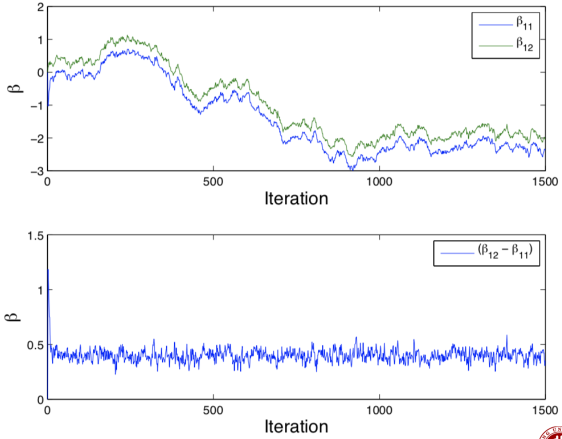
Priors for The Multinomial Logistic Model

- In the above example, while the values of η's changed the probabilities didn't. Note that for the multinomial logistic model, what really matters here is the difference between β's from one class to another.
- In statistics, when distinct parameter values give the same model, we say the model is unidentifiable.
- In classical statistics, this is bad, and to avoid this issue for multinomial logistic model, we could set one set of parameters (usually either β_1 or β_K) to zero.
- We do not do this in the Bayesian statistics since it would become difficult to set up symmetric priors based on β.
- Using the unidentifiable setting is totally fine with prediction. When inference is needed, we can use the posterior distribution of one of the β's as the baseline and subtract others from it to make it identifiable.

Example: Snoring and Heart Disease

- To show how we can set up an unidentifiable model and still perform inference, we use the snoring dataset for the first example (note that we can always use the multinomial logistic model regardless of whether the outcome is binary or multi-category).
- This time, β is a 2×2 matrix. The second row, β_{11}, β_{12} are the snoring effects on Class 1 (no heart disease) and Class 2 (heart disease).
- As before, we use a very wide $\mathcal{N}(0, 100^2)$ priors for β_{jk}, and use the slice sampler for simulating sample from the posterior distribution of β one parameter at a time.

- The first graph in the following figure shows the trace plots of β_{11} and β_{12}. The second graph shows the trace plot of $\beta_{12} - \beta_{11}$.



- While the absolute values of these parameters (and similarly the intercept parameters) do not converge to specific values due to non-identifiability, the identifiable parameters of the model, $\beta_{12} - \beta_{11}$, as shown in the second graph is converging with the posterior mean equal to 0.4 as we obtained using a logistic regression model.
- Therefore, we can continue our inference based on the identifiable parameters as we did before.

References

- A. Gelman and J. Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambriage University Press, 2007.
- Norton, P. G. and Dunn, E. V. (1985). Snoring as a risk factor for disease: an epidemiological survey. British medical journal (Clinical research ed.), 291(6496), 630–632. https://doi.org/10.1136/bmj.291.6496.630
- A. Agresti. Categorical Data Analysis. John Wiley & Sons, 2002