STATS 225: Bayesian Analysis
Linear and Generalized Linear Models

Babak Shahbaba

Department of Statistics, UCI

Bayesian Linear regression models

- Consider the following 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.

Bayesian linear regression models

- 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

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

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

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

Posterior distributions

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

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

- 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.

- That's why Bayesian models do not break down when $p > n$.

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 Inv-$\chi^2(\nu_0, \sigma_0^2)$.

- As we saw before, the posterior distribution of $\sigma^2|x, y, \beta$ is also scaled Inv-χ^2

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

Improper priors

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

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

- For β this is equivalent (in limit) to taking all $\tau_j^2 \rightarrow \infty$.

- The posterior distribution therefore becomes

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

Improper priors

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

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

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.

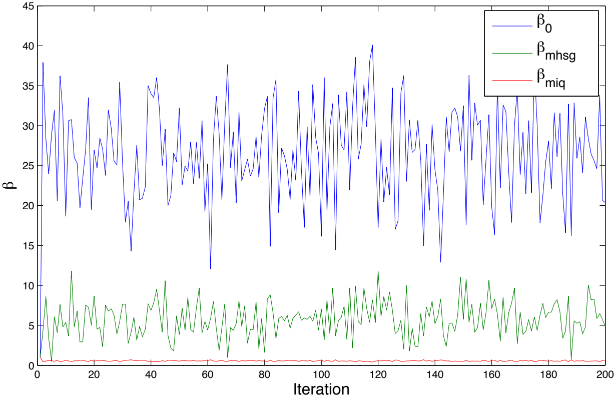
- For our Bayesian model, we use the following broad priors

$$\begin{aligned} \sigma^2 &\sim \text{Inv-}\chi^2(1, 0.5) \\ \beta &\sim N_{p+1}(0, 100^2 I) \end{aligned}$$

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

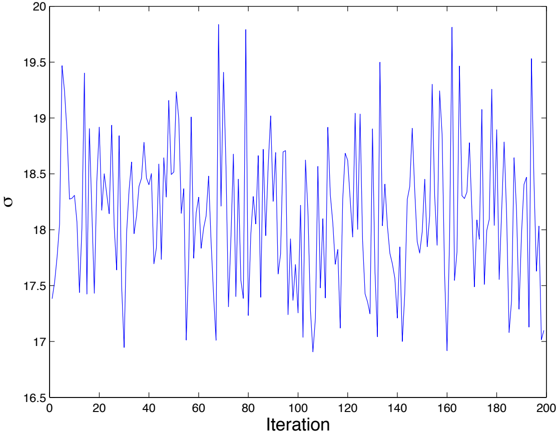
Example: Children's test score

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



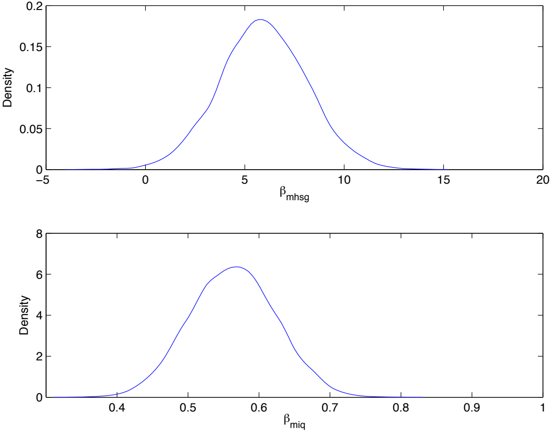
Example: Children's test score

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



Example: Children's test score

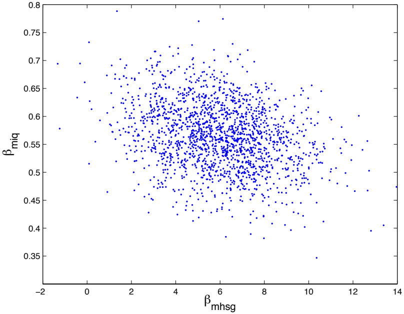
- Using the MCMC samples, we can also plot the posterior distribution of β's



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

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.

Example: Children's test score

- We can also summarize the result of our analysis as follows:

  The posterior estimates and 95% intervals for the regression parameters in the children's test score example.

