Random effects: When and why

Subhash Lele

2024-08-27

Review

- We reviewed the basic statistical concepts behind the likelihood inference and the Bayesian inference.
- We looked at how to write a JAGS model function for some linear and generalized linear regression models and use it in the package ‘dclone’ to get the Bayesian credible intervals as well as the frequentist confidence intervals based on the asymptotic normal distribution.
- We also discussed some of the reasons to use the MCMC approach to conducting statistical inference, either Bayesian or frequentist.

Random effects

- We will now generalize the models to make them relevant to some complex practical situations. - These are some of the situations where the analytical approaches to Bayesian and likelihood inference are difficult to impossible to implement.
- The question of estimability of the parameters becomes much more relevant but difficult to diagnose.
- The method of data cloning particularly is useful for diagnosing estimability of the parameters.
- Although the models are much more complex, the coding component does not increase in complexity.
- We will also discuss prediction of missing data.

Detection error in occupancy studies (Latent variables)

Let us visit the occupancy model again.

- In practice, the assumption that you observe the occupancy status correctly is somewhat suspect. For example, if we are looking for a bird species, if the bird never sings or gives some sort of a cue, it is extremely difficult to know they are there.
- Hence, even if the species is present, we may note that it is not present. This is called 'detection error'.
- How can we model this?

Statistical model for detection error

- Let W_i denote the true status of the i-th cell.
- In our previous notation, now $P(W_i = 1) = \phi$.
- The observed value, generally denoted by Y_i could be 1 or 0 depending on the true status.
- We assume that the species are never misidentified, then we can write $P(Y_i = 1|W_i = 1) = p$ and $P(Y_i = 0|W_i = 1) = 1 - p$. Moreover, $P(Y_i = 0|W_i = 0) = 1$.
- Probability of detection is p and probability of occupancy is ϕ.
- How can we infer about these given the data?

Hierarchical model

Notice that we only observe Y_i's and not the W_i's. *The unobserved variable W_i is called a latent variable.*

To write down the likelihood function, we need to compute the distribution of the observed data Y_i. It is easy to see that $P(Y_i = 1) = p * \phi$ and $P(Y_i = 0) = 1 - p * \phi$. We can write down the likelihood based on this. However, we are going to write this as a hierarchical model.

Hierarchy 1:

$$W_i \sim Bernoulli(\phi)$$

Hierarchy 2:

$$Y_i | W_i = w_i \sim Bernoulli(p * w_i)$$

Writing the hierarchical model in R and JAGS

Bananas!!!

Let us look at the model again.

- It is clear that we can estimate the product $p * \phi$ given the data.
- But decomposing this product in p and ϕ is impossible. This is called 'non-estimability'. In this case, this also is non-identifiability.
- There are several combinations of p and ϕ that lead to the same $p * \phi$ and hence the same distribution of the observed data.
- Such situations are not uncommon when dealing with the hiearchical models in general, and measurement error models in particular.

Bayesian learning and non-identifiability

- We should not make any inferences about the probability of occupancy based on these data. You can change the priors and see what happens to the posteriors. You might find it interesting and educational.

Non-estimability: If there two or more values in the parameter space that lead to identical likelihood value, such values are called 'non-estimable'.

Note: You may recall from the linear regression that if the covariates are perfectly correlated to each other, the regression coefficients are non-estimable. if covariate X_1 is perfectly correlated with X_2, these covariates separately give no additional information.

A Bayesian result and its data cloning version

If the posterior distribution converges to a non-degenerate distribution as the sample size increases, it implies that set of parameters is non-estimable.

If the posterior distribution converges to a non-degenerate distribution as the number of clones increases, it implies that set of parameters is non-estimable.

An immediate consequence of this result is that the variance of the posterior distribution does not converge to 0 (instead it converges to some positive number).

Let us modify our data cloning code to see what happens.

Data cloning diagnostics

- If the posterior distribution converges to a non-degenerate distribution as the sample size increases, it implies that set of parameters is non-estimable.
- If the posterior distribution converges to a non-degenerate distribution as the number of clones increases, it implies that set of parameters is non-estimable.
- An immediate consequence of this result is that the variance of the posterior distribution does not converge to 0 (instead it converges to some positive number).

