Chenhao Qian
ID: 001035275

Bayesian Regression Model for Abalone Data

Introduction

In this project, I aim to explore the regression analysis on a specific dataset from the Bayesian stand point. The data consists of observational records and label variable measure under traditional approach for abalone, a common seashore living creature. The goal is to establish a model to predict the age of abalone with the observational predictors.

Given the data, I considered two separate models: the regular linear regression model, with the Bayesian approach, and a finite mixture model of multi-variate normal distributions. The linear model works as expected and present problems that are caused by the structure of the data (multicollinearity) and the nature of the model setting (fixing a constant variance on residuals). A mixture model, in general, provides high level of flexibility that allows the mean to change nonlinearly. However, through fitting the model with Gibbs sampling approach, I found that the model is extremely hard to implement in the sense that the sampler fails in a few runs. I will discuss the reason of the failure and some potential approaches to fix the problem later.

Chenhao Qian
ID: 001035275

1. Look into the Data

The data comes from the online UCI Machine Learning Archive called “Abalone”. The original dataset consists of over 4000 rows of data with 8 predictor variables and one response variable of interest. The detailed information for all the variables is given below with a few rows of the data as a representation. The whole dataset don’t contain any missing data.

Name	Data Type	Unit	Description
Sex	nominal	--	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	--	+1.5 gives the age in years

The goal of the data is to build a model that can predict Y using Sex and X1 through X7. That is, we aim to predict the number of rings (which is linear to age) of the abalone based on information from easy-to-get measurements such as length, weight, etc.

2. Preprocessing Data

Notice that the categorical variable “Sex” is nominal with three different classes “M”, “F”, and “I”. A proper way to introduce a predictive model for such variable is to substitute it with two indicator variables. For example, let M=1 if the observation has Sex=”M”, and 0 otherwise; and F=1 if Sex=”F” and 0 otherwise. Then, in the case of both M and F being 0, the observation has Sex=”I”. Such setting prevents the risk of introducing non-existing incremental relationship among the three classes.

However, the method of dealing with categorical variable stated in the above paragraph actually indicates three similar but separable models, each for a given class. The fitting process for each category would be identical. Since our goal is to explore if a predictive model is plausible and how we can construct a proper model, fitting all three categories is redundant.

For this reason, I took the subsample of the data where Sex is labeled “M” as the actual data for analysis.

3. Regular Bayesian Linear Regression Model

The first and obvious model of choice is the simple multiple regression model. However, I did not do it with the usual non-Bayesian approach as that will be trivial using built-in functions of R. I set up a regression model.

a) Model Setting

$Let\ y^{(n\times1)}\ be\ the\ vector\ of\ the\ response\ variable\ and\ X^{(n\times8)}\ be\ the\ predictor\ matrix.$

Notice there is an additional column that has value 1 for each row to represent the intercept of the model. Thus the model is given below.

$$Y|\beta,\sigma,X \sim N(X\beta,\sigma^2 I)$$

Here $\beta,\sigma^2$are the parameters we aim to find through fitting the model. I is simply an 8-dimension identity matrix.

Since I am doing a regression, it is OK to condition every parameter on X to set up priors.

Let $p(\beta,\sigma^2|X)\sim\sigma^{-2}$ be the regular non-informative conjugate prior. The posteriors are derived in the next section.

b) Conditional Posterior

Since our only need is to be able to draw sample parameters from the posterior distribution, I will only derive the conditional posterior density.

Given data, the posterior density of σ^2 follows a scaled inverse-chi-squared distribution with degree of freedom n-8 and scale parameter s^2, where s^2 is the MSE of the fitted model. The conditional density of $\beta|\sigma^2$ is then just a multivariate normal with mean $\hat{\beta}$ and variance $V_\beta\sigma^2$. Here $\hat{\beta}$ is the MLE of the model and V_β is the covariance matrix of the fitted parameters.

$$\sigma^2|Y,X\sim Inv-\chi^2(n-k,s^2)=(n-k)s^2/\chi^2(n-k\,)$$

$$\beta|\sigma^2,Y,X\sim N(\hat{\beta},V_\beta\sigma^2)$$

c) Initial Value Estimate

Ideally we should choose crude values as initial values to generate the sample if we wish to perform a Gibbs sampler. However, since there are analytical forms of the posterior distribution, I skipped this step and directly draw from the posterior for inference.