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

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

- 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

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

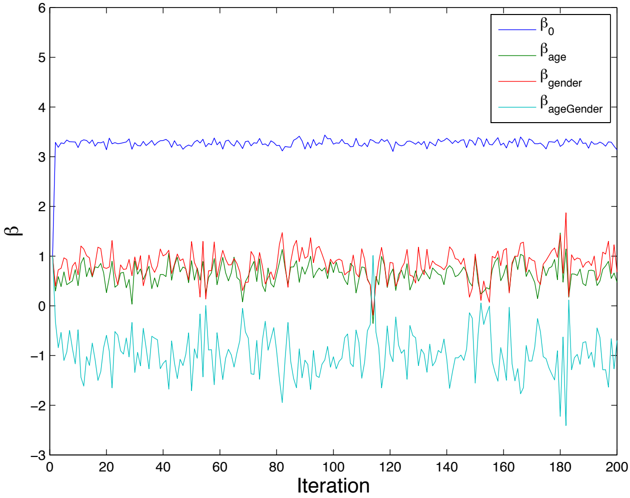
Example: Human, age and body fat

- 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:

$$\begin{aligned} \sigma^2 &\sim \text{Inv-}\chi^2(1, 0.5) \\ \beta_j &\sim N(0, 10^2) \end{aligned}$$

Example 2: Human, age and body fat

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



Example 2: Human, age and body fat

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

  The posterior estimates and 95% intervals for the regression parameters in the children's test score example.

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]

Model checking

- Once we develop a model and perform the required computation to obtain the posterior distribution of parameters, 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 or false (there is a famous quote that “all models are false but some are useful”), rather, how much our inference is affected by our simplifying assumptions.
- 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 (i.e., we do not include them in the model) and treat them as future observations.

Model checking

- 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 $y \sim N(y|\mu, \sigma^2)$. We first obtain the joint posterior distribution of (μ, σ^2), generate $l = 1, ..., L$ samples from this distribution, and for each ℓ, generate $y^{rep} \sim N(\mu^\ell, [\sigma^2]^\ell)$.

- 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 example, for rat tumors model, we first sample β^l, then sample $\theta_j|\mathrm{Beta}(\alpha + n_j, \beta^l + n_j - y_j)$, and then replicate n_j samples from Bernoulli(θ_j) in each group.

Model checking

- 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 $y^{rep} \sim N(x\beta^{\ell}, [\sigma^2]^{\ell})$.

- Note that y^{rep} is different from $\tilde{y}$ (i.e., future observations) since it 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.

- Unlike the frequentists methods where *test statistics*, $T(y)$, are function of data only, in the Bayesian framework, test quantities could be a function of both data and unknown parameters θ.

Model checking

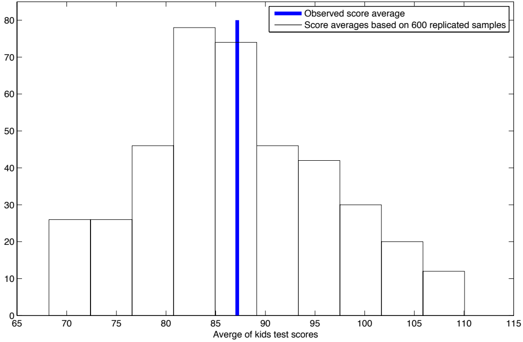
- Typical test quantities are mean, median, variance, min, and max.
- We can use multiple of these tests to evaluate different aspects of the model.
- We can calculate the tail probability

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

  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 $T(y^{rep_\ell}, \theta^\ell) \geq T(y, \theta^\ell)$, where $\ell, 1, ..., L$.
- The model is suspected if the tail probability is close to 0 or 1.

Model checking

- 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 know 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 $\tilde{y} \sim N(\tilde{x}\beta, \sigma^2)$.
- Since we used MCMC simulation, we already have samples from the posterior distribution, which we can use directly (after discarding the pre-convergence samples) to generate $\tilde{y}$.
- Finally, we can use the posterior predictive expectation of $\tilde{y}|y$ (i.e., by averaging the samples) to predict the outcome for future observation.

Prediction

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

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

  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 expectation of β) DO NOT DO THIS IN GENERAL. Always find the value of the function (in this case $\tilde{x}\beta$) over the posterior samples and then average.

Illustration with simulated data

- 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:

$$\begin{array}{lll} \beta_j & \sim & N(0, 2^2) \qquad j = 0, ..., p \\ \sigma^2 & \sim & \text{Inv-}\chi^2(5, 1) \end{array}$$

- 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.