Availability of the estimability diagnostics is one of the most important features of data cloning. It will warn you if your scientific inferences could be misleading.

What is the recourse?

- If the parameters are non-estimable, the only recourse one has is to change the model (add assumptions or collect different kind of data).
- There is always a possibility that, although the full parameter space may not be estimable, a function of the parameter might be estimable. If such a function is also of scientific interest, we can safely conduct scientific inferences based on estimates of such a function of the parameters.

One possible solution-Replicate surveys

Suppose we visit the same cell several times. Assume that the visits are independent of each other and the true occupancy status remains the same throughout these surveys, then we can estimate the parameters. The model can be written as a hierarchical model:

Hierarchy 1: $W_i \sim Bernoulli(\phi)$ Hierarchy 2: $Y_{ij}|W_i = w_i \sim Bernoulli(p * w_i)$

Notice that hierarchy 2 depends on hierarchy 1 result.

We can easily modify the earlier code to allow for multiple surveys. We will do such a modification with two surveys for each cell.

Replicate surveys and data cloning

# Random effects in regression: Why and when?

We have shown how hierarchical models can be used to deal with measurement error. Now we will look at a few examples where we use them to combine data across several studies.

We will start with a simple (but extremely important) example that started the entire field of mixture as well as hierarchical models (Neyman and Scott, 1949). Researchers in animal husbandary wanted to know how to improve the stock of animals such as milk cows and bulls. This played an important role in the 'white revolution' that lead to improving the nutrition in many countries. Following is a somewhat made up and highly simplified situation.

White revolution

Suppose we have n cows. We want to know which cows have good genetic potential that can be passed on to the next generation. Each cow might have only a few calfs. We measure the amount of milk by each calf.

Let $Y_i j$ be the amount of milk produced by the j-th calf of the i-th cow. We can consider a linear model that uses the 'cow effect' (genetic) and 'environmental effect' to explain the amount of milk. This is same as one way ANOVA model.

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Under the usual Gaussian error structure, we know that

$Y_{ij} \sim N(\mu_i, \sigma^2)$ where $i = 1, 2, ..., n$ and $j = 1, 2$. $(\mu_i = \mu + \alpha_i)$

Neyman-Scott problem

There are $2 * n$ observations and $n + 1$ parameters.

- The number of parameters increases at the same rate as the number of observations. Note that the ratio of parameters to observations converges to 0.5.
- Roughly speaking, for the MLE to work, this ratio needs to go to 0. Generally the number of parameters is fixed and hence this condition is satisfied. We simply do not have much information about each μ_i as there are only two observations corresponding to it.
- It also turns out that the ML estimator of σ^2 converges to $0.5 * \sigma^2$. Hence it is not consistent even though the number of observations corresponding to it do converge to infinity.
- This was a major blow to the theory of maximum likelihood. Although it turns out Fisher had implicitly answered it a decade before this paper.

How can we reduce the number of parameters?

1. If there are covariates such as weight of the mother, mother's milk production are available, we can model $\mu_i = X_i\beta$.
2. Suppose covariates are not available or difficult to assess what to use as a covariate. If we assume that cows are kind of similar to each other, then we can assume that they come from a population of cows such that $\mu_i \sim N(\mu, \tau^2)$. It turns out, under such an assumption, we can estimate the parameters (μ, σ^2, τ^2) consistently.

Hierarchical model

This model is a hierarchical model.

Hierarchy 1: $Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$ Hierarchy 2: $\mu_i \sim N(\mu, \tau^2)$

For a Bayesian approach, we put priors on the three parameters. This forms the third hierarchy. ## Further applications

This simple model can be used in many different situations.

- Measurement error in covariates in regression
- Random intercept regression model to account for missing covariates
- Kalman filter models for time series with measurement error are of the same kind with a bit more complexity as we will see the third part.

Let us see what makes it a difficult model to analyze using the likelihood approach.

Integration over the latent variables

In order to write down the likelihood function, we need to compute the marginal distribution of the observations. Remember μ_i are not observed. Hence we have to integrate over them.

$$f(y_{ij};\mu,\sigma^2,\tau^2) = \int f(y_{ij}|\mu_i)g(\mu_i)d\mu_i$$

Again, this is not a precise statement but it gives you the idea. This integral is one dimensional and hence can be computed analytically and also numerically. However, in many cases the dimension of the integral is quite large and hence neither of these solutions are available.