d) Inference on Parameters

The summary of 5000 posterior draws is given below along with the model summary of the regular frequentist linear model.

```
> summary(cbind(lm_beta,lm_sigma_square))
   INTERCEPT           X1                  X2                  X3               X4
 Min.    : 0.6127   Min.    :-25.8151   Min.    :-28.1649   Min.    :-13.59   Min.    : 0.4899
 1st Qu.: 4.0530   1st Qu.: -5.1004   1st Qu.: -0.1142   1st Qu.: 11.29   1st Qu.: 7.2897
 Median : 4.9304   Median : -0.3815   Median :  5.5125   Median : 16.40   Median : 8.9918
 Mean   : 4.9303   Mean   : -0.4271   Mean   :  5.4987   Mean   : 16.39   Mean   : 9.0200
 3rd Qu.: 5.7843   3rd Qu.:  4.2520   3rd Qu.: 11.1166   3rd Qu.: 21.46   3rd Qu.:10.7165
 Max.   :10.8457   Max.   : 27.8725   Max.   : 34.0925   Max.   : 41.58   Max.   :18.5930
       X5                 X6                 X7           lm_sigma_square
 Min.   :-31.170   Min.   :-25.106   Min.   :-6.322   Min.   :4.330
 1st Qu.:-21.081   1st Qu.:-13.301   1st Qu.: 6.901   1st Qu.:4.811
 Median :-19.232   Median :-10.246   Median : 9.538   Median :4.936
 Mean   :-19.208   Mean   :-10.226   Mean   : 9.553   Mean   :4.940
 3rd Qu.:-17.273   3rd Qu.: -7.089   3rd Qu.:12.227   3rd Qu.:5.060
 Max.   : -8.899   Max.   :  4.963   Max.   :22.974   Max.   :5.699
```

As we can see here the posterior mean resembles closely to the frequentist estimate which corresponds to the theoretical posterior. Also, the parameters for X1 and X2 include 0 in their central half of probability distribution, which indicates insignificant effect on the model. This again corresponds with the result of the frequentist model.

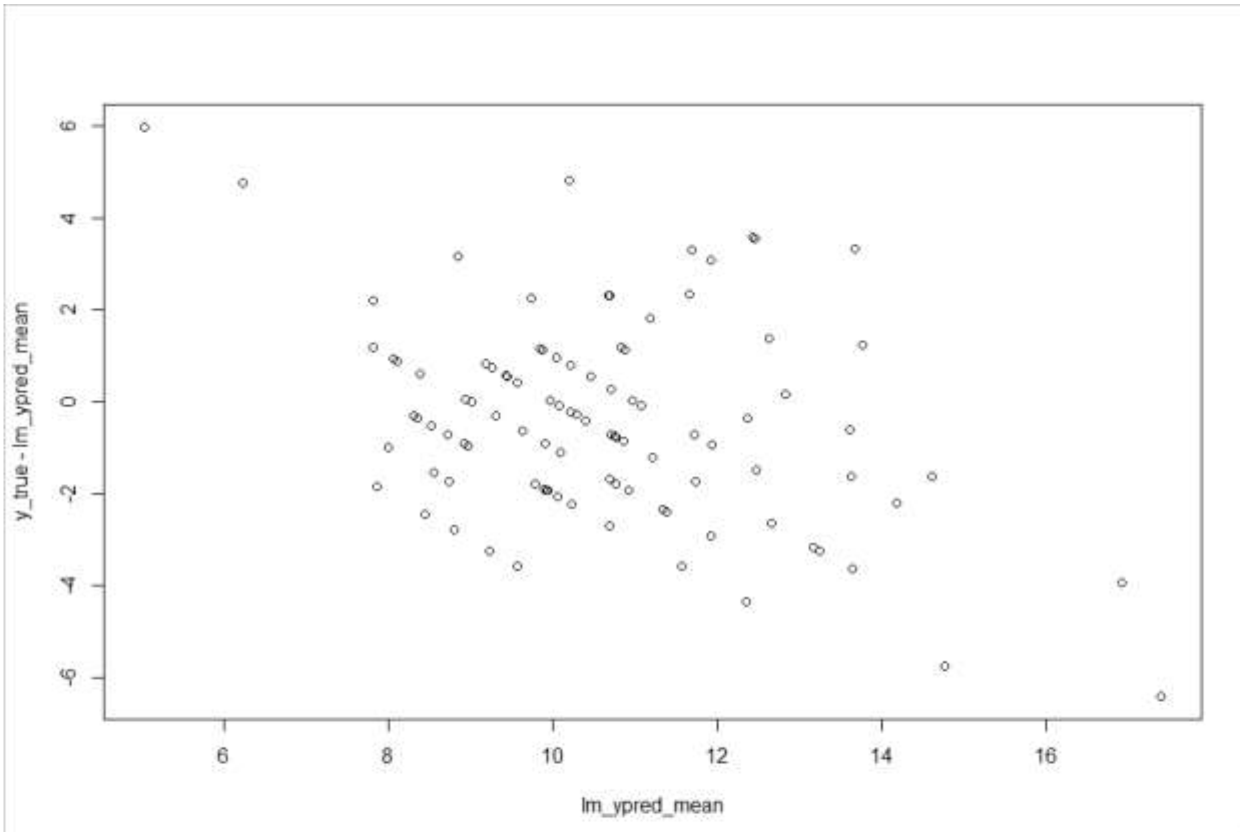
```
Call:
lm(formula = Y ~ ., data = abalone.train)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4081 -1.4408 -0.2878  0.9532 11.6063 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.9194     0.5664   8.685  < 2e-16 ***
X1           -0.4218     3.1966  -0.132    0.895    
X2            5.4761     3.8291   1.430    0.153    
X3           16.4316     3.4474   4.766 2.07e-06 ***
X4            9.0445     1.1747   7.699 2.55e-14 ***
X5          -19.2169     1.3011 -14.769  < 2e-16 ***
X6          -10.2736     2.0341  -5.051 4.97e-07 ***
X7            9.5448     1.8371   5.196 2.34e-07 ***
```

