STATS 212: Generalized Linear Models
Lecture 7: Bayesian GLM

Babak Shahbaba

UCI, Spring 2010

Introduction

- We saw that for for modeling more complex data (such as grouped observations and variability in the response variable beyond what explained by simple models) we used models where parameters where regarded as random with their own distribution.
- Such models fit more naturally in the Bayesian framework.
- This lecture provides a brief preview of Bayesian models in the context of generalized linear models.
- A high level understanding of Bayesian methods is assumed.

Bayesian Linear regression models

- Consider the following *ordinary liner regression* model:

$$y|x, \beta, \sigma^2 \sim N(x\beta, \sigma^2 I_n)$$

- 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 $(\beta_0, \beta_1, ..., \beta_p)$ where β_0 is the intercept and β_j represents the effect of the j^{th} predictor x_j on y.
- σ^2 is the conditional variance of $y|x, \beta$.
- Therefore, $E(y_i|x, \beta) = \beta_0 + \beta_1 x_{i1} + \beta_p x_{ip}$ and $\mathrm{Var}(y_i|x, \beta) = \sigma^2$. That is y_i's are conditionally independent with the same variance σ^2.

Bayesian linear regression models

- To perform Bayesian analysis, we need to obtain the posterior distribution of parameters based on the model and the prior.
- We discussed the model above

$$y|x, \beta, \sigma^2 \sim N(x\beta, \sigma^2 I)$$

- A common prior for parameters are

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$
$$\beta|\mu_0, \Lambda_0 \sim N_{p+1}(\mu_0, \Lambda_0)$$

where $\mu_0 = (\mu_{00}, \mu_{01}, ..., \mu_{0p})$ and $\Lambda_0 = \mathrm{diag}(\tau_0^2, \tau_1^2, ..., \tau_p^2)$.

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

Posterior distributions

- We now need to find the posterior distributions of β and σ^2.
- We first consider situations where $p + 1 < n$ and the rank of x is $p + 1$ (these conditions are not required in Bayesian analysis when we use informative prior).
- First do the following transformation: multiply y by $(x'x)^{-1}x$.
- Recall that if $Y \sim N(\mu, \Sigma)$ then $AY \sim N(A\mu, A\Sigma A')$. Therefore,

$$\begin{aligned}
(x'x)^{-1}x'y &\sim N((x'x)^{-1}x'x\beta, (x'x)^{-1}x' \; \sigma^2 I \; x(x'x)^{-1}) \\
(x'x)^{-1}x'y &\sim N(\beta, (x'x)^{-1}\sigma^2)
\end{aligned}$$

Posterior distributions of β

- Now set $z = (x'x)^{-1}x'y$ and regard it as the observed data.
- For a given σ^2, this reduces to a simple multivariate normal model with unknown mean, β, and conjugate prior.

$$\begin{aligned} z &\sim N(\beta, \Sigma_z) \qquad \Sigma_z = (x'x)^{-1}\sigma^2 \\ \beta &\sim N(\mu_0, \Lambda_0) \end{aligned}$$

- The posterior distribution of $\beta|z, \sigma^2$ is

$$\begin{aligned} \beta|z, \sigma^2 &\sim N(\mu_n, \Lambda_n) \\ \mu_n &= (\Lambda_0^{-1} + \Sigma_z^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + \Sigma_z^{-1}z) \\ \Lambda_n^{-1} &= (\Lambda_0^{-1} + \Sigma_z^{-1}) \end{aligned}$$

- Using this approach, we can see that, similar to the normal model, the posterior expectation is a weighted average between prior and the maximum likelihood estimate (assuming it exists).

Posterior distributions of β

- Notice that we could obtain μ_n and Λ_n as follows:

$$
\begin{aligned}
\mu_n &= (x_*'\Sigma_*^{-1}x_*)^{-1}x_*'\Sigma_*^{-1}y_* \\
\Lambda_n &= (x_*'\Sigma_*^{-1}x_*)^{-1} \\
x_* &= \begin{pmatrix} x \\ I_{p+1} \end{pmatrix} \qquad y_* = \begin{pmatrix} y \\ \mu_0 \end{pmatrix} \qquad \Sigma_* = \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \Lambda_0 \end{pmatrix}
\end{aligned}
$$

- 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. Everything else remains as before.
- The above approach works even if $n < p+1$.

