# ISI-BUDS
# Bayesian 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.

- $\beta$ is a column vector with $p+1$ elements $(\beta_0, \beta_1, ..., \beta_p)$ where $\beta_0$ is the intercept and $\beta_j$ 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

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

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

- $\mu_0$ is typically set to zero (unless we believe otherwise), $\Lambda_0$ should be sufficiently broad.

# Posterior distributions

- The posterior distributions of $\beta$ 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 $\Lambda_0$.

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

# Posterior distributions of $\sigma^2$

- Now, we want to obtain the posterior distribution of $\sigma^2$
- Given $\beta$, again we have a simple normal model with observations $y_i$ with known mean $(x\beta)$, unknown variance $\sigma^2$, 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-$\chi^2$

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

# Improper priors

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

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

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

- The posterior distribution therefore becomes

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

# Improper priors

- The posterior distribution of $\sigma^2$ also has a closed form

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

# Example: Children's test score

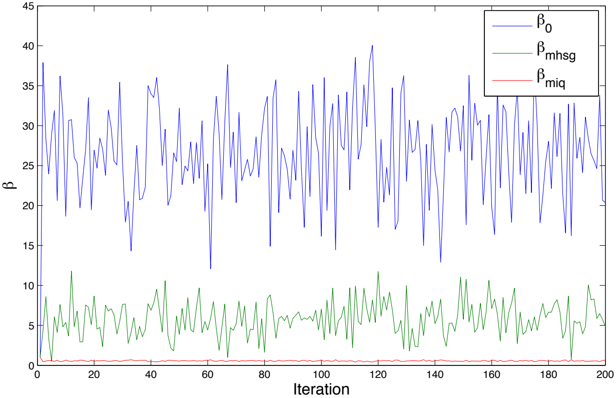
- Consider the children's test score example discussed by Gelman and Hill (2007).
- In this example, we are interested in the effect of mother's education (mhsg) and her IQ (miq) on the cognitive test score of 3 to 4 year old children.
- For our Bayesian model, we use the following broad priors

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

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

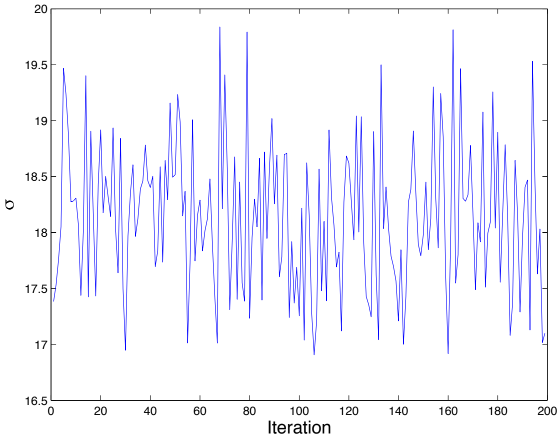
# Example: Children's test score

- The following plot shows the trace plot of posterior samples for $\beta$'s



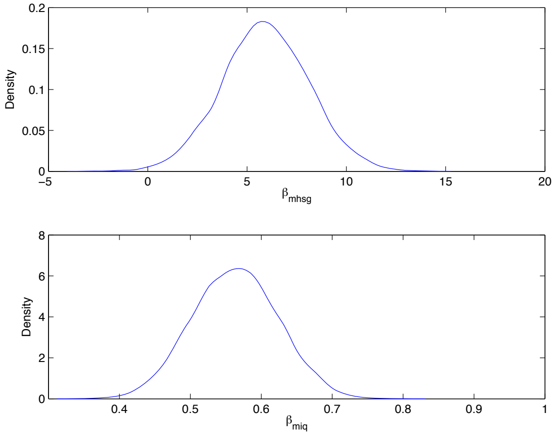
# Example: Children's test score

- The following plot is the trace plot of posterior samples for $\sigma$



# Example: Children's test score

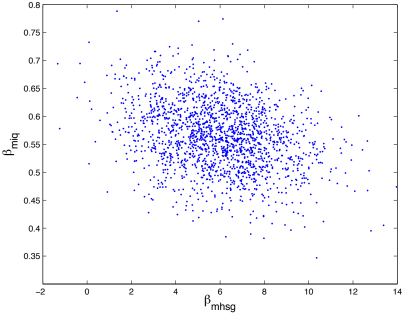
- Using the MCMC samples, we can also plot the posterior distribution of $\beta$'s