Illustration with simulated data

- Next, we create n samples of predictors x. Since in our model we assume x are independent, we sample them independently form some distribution. Here, we set $n = 150$ and generate $x_{ij} \sim N(0, 1)$ for $i = 1, ..., n$ and $j = 1, 2, 3$. Note that for $j = 0$, we use a column of 1's.

- Now, we can simulated y_i using the assumed linear model with sampled predictors and true parameters

$$y_i | x, \beta, \sigma^2 \sim N(x\beta^*, [\sigma^*]^2)$$

- We regard the first 100 samples as observed data to build our model (we call them the training set), and use the remaining 50 data (we call them the 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.

Illustration with simulated data

- 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 measures (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 situations, 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

Generalized linear model

- The random component identifies the response variable and its probability distribution.
- In most situations, we assume some sort of exchangeability for the set of observed outcome values $y_1, ..., y_n$, and regard them as iid given a parametric model $p(y|\theta)$ from the exponential family.
- Recall that the exponential family includes most of the well-known distributions 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 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.

Generalized linear model

- The systematic component specifies the set of predictors (i.e., explanatory variables) $x = (x_1, ..., x_p)$ 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 $\beta = (\beta_0, \beta_1, ..., \beta_p)$.
- Alternatively,for each observation i, where $i = 1, ..., n$, the linear predictor function is $\eta_i = \beta_0 + \sum_j^p x_{ij}\beta_j$.
- Also, as before, some of predictors could be a transformation (e.g., x^2) of original predictors.

Generalized linear model

- The link function is a monotonic differentiable function that connects the random and systematic components.
- More specifically, if $\mu = E(y|x)$, 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 $\mu_i = \eta_i = x_i\beta$.

Logistic regression model

- As mentioned before, for binary outcome variable, we use the binomial distribution.

$$y_i | n_i, \mu_i \sim \mathrm{binomial}(n_i, \mu_i)$$

  with the Bernoulli distribution as its special case when $n_i = 1$.

- As usual, we define the systematic part of the model $\eta_i = x_i\beta$ (where x_i is a row vector of all observed values for subject i, and β is a column vector of size $p + 1$).

- A common link function for this model is the *logit* function *logit* and defined as

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

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

- Therefore,

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

Logistic regression model

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

$$
\begin{aligned}
P(y|\mu) &\propto \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{n_i-y_i} \\
P(y|\beta) &\propto \prod_{i=1}^{n} \left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(x_i\beta)}\right)^{n_i-y_i}
\end{aligned}
$$

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

$$var(y_i|x_i) = \mu_i(1-\mu_i)$$

Multinomial logistic model

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

  $$y_i | n_i, \mu_{i1}, ..., \mu_{iK} \sim \mathrm{multinomial}(n_i, \mu_{i1}, ..., \mu_{iK})$$

  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 of K elements with $\sum_{k=1}^{K} y_{ik} = n_i$.

- The systematic part is now a vector $\eta_{ik} = x_i \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a matrix of size $(p+1) \times K$.

Multinomial logistic model

- Each column k (where $k = 1, ..., K$) 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 (where $k = 1, ..., J$) can be set to zero without changing the set of relationships expressible with the model.

- Usually, either the parameters for $k = 1$ (the first column) or for $k = K$ (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 general, nonidentifiability does not create problem for Bayesian models). 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

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

- The likelihood in terms of β is as follows:

$$\begin{aligned} P(y|\mu) &\propto \prod_{i=1}^{n} \prod_{k=2}^{K} \mu_{ik}^{y_{ik}} \\ P(y|x, \boldsymbol{\beta}) &\propto \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{\exp(x \boldsymbol{\beta}_k)}{\sum_{k'=1}^{K} \exp(x \boldsymbol{\beta}_{k'})} \right)^{y_{ik}} \end{aligned}$$

- Here $\boldsymbol{\beta}_k$ is a column vector of $p + 1$ parameters corresponding to class k.

Multinomial logistic model

- $\boldsymbol{\beta}$ in general is a $(p+1) \times K$ matrix. The first row, $(\beta_{01}, ..., \beta_{0K})$ are intercepts, and $(\beta_{j1}, ..., \beta_{jK})$ in row $j+1$ are regression parameters associated with the j^{th} predictor.

- x_i is the row vector of predictors value for observation i (including the constant 1 at the beginning).