Posterior distributions of σ^2

- 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 $\text{Inv-}\chi^2(\nu_0, \sigma_0^2)$.
- As we saw before, the posterior distribution of $\sigma^2 | x, y, \beta$ is also scaled $\text{Inv-}\chi^2$

$$\begin{aligned} \sigma^2 | x, y, \beta &\sim \mathrm{Inv}\text{-}\chi^2(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}) \\ \nu &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 \end{aligned}$$

Noninformative prior distribution

- If we do not have an informative priors, we can instead use the following noninformative prior for regression models (assuming $n > p + 1$ and the rank of x is $p + 1$)

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

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

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

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

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

Logistic regression model

- Recall that the likelihood for logistic regression models in terms of β is as follows:

$$f(\beta) \quad \propto \quad \prod_{i=1} \left(\frac{\exp(\alpha + x_i\beta)}{1 + \exp(\alpha + x_i\beta)} \right)^{y_i} \left(\frac{1}{1 + \exp(\alpha + x_i\beta)} \right)^{n_i - y_i}$$

- Now similar to the ordinary linear regression model, we need to specify a prior distribution for α and β.
- Note that we separated the intercept here to emphesize that it should have it's own separate prior even if we decide to use a common prior for regression coefficients.

Logistic regression model

- It is common to set the prior for regression coefficients to $\beta_j \sim N(\mu_{0j}, \tau_{0j}^2)$. We mostly set $\mu_{0j} = 0$ unless we believe otherwise. For α, we can use $\alpha \sim N(0, \tau_\alpha^2)$
- Unlike the ordinary linear regression model, this would not be a conjugate prior, so we cannot use the Gibbs sampler directly.
- Alternatively, we might want to standardize the covariates to have mean zero and standard deviation 1, and use $\beta_j \sim N(0, \tau_\beta^2)$; i.e., one parameter τ_β for all covariates.

Poisson model

- For the Poisson model, the likelihood in terms of β is obtained as follows:

$$f(\beta) \quad \propto \quad \prod_{i}^{n} \exp[-\exp(\alpha + x_i\beta)][\exp(\alpha + x_i\beta)]^{y_i}$$

- We again use a nonconjugate normal $N(\mu_{0j}, \tau_{0j}^2)$ prior for β_j. As before, we set $\mu_0 = 0$ unless we believe otherwise.
- Again, we might want to standardize the covariates to have mean zero and standard deviation 1, and use $\beta_j \sim N(0, \tau_\beta^2)$ prior.
- The posterior sampling for β's can be performed using the Metropolis algorithm with Gaussian jumps, or more advanced method such as the slice sampler.

Multinomial logistic model

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

$$\mu_{ik} = \frac{\exp(\alpha_k + x_i\beta_k)}{\sum_{k'=1}^{K} \exp(\alpha_{k'} + x_i\boldsymbol{\beta}_{k'})}$$

- The likelihood in terms of β is as follows:

$$f(\beta) \quad \propto \quad \prod_{i=1}^{n}\prod_{k=1}^{K}\Big(\frac{\exp(\alpha_k + x\beta_k)}{\sum_{k'=1}^{K} \exp(\alpha_{k'} + x\beta_{k'})}\Big)^{y_{ik}}$$

- Here β_k is a column vector of p parameters corresponding to class k.

Setting up 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, $\mu_2, ..., \mu_K$). The first one would be determined based on these $K-1$ parameters since $\sum_{k=1}^{K} \mu_{ik} = 1$, i.e., $\mu_{i1} = 1 - \sum_{k=2}^{K} \mu_{ik}$.
- Without this constraints, we can have different set of parameter values giving the same probability. For example,

