Prediction on Insurance Cost with Bayesian Statistics and Machine Learning Methods

MingZhong Gao (mgao24@jhu.edu), Chenyang Li (cli110@jhu.edu),
Ziyan Lin (zlin25@jhu.edu), Han Wang (hwang178@jhu.edu),
Weizhuo Wang (wwang111@jhu.edu)
Department of Applied Mathematics and Statistics
Johns Hopkins University

December 17, 2018

Abstract

In order to make predictions on insurance costs, we fit linear regression models and classification models. We use Bayesian method with conjugate and non-conjugate priors to find appropriate values of coefficients for our regression models. Next, we apply classification methods to our data, aiming to find a best model that can predict the class of insurance charge.

1 Introduction

Health insurance has been a popular society issue over the country for many years. It is also highly related with our daily life in terms of health and finance. It is always ideal for a family to have an affordable and well-covered health insurance. However, clients are sometimes confused or discriminated by the policy introduced by an insurance company. In this project, our primary objective is to predict the insurance charge with respected to individual features such as age, gender, smoking habit, number of children in family and geographical locations. We would construct models both on predicting the numerical values of expenditure and on classifying the insurance types. Method we used in this study includes Linear Regression and Bayesian Statistics, which helps us to find proper coefficients. Classification with Machine Learning models is also involved.

1.1 Data Description

The data we used is retrieved from

`https://www.kaggle.com/mirichoi0218/insurance/data`.

It incorporates 1338 observations with 6 predictive variables (Age, Sex, BMI, Children, Smoker, Region) and one response variable (Charges). We split the

data into Train and Test sets with a ratio of 8:2. We also manually created a binary variable called Charge_bi, it equals to 1 if the insurance charge is higher than 14000 and 0 if not. Since our response variable suggests a highly right skewed Normal distribution, we take its Log form when we apply linear regression.

2 Linear Regression

Since the graph of data is like a log-normal distribution, we transformed the response variable to the form $\log(Y) = Z$. We apply a linear regression model to simulate the relationship between variables:

$$Z = X\beta + \varepsilon$$

Here, Z is a $n \times 1$ column vector with known observations, $Z \sim \text{MVN}(X\beta, \sigma^2)$ under our assumption. X is a $n \times p$ matrix with known observations, β is a $p \times 1$ column vector with unknown parameters and ε is a $n \times 1$ column vector also with unknown parameters, where n is the number of observations in training data and p is the number of variables we included in the model. We first perform variable selection via BIC methods in order to find the most efficient model. The model given by BIC procedure is

lcharges $\sim$ age + sex + bmi + smoker + children + region + age×smoker + age×children + bmi×smoker (lcharges here is the log form of variable charges)

MSE for this model is 31.69.

2.1 Conjugate Prior

We first set conjugate prior for β and σ^2 to get Bayesian estimations of the parameters. Priors are given as follows:

$$\begin{aligned} \pi(\beta \mid \sigma^2) &\sim \text{Norm}(\mu_\beta, \sigma^2 V_\beta) \\ \pi(\sigma^2) &\sim \Gamma^{-1}(a, b) \end{aligned}$$

where we set $\mu_\beta = 0$, $V_\beta = 10^4 I_{20}$, $a = 1$, $b = 1$. This leads to the following posterior distribution

$$\begin{aligned} p(\beta \mid \sigma^2, Z) &\sim \text{Norm}(\mu_\beta^*, V_\beta^*) \\ p(\sigma \mid Z) &\sim \Gamma^{-1}(a^*, b^*) \end{aligned}$$

where

$$\begin{aligned}
\mu_\beta^* &= (V_\beta + X^T X)^{-1}(V_\beta^{-1}\mu_\beta + X^T Z) \\
V_\beta^* &= \sigma^2(X^T X + V_\beta^{-1})^{-1} \\
a^* &= a + \frac{n}{2} \\
b^* &= b + \frac{1}{2}[\mu_\beta^T V_\beta^{-1}\mu_\beta + Z^T Z - \mu_\beta^T(X^T X + V_\beta^{-1})\mu_\beta^*]
\end{aligned}$$

Mean Squared Error for this estimation is 0.149. This value is much smaller than that of the direct estimation of BIC model.

One advantage that we choose conjugate prior is that we can get a close form posterior, from which we can sample and estimate parameters directly. In this model, we choose noninformative priors with large variations 10^4, as we no nothing about the parameter. The distribution of the prior is called Normal-Inverse-Gamma (NIG) prior. To ease our computation, V_β is specified with diagonal structured. Another advantage is that we can get the posterior with invariant procedure regardless of the scale of the regressors. In our model, we have 5 parameters and it will be difficulted to get their posteriors without a known conjugate form.