- 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 $\beta_{mhsg}$ and $\beta_{miq}$




- Note that in general, $\beta$'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 |
|---|---|---|
| $\beta_0$ | 25.7939 | [14.4, 37.2] |
| $\beta_{\mathrm{mhsg}}$ | 5.9278 | [1.6, 10.3] |
| $\beta_{\mathrm{miq}}$ | 0.5633 | [0.4, 0.7] |
| $\sigma$ | 18.2 | [16.9, 19.4] |

# 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 $(\mu, \sigma^2)$, generate $l = 1, ..., L$ samples from this distribution, and for each $\ell$, 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.

# Model checking

- For linear regression models, we generate samples $(\beta^{\ell}, [\sigma^2]^{\ell})$ from the posterior distribution of $(\beta, \sigma^2)$, 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 $\theta$.

# Model checking

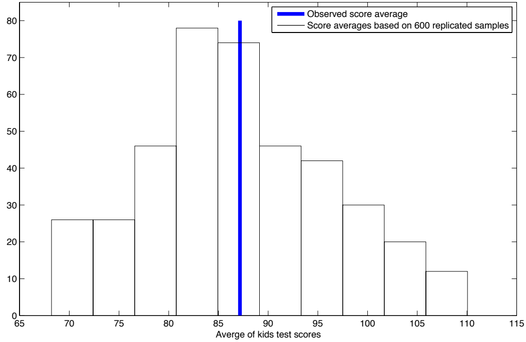
- Typical test quantities are mean, median, variance, min, and max.
- We can use multiple of these tests to evaluate different aspects of the model.
- We can calculate the tail probability

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

  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 $(\beta, \sigma^2)$ 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:

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

  where $L$ is the number of posterior samples $\beta^{\ell}$ after convergence.

- Although for the above model, we could use $\tilde{x}\hat{\beta}$ (where $\hat{\beta}$ is the posterior expectation of $\beta$) 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.

# Generalized linear model

- Recall that for generalized linear models we need to specify three components:
  - A random component
  - A systematic component
  - A link function

# Prior

- Within the Bayesian framework, we also need to specify priors on model parameters.
- A common prior for $\beta$ 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 $\beta$'s can be performed using the Metropolis algorithm with Gaussian jumps.

# Posterior

- Here, we discuss a logistic regression model with normal priors for $\beta$.
- 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 $\beta_j$, the log-prior probability given $\tau_0^2$ 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 $\beta$

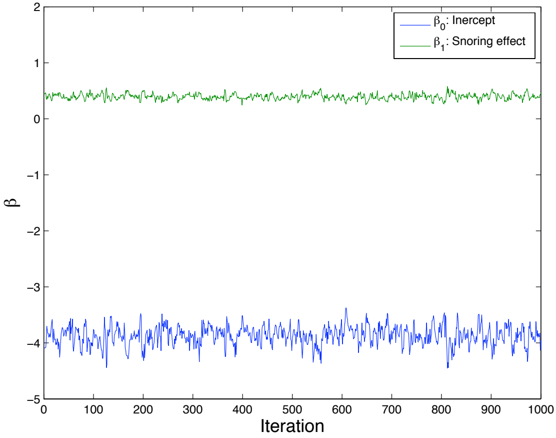
$$\beta_j \sim N(0, 100^2) \qquad j = 0, 1$$

- The role of prior here is mainly to provide a reasonable range for possible values of $\beta$ (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

- 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 |
|---|---|---|
| $\beta_0$ | -3.87 | [-4.24, -3.53] |
| $\beta_1$ | 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.

# Bayesian GLM in R

- We can fit Bayesian GLM models using the function `stan_glm()` from the package `rstanarm`.
- You can find more information at

  `https://mc-stan.org/rstanarm/articles`

- The general form is similar to `glm()`

  `stan_glm( formula, family, data)`

- For the prior, you can use the default setting or specify your own prior (preferred).