$$\eta_{i1} = 2, \eta_{i2} = -3, \eta_{i3} = 0.5 \Rightarrow$$

$$p(y_i = 1|\eta) = \frac{\exp(2)}{\exp(2) + \exp(-3) + \exp(0.5)} = 0.8131$$

$$\eta_{i1} = 3, \eta_{i2} = -2, \eta_{i3} = 1.5 \Rightarrow$$

$$p(y_i = 1|\eta) = \frac{\exp(3)}{\exp(3) + \exp(-2) + \exp(1.5)} = 0.8131$$

Setting up priors for the multinomial logistic model

- In the above example, while the values of η's changed the probabilities didn't. This is because we kept the difference between η's as before (we added 1 to all η's). Therefore, for the multinomial logistic model what really matters 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 in *unidentifiable*
- In classical statistics, this is bad, and to avoid this issue for the multinomial logistic model, we could set one set of parameters (usually either β_1 or β_K) to zero.

Setting up priors for the multinomial logistic model

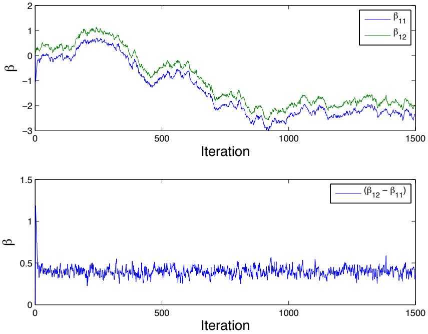
- We do not do this in the Bayesian statistics since it would become difficult to set up symmetric priors (i.e., when in prior all classes have equal probability) based on β.
- If, for example, we assume all categories are equally probable in prior and use $N(0, \tau_\beta^2)$ for all β's, after transformation according to the identifiable multinomial logistic model, the probabilities would not be the same (write down the probability of all classes according to the identifiable model to see this).
- For the multinomial logistic model, we use the unidentifiable setting (no β will be set to zero).
- This does not matter if our goal is prediction.
- If our goal is inference, we can use the posterior distribution of one of the β's (say $\boldsymbol{\beta}_1$, i.e., the first column) as the baseline and subtract other β's (columns 2 to K) from it to make it identifiable.

Example: Snoring and heart disease

- To show how we can set up a unidentifiable model and still perform inference, we use the the snoring and heart disease dataset for the first example (discussed in Agresti, 2002).
- Although the response variable is binary (heart disease/no heart disease), we use a multinomial logit model.
- There would be two regression coefficients now, $\beta_{11} and \beta_{12}$, which are the snoring effects on Class 1 (no heart disease) and Class 2 (heart disease).

Example: Snoring and heart disease

- Because of non-identifiablity the parameters would not converge.
- However, the actual values of β_{11} and β_{12} are not important, rather, the difference between these parameter (which is identifiable parameter in the model) is what matters.
- The difference as we can see converges.



Hierarchical Bayesian models

- Hierarchical Bayesian models are one of the main reasons Bayesian methods have gained increasing popularity.
- To understand these models, we need to understand the concept of exchangeability.

Exchangeability

- Informally, a set of observations $y = (y_1, ..., y_n)$ are exchangeable if in constructing their joint distribution, we believe that the indices are uninformative.
- Exchangeability is important since according to deFinetti's representation theorem if we can judge an infinite sequence of observations to be exchangeable, we can *model* any subset of them as independent and identically distributed (iid) samples from a parametric distribution $p(y|\theta)$, and there is a prior distribution for $p(\theta)$.
- Moreover, there exists a *prior* probability distribution $p(\theta)$ over the parameters of the model such that we can find the unconditional (or marginal) joint distribution of observations.

Exchangeability

- Therefore, we have

$$\begin{aligned} P(y|\theta) = P(y_1, y_2, ..., y_n|\theta) &= \prod_{i=1}^{n} P(y_i|\theta) \\ P(y) = p(y_1, y_2, ..., y_n) &= \int_{\Omega} \prod_{i=1}^{n} P(y_i|\theta) p(\theta) d\theta \end{aligned}$$

- Note that the above theorem is an *existence* theorem. We still need to specify the form of these distributions.