2.2 Non-Conjugate Prior

Following the non-conjugate prior assumption, we can have that

$$\begin{aligned}
Y \mid \beta, \sigma^2 &\sim \text{Norm}(X\beta, \sigma^2 I_n) \\
\beta_i &\sim \text{Norm}(0, \sigma_0^2), \quad i = 1, 2, 3, \cdots, 20 \\
\sigma^2 &\sim \Gamma^{-1}(c, d)
\end{aligned}$$

Next, we can have the likelihood function and prior as follows

$$\begin{aligned}
L(Y \mid \beta, \sigma^2) &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\{-\frac{1}{2\sigma^2}(X\beta - Y)^T(X\beta - Y)\} \\
p(\beta_i) &\propto \exp\{-\frac{\beta_i^2}{2\sigma_0^2}\}, \quad i = 1, 2, 3, \cdots, 20 \\
p(\sigma^2) &\propto (\sigma^2)^{-c-1}\exp\{-\frac{d}{\sigma^2}\}
\end{aligned}$$

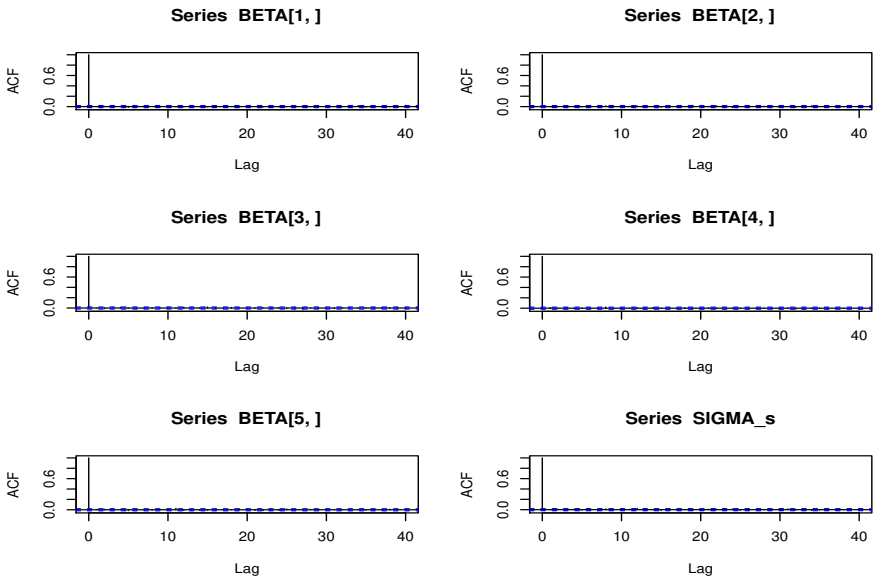
Then, we can have the posterior function and derive several parameters for the full conditional posterior function

$$
\begin{aligned}
p(\beta, \sigma^2) &\propto (\sigma^2)^{-(\frac{n}{2}+c)-1} \exp\{-\frac{1}{2\sigma^2}[(X\beta - Y)^T(X\beta - Y) + 2d] - \frac{\sum_{j=1}^{20} \beta_j^2}{2\sigma_0^2}\} \\
a_i &= \sigma_0^2 \sum_{k=1}^{n} x_{ki}^2 + \sigma^2 \\
\tilde{\mu}_i &= -\sigma_0^2[\sum_{j=1} x_{ji}(\sum_{k \neq i} x_j k \beta_k) - Y_j]/a_i \\
\tilde{\sigma}_i^2 &= 2\sigma_0^2\sigma^2/a_i \\
\tilde{\alpha} &= \frac{n}{2} + c \\
\tilde{\beta} &= \frac{(Y - X\beta)^T(Y - X\beta)}{2} + d
\end{aligned}
$$

Finally, we can get the closed form of full conditional posterior functions of β and σ^2

$$
\begin{aligned}
p(\beta_i \mid \beta_{j \neq i}, \sigma^2) &\sim \text{Norm}(\tilde{\mu}_i, \tilde{\sigma}_i^2) \\
p(\sigma^2 \mid \beta) &\sim \Gamma^{-1}(\tilde{\alpha}, \tilde{\beta})
\end{aligned}
$$

Since we have the closed form, we implemented Gibbs Sampling method to sample parameters β and σ^2. We took $\sigma^2 = 10$, $c = 100$ and $d = 100$, and we get our results. ACFs of some of these parameters are given below.



