Proposal of STATS 551 Project

Minxuan Chen, Ruixuan Deng, Yizhan Zhang

November 1st, 2023

1 Introduction

Many models of great interest in Mathematics and Statistics have been motivated by the goal of have a goodness of fit and good prediction of their future values.

1.1 Dataset and Background

Diabetes mellitus (DM), a prevalent non-communicable disease, represents a significant global health concern, necessitating the development of accurate early detection methods based on established risk factors. The Diabetes dataset, a widely utilized benchmark dataset in regression analysis, was originally introduced in "Least Angle Regression" by *Efron et al.* in 2004, published in the *Annals of Statistics*. This dataset comprises 442 instances and encompasses 10 baseline variables, including age, sex, body mass index (BMI), average blood pressure, and six blood serum measurements deemed pertinent to predicting diabetes progression.

The primary objective of our project is to compare **Ordinary Least Squares (OLS)** regression and **Bayesian regression** techniques using this specific dataset. Our focus lies in identifying the most predictive variables associated with diabetes progression. Specifically, we aim to develop a model capable of predicting diabetes progression one year after the baseline measurement, quantified by a response of interest measure.

It is worth noting that the original literature does not provide information regarding the dataset's collection process, precluding discussions about potential biases introduced during sampling or suggestions for improving data collection methods. Nevertheless, a general best practice in data collection is to include a diverse range of samples to ensure the broad applicability of analytical conclusions. In our context, the absence of specific details about the data collection process does not impact our selection of prior distributions. Given the dataset's widespread applicability in regression problems, choosing it as our dataset remains a favorable and pragmatic choice.

1.2 Exploratory Data Analysis

2 Plan and Timeline

First and foremost, we need to choose an appropriate data set to do bayesian analysis. On 21st October, we first met each other and we formed a group. On 23rd October, we talked about what we should do, then We agreed to combine Bayesian analysis and linear regression methods to do a project on Bayesian linear regression.We observe and discuss the results of Bayesian linear regression and classical linear regression, observe the similarities and differences between them, and see if the results of applying the theory of Bayesian analysis can explain the model better.Below is our timeline of the final project.
Discuss topics and choose an appropriate data set. (week 1 Oct 23-Oct 29):

- Ruixuan Deng takes the lead in acquiring and preprocessing the dataset.
- Based on the data, Minxuan Chen identified the model to analyze and study the data and the methods we used
- According to the model and method Minxuan Chen proposed, Yishan Zhang put forward some additions to help him perfect the use of the model

Propose and Fit a Preliminary Model (Weeks 2-3 Oct 30-Nov 12):
Starting from the second week, we will organize discussions twice a week, report the progress and results of each other, share the problems encountered in the project, and then discuss to solve these problems together.

- Minxuan Chen will take the lead in proposing and fitting the preliminary Bayesian model based on the chosen topic and dataset.
- Ruixian Deng will support by ensuring the data is correctly prepared and fed into the model.
- Yishan Zhang will assist in interpreting the initial model results and providing feedback.

And we make sure that we have set 6 prior distributions. Each one will finish two of them. After that, we will discuss together and help each other modify their results and finally combine all of them.

Potential Difficulties:

Limited time for extensive model exploration and validation. Data quality issues that may affect model performance.

Examine Initial Results and Adjust Models Further if Needed (Week 4 Nov 13-Nov 19):

- Yishan Zhang will carefully examine the initial model results, identifying strengths and weaknesses.

- Minxuan Chen will assist in conducting sensitivity analyses and identifying potential areas for improvement.
- Ruixuan Deng will start preparing the presentation, focusing on key findings and insights.

Potential Difficulties:

Time constraints may limit the depth of the analysis and the number of models that can be considered.

Complete the model design and analysis of the results, and prepare for class presentation (Week 5 Nov 27- Dec 3):

- All the team member will obtain the final results and start model checking and model comparing
- Think about why does this model work well and the other model doesn't
- Think about how to interpret and present the results and prepare for the final presentation

Potential Difficulties:

Balancing technical depth with audience comprehension in the presentation. Ensuring all team members have a clear understanding of the material for the presentation.

Summarize and Write a Final Report (Week 6 Dec 4-Dec 10):

- All team members will collaborate on summarizing the project's key findings, recommendations and results in the final report.
- Each team member will contribute their expertise to ensure a comprehensive and accurate report. we will write down 2 prior distributions of the results each other and try our best to give an accurate and clear presentation.

3 Methods

3.1 Basic Settings

The analysis of our project is based on Bayesian linear regression model. Here, we list some basic settings.