- y_{ik} is the number of cases in observation i that are in class k.

Poisson model

- When the outcome variable, y, represents counts, we use the Poisson model.

$$y_i|\mu_i \sim \mathrm{Poisson}(\mu_i)$$

- The systematic components are defined as before: $\eta_i = x_i\beta$.
- The usual link function for this model is the log link:

$$g(\mu_i) = \log(\mu_i) = \eta_i$$

- We therefore have

$$\mu_i = \exp(\eta_i) = \exp(x_i\beta)$$

Poisson model

- The likelihood in terms of β can obtained as follows:

$$
\begin{aligned}
P(y_i|\mu_i) &\propto \prod_i^n \exp(-\mu_i)\mu_i^{y_i} \\
P(y_i|\beta) &\propto \prod_i^n \exp[-\exp(x_i\beta)][\exp(x_i\beta)]^{y_i}
\end{aligned}
$$

- Similar to logistic and multinomial models, the variance of $y|x$ in Poisson model depends on the mean and therefore will not be constant

$$var(y_i|x_i) = \mu_i$$

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 $N(\mu_{0j}, \tau_{0j}^2)$.
- We usually 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 the Metropolis algorithm with Gaussian jumps, or more advanced method such as the slice sampler.

Posterior

- Here, we discuss a logistic regression model with normal priors for β.
- Similar approach can be used for multinomial and Poisson models.
- For logistic model, log-likelihood is obtained as follows:

$$\begin{aligned}
\eta_i &= x_i\beta \\
P(y|\beta) &\propto \prod_{i=1} \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\eta_i)}\right)^{n_i - y_i} \\
\log(p(y|\beta)) &= \sum_i \Big[y_i \log[\exp(\eta_i)] - y_i \log[1+\exp(\eta_i)] + \\
&\quad -(n_i - y_i)\log[1+\exp(\eta_i)] \Big] + C_l \\
\log[P(y|\beta)] &= \sum_i \Big[y_i\eta_i - n_i \log(1+\exp(\eta_i)) \Big] + C_l
\end{aligned}$$

Posterior

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

$$\log[P(\beta_j|\tau_0^2)] = -\frac{\beta_j^2}{2\tau_0^2} + C_p$$

- Note that when we are sampling one parameter at a time, since all other parameters are fixed at their current values, their prior probability would be treated as constant and absorbed into C_p (i.e., we don't need to calculate them).

- The log-posterior is therefore:

$$\log[P(\beta_j|y)] = -\frac{\beta_j^2}{2\tau_0^2} + \sum_i \Big[y_i\eta_i - n_i \log(1 + \exp(\eta_i)) \Big] + C$$

Example: Snoring and heart disease

- The objective of this study (Norton and Dunn, 1985, British Medical Journal; 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 sever) is the predictor or explanatory variable.

- The outcome variable is binary (i.e., heart disease = 1, no heart disease = 0).

Example: Snoring and heart disease

- 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

$$\beta_j \sim N(0, 100^2) \qquad j = 0, 1$$

- 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.

Example: Snoring and heart disease

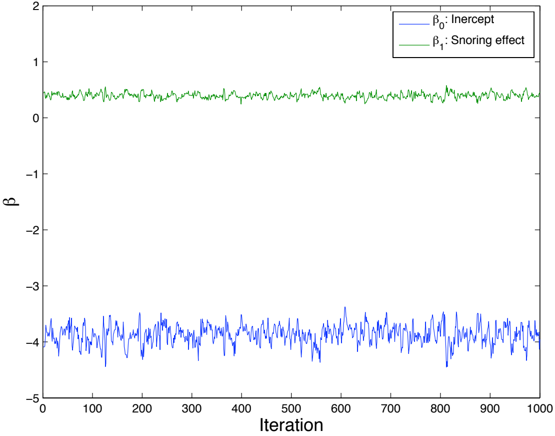
- After constructing the posterior distribution, we sample one parameter at a time using the slice sampler (stepping out and shrinkage), with $w = 2$ and $m = 10$.

- The computer program would be available from the course website, but we explain some computational aspects of it here.

- As usual, it is better to work with the log of posterior probability which is then log(likelihood)+log(prior) up to some constant.

Example: Snoring and heart disease

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



Example: Snoring and heart disease

- 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. With some precautions, we might interpret this as a causal effect.