Means of β are 7.6314e-2, 3.5369e-4, 1.0153e+1, 2.8449e-4, -1.0760e-1, -3.8343e-2, -1.6343e-1, 2.0421e-1, -1.6979e-1, 5.6347, 1.0427e+1, -1.2256, 1.4104e+1, 1.6021e-3, 3.1598e-5, -4.4675e-5, -1.2068e-4, -1.0690e-3, 4.0921e-4, 1.3306e-3 and mean of σ^2 is 70.9268. The residual of the train set is 81.5046 and that of the test set is 80.4123. The graph of acf shows that our method converges quite well.

2.3 Conclusion

In general, we prefer the conjugate prior setup to the non-conjugate one in term of efficiency and accuracy. From the result, we can see that the sum of squared residuals of conjugate Bayesian model is less than BIC model, which indicates conjugate Bayesian model is better. This is reasonable as we have chosen the parameters with the BIC model and then do the conjugate Bayesian. The information of prior distribution improves the estimations of the parameters.

As for Gibbs Sampling method, although the result converges, the process is slow because of the large number of parameters. Also, the results of this method are not ideal because the residuals are too large compared with those derived from the basic BIC model and from conjugate prior hypothesis. We analyzed this and found it as a result that the method might converge to the local optimum rather than the global optimum. We tried different initial values but the problem was not fixed. We think there shall be more advanced methods to deal with this kind of situations.

3 Classification

In real life, insurance company usually provide some different types of insurance, such as basic cover or full cover. Therefore, from the perspective of insurance company, classification methods are much more useful than linear regression. The mean of charges is around 14,000, and as a result, we set this value as the threshold and divide the dataset into two groups. The high group including all charges above 14,000 is assigned with 1, and the low group including all charges below 14,000 is assigned with 0.

3.1 Methods

We use seven different classification approaches on the dataset: random forest, logistic regression, quadratic discriminant analysis, linear discriminant analysis, naïve Bayes, support vector machine, and k-nearest neighbor. Based on predictor variables (age, sex, bmi, children, smoker, region, age:smoker, age:children, and bmi:smoker) that are selected by BIC, we apply all seven approaches. Moreover, each approach involves a step of cross validation with cv = 10 in the end. Accuracy of each approach is calculated by the mean of cross validation technique. The results are in the following.

Table 1: Results of Classifications

Classification	Accuracy
Random Forest	0.9215351812366738
Logistic Regression	0.9095892716866795
QDA	0.7528616316911682
LDA	0.908843003029963
Naïve Bayes	0.8991358994501178
SVM	0.908843003029963
KNN (k = 5)	0.9222702278083268

3.2 Conclusion

K-nearest neighbor has relatively higher accuracy, and QDA has the lowest accuracy. Except for QDA, all other approaches have accuracies around 90%. Since these classification approaches generate similar output, it is hard to say which approach has best predictive ability. Naïve Bayes, a classifier applying Bayes theorem, has a high accuracy as well, and therefore we can conclude that Bayesian method is useful for classification on the dataset.

The high accuracy reflects that our model, selected by BIC, is appropriate for the dataset. According to these predictors, classification algorithms have enough information to predict charge levels. Thus, we strongly suggest that insurance company should collect personal information, such as age, sex, bmi, children, smoker, and region, and then deal with interaction terms, such as age:smoker, age:children, and bmi:smoker, to accurately recommend appropriate insurance plan for customers. Random forest, logistic regression, linear discriminant analysis, naïve Bayes, support vector machine, and k-nearest neighbor are important classification approaches that should be considered carefully by insurance company.

4 Summary

Based on our results above, we conclude that the best regression model for predicting charges is the BIC model with conjugate prior setup, and the best classification method to distinguish insurance classes is KNN. Apart from theoretical interpretations, our study indeed has some real-life applications. For example, if people are confused about the annual budget on their health insurance, or when they feel discriminated by the policy introduced by an insurance company, this study can be useful. As long as the basic personal information (age, gender, BMI, number of children, smoking habit and region) is given, one can easily decide the approximate expenditure on insurance, as well as whether should be treated as high insurance class.

5 Improvement

One of the biggest restrictions on our study is related to Bayesian Statistics. We applied Bayesian method to improve the estimation accuracy of coefficients since MCMC method is ideal for parameter estimation, but the result of Gibbs Sampler was not satisfactory. There are mainly two reasons that we can think of, one is the large number of variables and the other one is due to the lack of information on parameter distributions. Having too many variables makes the Gibbs procedure runs slow, but it is due to the nature of the data, we cannot make improvement on that. However, what we can do is to do more research on the distributions of the parameters, once we obtain a more proper assumption about the priors, we would have a better result on Bayesian estimation.