In a linear regression model, we are interested in the relationship between X and y. Generally, we assume

$$X|\phi \sim p(X|\phi), y|X, \theta \sim p(y|X, \theta)$$

in which ϕ, θ are parameters, and they have no common components.

Specifically, for the likelihood, we adopt the assumptions of classical linear regression, which include linearity, independence, normality, and constant variance.

$$y|X, \theta \sim N(X\beta, \sigma^2 I).$$

Here we take $\theta = (\beta, \sigma^2)$. Moreover, we assume there is no error when measuring X.

For the prior of ϕ, θ, we assume

$$p(\phi, \theta) = p(\phi)p(\theta)$$

i.e. ϕ and θ are independent apriori.

Since we are more interested in the relationship between X and y, we focus on the parameter θ, From these assumptions, we have

$$p(\phi, \theta|X, y) = p(\phi|X)p(\theta|X, y)$$

So, to learn about θ, we only need to look at $p(\theta|X, y)$. In other words, our models can be conditional on X.

3.2 Choices of Prior

For the specific prior of θ, we plan to analyze the following choices

3.2.1 Standard Noninformative Prior (M_1)

$$p(\theta) = p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

3.2.2 Simple Conjugate Prior (M_2)

$$\beta|\sigma^2 \sim N(m_0, \sigma^2 C_0)$$
$$\sigma^2 \sim Inv - \chi^2(v_0, s_0^2)$$

3.2.3 Complicated Conjugate Prior

$$\beta|\sigma^2 \sim N(b_0, \sigma^2 B)$$
$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

in which B can be

1. Diagonal matrix
2. $B = gI \qquad (M_3)$
3. $B = g(X^T X)^{-1}$

in which, g is a hyperparameter.

We are more interested in the third choice, which is also known as Zellner's g-prior. There are 3 appealing ways to set the value of g.

1. Cross Validation (M_4)
2. Empirical Bayes Estimation (M_5)

$$\hat{g} = \text{argmax}_g \, p(y|g)$$

3. Full Bayesian Model: Assign a hyper prior to g. (M_6)
   One choice is

$$p(g) \propto g^{-3/2} e^{-n/2g} \quad (g \sim \text{Inv} - \Gamma(\frac{1}{2}, \frac{n}{2}))$$

3.3 Analysis Pipeline

Now, we have six models, denoted as $M_1, M_2, \ldots, M_6$, which we intend to apply to our data.

For these models, our approach is to utilize both analytical and simulation techniques to investigate the posterior properties. Specifically, we will analyze the posterior distribution of both the model parameters and hyperparameters, focusing on quantities such as posterior means and variances. We anticipate that a substantial portion of the work will involve Markov Chain Monte Carlo (MCMC) methods.

Upon obtaining results from these models, our next steps will encompass model checking and model comparison.

For model checking, our plan involves identifying a test statistic, denoted as T, determined through Exploratory Data Analysis (EDA).

In terms of model comparison, we will divide this process into two distinct parts. First, we intend to use criteria like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to discern the best model. Second, we aim to perform variable selection from a Bayesian perspective, focusing on Zellner's g-prior (without a hyperprior on g).

Specifically, considering two design matrices X_1 amd X_2, we want to compare Model 1

$$\begin{aligned} y &\sim N\left(\alpha 1_n + X_1\beta_1, \sigma^2 I\right) \\ \beta \mid \sigma^2 &\sim N\left(b_1, g_1\sigma^2\left[X_1^T X_1\right]^{-1}\right) \\ p\left(\alpha, \sigma^2\right) &\propto 1/\sigma^2 \end{aligned}$$

and Model 2

$$\begin{aligned} y &\sim N\left(\alpha 1_n + X_2\beta_2, \sigma^2 I\right) \\ \beta \mid \sigma^2 &\sim N\left(b_2, g_2\sigma^2\left[X_2^T X_2\right]^{-1}\right) \\ p\left(\alpha, \sigma^2\right) &\propto 1/\sigma^2 \end{aligned}$$

in which g_1 and g_2 are determined by some methods such that they are "best" for the corresponding design matrix.

For simplicity, we may opt to perform this selection among nested models, where the column vectors of X_1 form a subset of those in X_2. In this process, we intend to leverage Bayes factors.

3.4 Summary

In summary, for the purpose of these analysis, we expect to

1. Apply various models to the diabetes dataset.
2. Comparing the pros and cons of these models and identifying the "best" model for this problem.
3. Explore variable selection within a Bayesian framework for linear regression.