Within-group exchangeability

- Now, assume that we are modeling the housing price, y_i, for a sample 4 bedroom houses in the US.
- We might regard this sample as exchangeable if all we know is the price.
- However, if we also know in which state the house is located, it might be more appropriate to assume exchangeability only within each group since the age distribution would probably be different from one state to another.
- In this case, the price is represented by y_{it}, where i is an index for the states.
- Now the index is not completely uninformative anymore, since we expect different distributions for different i.
- However, we can still use the above theorem and consider each sub-sample, (i.e., for a fixed i) as iid given their own specific parametric model with parameter θ_i.

Within-group exchangeability

- For the above example, for each state j we have

$$P(y_{i.}|\theta_i) \quad = \quad P(y_{i1}, y_{i2}, ..., y_{in_i}|\theta_j) = \prod_{i=1}^{n_i} P(y_{it}|\theta_i)$$

- Therefore, the joint distribution of all samples is

$$P(y|\theta) = \prod_{i=1}^{n} \prod_{t=1}^{n_i} P(y_{it}|\theta_i)$$

- Assuming a normal $N(\mu_j, \sigma_j^2)$ for each state, we have

$$P(y|\mu, \sigma^2) = \prod_{i=1}^{n} \prod_{t=1}^{n_i} N(y_{it}|\mu_i, \sigma_i^2)$$

- We can assume all states have the same variance

$$P(y|\mu, \sigma^2) = \prod_{i=1}^{n} \prod_{t=1}^{n_i} N(y_{it}|\mu_i, \sigma^2)$$

Hyperprior

- Now, as we mentioned before, there exists a prior distribution over parameters, $\theta_1, \theta_2, ..., \theta_n$.
- Similar to y, if we could imagine the infinite sequence of such θ's being exchangeable, we can regard them as being iid samples given the prior distribution $p(\theta|\phi)$ with the parameter ϕ

$$P(\theta|\phi) = P(\theta_1, ..., \theta_n|\phi) = \prod_{t=1}^{n} P(\theta_i|\phi)$$

- ϕ is referred to as *hyperparameter*, which since it is unknown, we need to express our uncertainty using a probability distribution $p(\phi)$.
- Additionally, the joint prior distribution $p(\phi, \theta) = p(\phi)p(\theta|\phi)$
- And the posterior distribution of parameters is

$$P(\phi, \theta|y) \propto P(\phi, \theta)P(y|\phi, \theta) = P(\phi)P(\theta|\phi)P(y|\theta)$$

Hyperprior

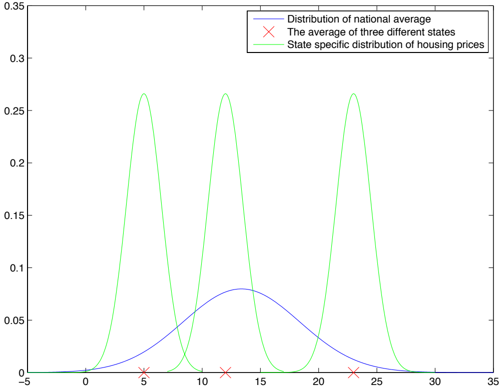
- Note that given θ (i.e., if we fix θ), y becomes independent of ϕ.
- For the housing prices example, we can assume the following priors

$$
\begin{aligned}
\sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)\\
\mu_i &\sim N(\mu_0, \tau_0^2)\\
\mu_0 &\sim N(M, V^2)
\end{aligned}
$$

- Here, we are assuming that τ_0^2 is fixed and only μ_0 is the hyperparameter.
- Moreover, we are assuming that the variance σ^2 is the same for all states, therefore, it is not hierarchical.
- In general, we can set up a hierarchical for some parameters and and not for others. We usually do this if it makes sense and results in a rather simpler model.

Hyperprior

- This graph shows the relation between the distribution of the national average and state specific distributions for three states.



Hierarchical Bayesian GLM

- We can use hierarchical priors for generalized linear models.
- Recall that we mentioned the possibility of using $\beta_j \sim N(0, \tau_\beta^2)$ prior, where all regression coefficients are controlled by one parameter τ_β^2.
- Instead of fixing τ_β^2 at a constant value, we can regard it as an unknown parameter with its own prior distribution.
- In this case, τ_β^2 is the hyperparameter and its prior is the hyperprior.
- For example, we can use the following hyperprior for τ_β^2:

$$\tau_\beta^2 \sim \text{Inv-Gamma}(a_0, b_0)$$

- In this setting, if the covariates are not relevant to the response variable, instead of forcing the individual parameters β_j to become small, the model shrinks τ_β^2 (which is only one parameter) towards zero, which in turns results in shrinkage of β's towards zero.

