BIOSTAT625 Final Project

Xiangeng Fang, Xinjue Li, Yixuan Zeng

Abstract:

Insurance claim is an effective tool to decrease the cost of loss caused by various risks of illnesses. It is of great significance for us to analyze the data related to insured people in all aspects and calculate the insurance claim efficiently and accurately to provide reasonable decisions to choose appropriate claims for insured and financial decisions for insurers. Insurance claims are jointly influenced by various factors such as the demographic information of the beneficiary group and the type of insurance patients chose, which might be complicated and time-consuming to estimate the charge by traditional actuarial approaches. This report illustrates how regression models and logistic models can be applied to analyze the data more efficiently and improve the process of predicting insurance claims for different insured people.

Introduction:

In recent times, to obtain a more comprehensive understanding of the beneficiary group, electronic health data appears in health-care systems such as insurance companies, hospitals and government departments frequently. Our method might sum up the work or technique to define the strategy. Regression models and logistic models are all very efficient ways to analyze data and give predicted values.

Our found dataset consists of the data of the insured group covering 2008 to 2010. There are 8 csv files in the dataset, which includes 3 beneficiary summary files, 2 different carrier claims files and 3 different kinds of patients files. To be more specific, the 3 beneficiary summary files include the basic information such as what diseases the insured are suffering from. The 2 carrier files have the information from the perspective of the insurance companies that analyze the claim of the beneficiary. The 3 patients files contain the inpatients, outpatients and patients under prescription drug events. It's adequate for us to analyze and predict from several dimensions.

Medicare is the federal health insurance program for three specific categories of the people. First, people who are 65 or older could have the access to get the insurance. Second, people who are younger but with disabilities could obtain medicare insurance. Third, people with end-stage renal disease(permanent kidney failure requiring dialysis or a transplant, sometimes called ESRD) also have the right to get insurance. There are four parts of medicare. Part A refers to the hospital insurance, which covers inpatient hospital stays, care in a skilled nursing facility, hospice care, and some home health care. Part B is regarded as medical insurance, which covers certain doctors' services, outpatient care, medical supplies, and preventive services. Part C could be interpreted as the insurance provided by the Health Maintenance Organization, which are initiated and formed by local doctors, hospitals, and other health care providers. Part D is prescription drug coverage, which helps cover the cost of prescription drugs (including many recommended shots or vaccines).

Explanatory data analysis

Data preprocessing

Since the size of our data is quite big, we apply some big data computing skills during our data preprocessing. For instance, we use *fread* from *data.table* package to read data instead of using *readcsv* function. Furthermore, we also try to avoid loops and use more vectorized function through the whole process. The detailed data preprocessing procedure is presented as follow:

1. The three beneficiary tables, which contain beneficiary information from 2008 to 2010, are selected and combined. The variables of interest, which includes chronic information, birth date, race, sex and DESYNPUF_ID are retained, while all others are screened.
2. The variable AGE is defined, based on the birth date and corresponding year. Note that there might be duplicated beneficiary for these three years.
3. For those categorical variables of interest, rearrange their values. For those binary variables corresponding to chronic condition, value = 0 means without this disease, while value = 1 means having this disease.
4. Load the inpatient claims information. Select the variables of interest.
5. For the inpatient claims, we find that there exists multiple rows with same claim ID but different segment. Considering these segments, they might have different providers and claim payment amount. Hence, we regard them as two different segments. However, the second segment with the same claim ID usually lack of information about claim dates. We fill the missing values with the claim dates from the first segment.
6. The tables with inpatient claims information and the table with beneficiary information are combined based on DESYNPUF_ID and YEAR (i.e. attach the corresponding year's beneficiary information with each claim)
7. Considering providers, we calculate provider size for each providers (number of claims for each provider), and we delete those providers with number of claims less than 5.