- 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 the 95% interval does not include 0), we believe the observed effect is statistically significant.

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 the same (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 $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_K$) 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 $N(0, \tau_0^2)$ 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.

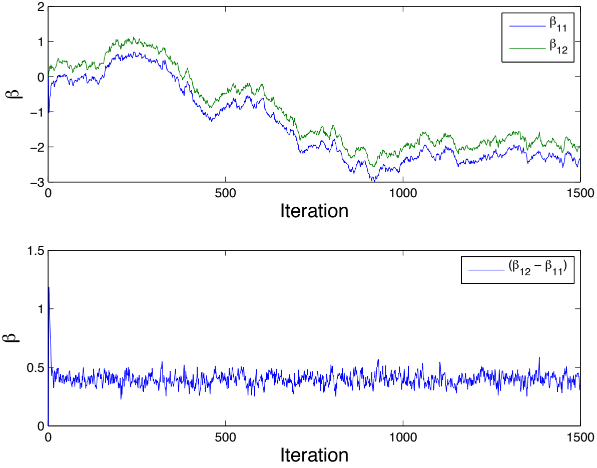
- If our goal is inference, we can use the posterior distribution of one of the β's (say $\boldsymbol{\beta}_1$, i.e., the first column) as the baseline and subtract other β's (columns 2 to K) from it to make it identifiable.

Example: Snoring and heart disease (revisited)

- To show how we can set up a unidentifiable model and still perform inference, we use the 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.
- Recall that the posterior expectations for β_0 and β_1 were -3.8 and 0.4 respectively.
- 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 $N(0, 100^2)$ priors for β_{jk}, and use the slice sampler (stepping out and shrinkage) for simulating samples from the posterior distribution of β one parameter at a time.

Example: Snoring and heart disease (revisited)

- 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 $\beta_{12} - \beta_{11}$.



Example: Snoring and heart disease (revisited)

- 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 these model, $\beta_{12} - \beta_{11}$, shown in the second graph is converging with the posterior expectation 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.

- If our goal was prediction, as it is the case in the next example, we do not need to use the identifiable parameters.

Computation

- Sometimes, we might face the overflow problem when calculating the log-likelihood of multinomial logistic model.
- In general, to avoid overflow when calculating

  $$A = \log(\exp(a_1) + \exp(a_2) + ... + \exp(a_s))$$

  use the following trick (by Radford Neal)

  $$\begin{aligned} m &= max(a_1, ..., a_s) \\ A &= m + \log(\exp(a_1 - m) + \exp(a_2 - m) + ... + \exp(a_s - m)) \end{aligned}$$

Evaluating performance

- To compare the performance of different classification models (e.g., logistic, multinomial), we use average log-probability, accuracy rate, precision and F_1.
- We discussed average log-probability and accuracy before.
- While accuracy measurements are based on the top-ranked (based on the posterior predictive probabilities) category only, precision measures the quality of ranking and is defined as

$$precision \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|y : P(y|x^{(i)}) \geq P(y^{(i)}|x^{(i)})|}\right)$$

Here, y ranges over all classes and $y^{(i)}$ is the correct class of case i. The denominator is, therefore, the number of classes with equal or higher rank compared to the correct class.

Evaluating performance

- F_1 is a common measurement in machine learning

$$F_1 = \frac{1}{K}\sum_{k=1}^{K}\frac{2A_k}{2A_k + B_k + C_k}$$

- Here, A_k is the number of cases which are correctly assigned to class k.
- B_k is the number cases incorrectly assigned to class k.
- C_k is the number of cases which belong to the class k but are assigned to other classes.
- The higher the F_1 measure the better the model.

Evaluating performance

- It is always recommended to compare your results to a baseline model.
- The baseline model in this case is the model that does not use the predictors and instead uses a simple multinomial distribution to model y.
- For this model, we use a noninformative Dirichlet distribution with $\alpha_j = 1$ where $j = 1, ..., K$.
- Based on this model, $(\theta_1, ..., \theta_K | y)$ has a Dirichlet$(1 + y_1, 1 + y_2, ..., 1 + y_K)$ distribution, where θ_k is the probability of observing the k^{th} category, and y_k is the number of training cases that belong to the k^{th} category.
- Using this model, we simply assign all test cases to the category with the highest posterior probability

$$P(\tilde{y} = k | y) = \frac{1 + y_k}{K + \sum_1^K y_k}$$