e) Predicted Value on Test Data

To check the predictive power of the model, I fitted the reserved test data to each of the posterior draws and obtained the predicted Y value. Instead of using each y_pred as a point estimate, I took the sum of all predictive draw to eliminate the randomness induced by drawing single data point from a distribution. Looking at the residual plot, I notice that the residuals are distributed reasonably close to 0 with an acceptable disperse. There are a little drift-away's when the predicted value is near boundary but still doesn't affect the effectiveness of the model.



4. Finite Gaussian Mixture
   a) Model Layout

The layout of the finite mixture of multi-variate normal model is given below.

Let $w_i = (x_i, y_i)$

Assume w_i has H multivariate normal components.

We can introduce latent variable z_i, vector indicator of which component w_i belongs to

$$w_i | z_i = h \sim N(\mu_h, \Sigma_h),\ p(z_i = h) = \pi_h$$

$$f(w_i) = \sum_{h=1}^{H} \pi_h N_{p+1}(w_i | \mu_h, \Sigma_h)$$

This will induce to which is a convenient linear regression within each component.

$$f(y_i|x_i) = \sum_{h=1}^{H} \pi_h(x_i) N(y_i|\beta_{0h} + x_i\beta_{1h}, \sigma_h^2)$$

$$\pi_h(x_i) = \frac{\pi_h N_p(x_i|\mu_h^{(x)}, \Sigma_h^{(x)})}{\sum_{h'=1}^{H} \pi_{h'} N_p(x_i|\mu_{h'}^{(x)}, \Sigma_{h'}^{(x)})}$$

b) Model Parameters

The parameters to be estimated are

- Latent indicator z_i
- $\pi = (\pi_1, \cdots, \pi_H)$
- For each $h \in \{1, \cdots H\}$
- μ_h: $mean\ vector\ of\ the\ multi-normal\ component$
- Σ_h: $covariance\ matrix\ of\ the\ multi-normal\ component$

These parameters outline the joint model for the multivariate normal mixture setting while the regression Coefficient β_h, variance σ_h^2 will not be estimated using Bayesian inference but just a simple regression with all data related to the component.

c) Choices of Prior

The prior distributions are chosen as below:

- $\pi = (\pi_1, \cdots, \pi_H) \sim Dirichlet\left(\frac{1}{H}, \cdots, \frac{1}{H}\right)$
- $p(\mu_h, \Sigma_h) \propto |\Sigma_h|^{-\frac{d+1}{2}}\ Jeffery's\ Prior$
- $\Sigma_h \sim Inv-Wishart(\Lambda_0^{-1})$
- $\mu\_h|\Sigma_h \sim mvN(\mu_{0h}, \Sigma_h)$

Note that there is no need to specify prior for regression parameters since these parameters will be estimated with regular regression with data limited within component

d) Conditional Posterior

The full conditional posterior distributions are given below with all distributions being known ones.

- $\Pr(z_i = h|w_i, \pi, \mu, \Sigma) = \frac{\pi_h N_{p+1}(w_i|\mu_h,\Sigma_h)}{\sum_{h'=1}^{H} \pi_{h'} N_{p+1}(w_i|\mu_h,\Sigma_h)}$
- $\pi|w, z \sim Dirichlet(1 + n_h),\ n_h is\ occurance\ of\ z_i = h$
- $\Sigma_h|w, z \sim Inv - Wishart_{n_{h-1}}(S_h^{-1})$
- $\mu_h|w, z, \Sigma_h \sim mvN(\overline{w_h}, \frac{\Sigma_h}{n_h})$
- Where $S_h = \sum_{i:z_i=h}(w_i - \overline{w_\iota})(w_i - \overline{w_\iota})^T$

e) Model Failure and Difficulties

I experienced much difficulty in this model of the project and what's worse is that the model still failed. One thing that was hard for me was how to relate the clustering nature of components to the regression output of the variable of interest. Another difficulty if caused by the large number of parameters in the model, especially when the number of components was set to be large. A last issue is with the dimensionality of the model parameters. This problem is most outstanding when trying to set up the Gibbs sampling process to make sure that I get all the subscripts correctly.

5. Discussion

The main reason for the model to fail, I believe is due to the nature that the data doesn't have clusters existing. For that reason, the assignment of components for each row of data is extremely unstable and varies greatly with no patterns. Furthermore, some of the component weights tend to 0 in just a few iterations of Gibbs sampling as the computer learns that there aren't that many components within the data. This will lead to the matrix becoming singular and prevent the process of Gibbs sampling being continued.

If we really wish to proceed with the model, that is, insisting there exist a certain number of clusters within data, we probably have to add strong assumptions that induce strict constrains to prevent the weights to decrease beyond a certain threshold. However, even if we obtain result from such model, the robustness of the result may be questioned.