Possible solutions

- In some cases, one can obtain Laplace approximation to this integral.
- INLA and related methods rely on this approximation. But I have a manuscript that is ready to be submitted that shows INLA and related approximation methods are seriously flawed.
- The most general approach is based on the MCMC algorithm.

MCMC based solution (Bayes and ML): R and JAGS program

Generalization

The best thing about the MCMC approach is that we can modify the prototype program to do Generalized linear mixed models.

Let us see how we can change the program to do Poisson regression with random intercepts. This is useful for accounting for missing covariates in the usual Poisson regression. This can also be used to account for site effect in abundance surveys.

Mathematically the model is:

Hierarchy 1: $log(\lambda_i) \sim N(log(\lambda), \tau^2)$

Hierarchy 2: $Y_i|\lambda_i \sim Poisson(\lambda_i)$

R code for GLMM

Estimating the effect of a treatment from multi-center clinical trials

- In clinical trials, we are interested in estimating the effect of the treatment.
- One of the simplest forms of clinical trial is where we split a group of patients in two groups randomly. One of the group gets the treatment and the other gets a placebo. We can then estimate the difference in the outcomes. This may be done using a simple t-test if the outcome is a continuous measurement.
- If the patients are quite different from each other in terms of say age, Blood pressure or some such physical characteristics that may affect the outcome, we adjust them by using a regression approach.
- We include these other covariates and the treatment/control indicator variable in the model. The effect of the treatment after adjusting for other covariates can be studied using such a regression model.

Random intercept in LM and GLM

Often in practice, we may not have access to such covariates. Using random intercept model in regression is one way out of such a situation. In this case, we do consider the differences between patients but without ascribing them to any specific, known values of the covariates. The model one may consider is:

$$Y_i = \beta_{0i} + \beta I_{(Treatment)} + \epsilon_i$$

As we have seen in the previous example (Neyman-Scott problem), this model is non-estimable. One way to make it estimable is by using a hierarchical structure:

Hierarchy 2: $\beta_{0i} \sim N(\beta_0, \tau^2)$

Accounting for random effects

Homework: Check the validity of the following statement without doing any mathematics. You can use data cloning to do that.

This leads to estimability for β, the parameter of interest. Although it does not lead to estimation of the variances $c(\tau, \sigma)$.

An approach not described in this course: We can use profile likelihood for the parameter β. This eliminates the 'nuisance parameters' β_0, τ, σ. Computing the profile likelihood and quantification of uncertainty for inferences based on it for hierarchical models can be tackled using data cloning. This could be another course!

Another example

Suppose the outcome is binary, survival for 5 years vs failure before 5 years. In this case, a convenient model is a binary regression model such as a Logistic regression model.

Hierarchy 1:

$$P(Y_i = 1) = \frac{exp(\beta_{0i} + \beta I_{(Treatment)})}{1 + exp(\beta_{0i} + \beta I_{(Treatment)})}$$

Hierarchy 2: $\beta_{0i} \sim N(\beta_0, \tau^2)$

This is an example of a Generalized Linear Mixed Model. Fortunately, for this model all parameters are estimable as we will see using data cloning. The random intercept here could be accounting for differences in the clinical centers (hospitals), assuming we have only two patients, one in control and one in the treatment group.

More complex example

If we have multiple patients in each group, we may include another random effect to account for differences in the patients within each group. For example, we may consider a model:

Hierarchy 1:

$$P(Y_{ij}=1)=\frac{exp(\beta_{0i}+\beta I_{(Treatment)}+\beta_{1j})}{1+exp(\beta_{0i}+\beta I_{(Treatment)}+\beta_{1j})}$$

Hierarchy 2: $\beta_{0i}\sim N(\beta_0,\tau^2)$ $\beta_{1j}\sim N(\beta_1,\tau^2)$

We can include interactions between random effects and so on. We will not go into these complex models in this course.

R code modification

Let us see how one can modify the prototype program to analyze the random intercept Logistic regression model.

Wait, wait .. there is more!

It should be clear by now that we can modify any Bayesian analysis to get Maximum likelihood estimate quite easily by adding one dimension to the data and a do loop over the clones.

In the next part, we will discuss how to analyze time series data.