Hierarchical Bayesian GLM

- This is specially useful in multinomial logit models, where we can use one hypeparameter, τ_j for all the coefficients related to $\beta_{j1}, ..., \beta_{jK}$ related to covariate x_j.
- This way, if a covariate is irrelevant, the corresponding hyperparameter will tend to be small, forcing the coefficients for that covariate be near zero.
- This method is called Automatic Relevance Determination (ARD), and was suggested by Neal (1996).

Model evaluation based on deviance

- Previously, we mentioned that with a 0-1 loss function, we choose the model with a higher posterior probability.
- It turns out (as discussed in Appendix B in Gelman et. al., 2002), the model with the highest posterior probability would have the lowest KL (Kullback-Leibler) information, and as the result the lowest expected deviance.
- As we discussed before, deviance, which is defined as $D(y, \theta) = -2\log(P(y|\theta))$, is a measure of discrepancy (i.e., lack of fit, therefore lower deviance is better).

Model evaluation based on deviance

- The deviance measure as described above, depends on both y and θ.
- If we want to use a measure that depends only on y, we can integrate the deviance over the posterior distribution

$$D_{avg}(y) = E(D(y,\theta)|y)$$

- We can estimate this by using simulated samples from the posterior distribution

$$\hat{D}_{avg}(y) = \frac{1}{L}\sum_{\ell=1}^{L} D(y,\theta^{\ell})$$

Model evaluation based on deviance

- Deviance is especially used when we compare nested models; that is, when we are deciding whether to include the predictor x in the model or not, i.e.:

$$\begin{aligned} M_0 : \quad & y = \beta_0 + \epsilon \\ M_1 : \quad & y = \beta_0 + \beta_1 x + \epsilon \end{aligned}$$

- However, we could decrease deviance by arbitrarily increasing the complexity of model, for example, by adding more predictors into the model.
- In general, it is recommended to use more complex models only when they result in substantial (i.e., statistically significant) improvement in performance (i.e, substantial decrease in deviance).
- The above principle is widely known as Occam's razor stating that "entities should not be multiplied beyond necessity", or in simple words: "everything equal, we should use the simplest solution".

Deviance Information Criterion (DIC)

- When we are relying on deviance, we need a measurement that accounts for the trade-off between complexity and goodness-of-fit.
- In a decision model, this could be done by using a loss function that penalizes larger models (i.e., everything equal, we favor simplicity).
- A simple measure, which does this automatically, is called *deviance information criterion* (DIC) defined as follows (Spiegelhalter et. al. 2002):

$$DIC = \hat{D}_{avg}(y) + p_D$$

- p_D is called *effective number of parameters* and is a measure of complexity

$$p_D = \hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

Deviance Information Criterion (DIC)

- Here, $D_{\hat{\theta}}(y)$ is the deviance when we first average posterior parameters and then calculate deviance (as opposed to integrating deviance over posterior parameters).
- Therefore, we can obtain DIC as follows:

$$DIC = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

- Caution! Although it is easy to use DIC for model evaluation, remember that the best approach is still to use problem specific loss function, and based on the posterior risk, to find the optimal decision rule. Use DIC only when you don't have a better loss function or you simply want to report your findings.

Example: Titanic survival

- Recall the Titanic dataset.
- We consider two nested logistic regression models: Model M_0, which does not include the social class predictor (i.e., only the intercept, age and gender are included), and Model M_1, which includes the social class as well as other variables.
- We fit these two models separately and present the results in the following table

Model	$\hat{D}_{avg}$	$D_{\hat{\theta}}$	p_D	DIC
M_0	2331.6	2329.1	2.5	2334.1
M_1	2216.2	2210.1	6.1	2222.4

- As we can see, M_1 has a smaller DIC, and therefore, provides a better fit. This could be interpreted as statistical significance of social class.