Visualization

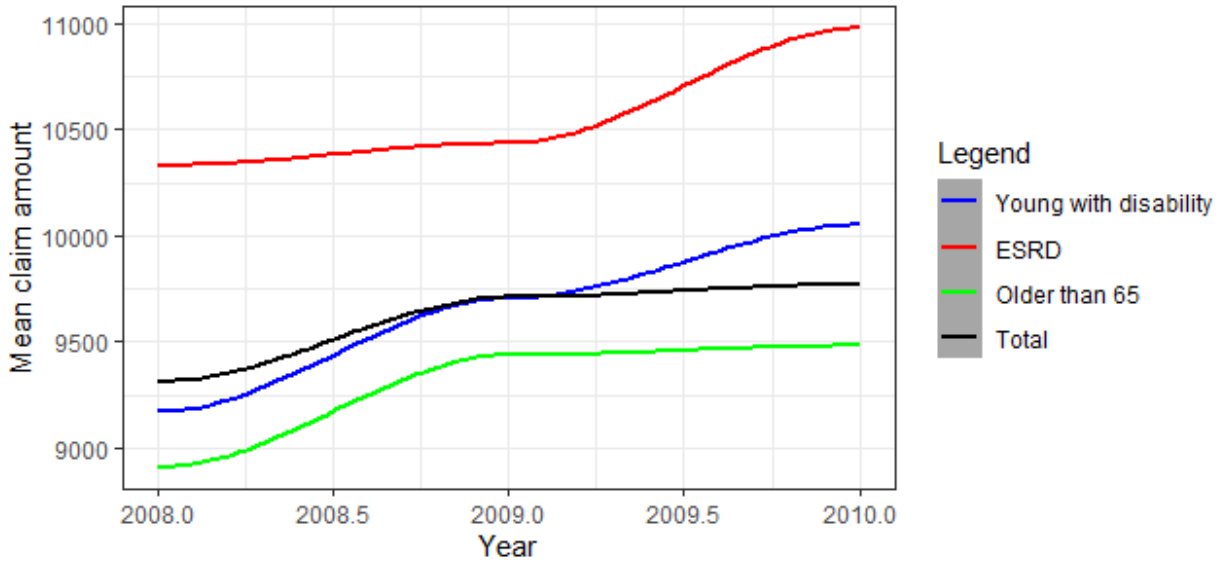

Figure 1: Trends of different group in mean claim amount from 2008 to 2010

We obtained a line graph to examine the relationship between several different groups in mean payment amount from 2008 to 2010. The red line represents the trend of beneficiaries with end-stage renal disease(ESRD), which has the highest mean payment amount among these several groups. It first grows steadily from 2008 to 2009 then rises faster from 2009 to 2010. The blue line shows the changes of beneficiaries with disabilities, but the age of whom is less than 65 years old. It increased from 2008 to 2009 while decreasing since 2009. The green line demonstrates the relationship between mean claim amount and the left group of beneficiaries whose age is more than 65 years old through years. And this line has the smallest mean claim amount among different groups. Its growing pace slowed down in 2009 to 2010 compared to 2008 to 2019. The black line is the expression of the total people which are the whole group that has been researched.

From the general view of this line graph, we can find that the people who are suffering from end-stage renal disease show more necessity to access the insurance. This could suggest that people who have some severe illnesses tend to concentrate more on medicare than other groups. This is a significant conclusion that we could draw from the graph, which is really helpful for the insurers to use. For insurance companies, it is critical to use mathematical and statistical tools to assess the risk of financial loss and predict the cost that insurance companies should pay according to these individual characteristics, which is related to the future development prospect and financial condition of insurance companies.

Model

From the above analysis, we can see that different diseases might significantly affect the use of Medicare. The treatment of a disease is highly dependable on healthcare provider. Motivated by He, Kalbfleisch, Li & Li (2013), we apply the GLM with fixed providers' effect to analysis the claim of Medicare. Two responses of interest are payment amount for each claim (continuous) and reclaim rate in the next 3 months (binary). GLM with identity link and logit link can be applied separately to analysis the relationship between response and risk factors considering the providers effect.

GLM with fixed providers' effect

Denote the total number of providers as m, and denote the number of subjects from provider i ($i = 1, \ldots, m$) as n_i. Let $n = \sum_{i=1}^{m} n_i$ be the total count. For subject j of provider i, let Y_{ij} denote the response, and $X_{ij} \in \mathbb{R}^p$ denote the explanatory variables. Under the GLM assumption, given $\boldsymbol{X}_{ij}$, outcome Y_ij follows a exponential family distribution with parameters θ_{ij} and ϕ

$$f(Y_{ij}|\boldsymbol{X}_{ij}; \theta_{ij}, \phi) \propto \exp\left\{\frac{Y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)}\right\},$$

where functions $a(\cdot)$ and $b(\cdot)$ are both known function, and $\theta_{ij} = \gamma_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}$ is a linear predictor relating to provider effect γ_i and coefficients $\boldsymbol{\beta}$. Note that γ_i here is provider-specific, which means that for different providers, we have different intercepts. The function a and b are subject to the specific distribution that the outcome of interest follows. In this project, we focus on the normal (claim payment amount) and binary (3-month reclaim rate) outcomes corresponding to the Medicare data. Given observed data $\{(Y_{ij}, \boldsymbol{X}_{ij}), i = 1, \ldots, m, j = 1 \ldots, n_i\}$, the log-likelihood can be written as

$$l(\boldsymbol{\gamma}, \boldsymbol{\beta}) \propto \sum_{i=1}^{m}\sum_{j=1}^{n_i}\{Y_{ij}(\gamma_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}) - b(\gamma_i + \boldsymbol{X}_{ij}\boldsymbol{\beta})\}$$

where $\gamma = [\gamma_1, \ldots, \gamma_m] \in \mathbb{R}^m$. Based on the above information, we can calculate the gradient and hessian matrix regarding the data. Then, we can apply Newton-Raphson method to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

Model evaluation

Table 1: Model RMSE

	Linear regression		Logistic regression	
Evaluation	Training	Testing	Training	Testing
Without provider effect	9242.313	9146.822	0.4541	0.4542
With provider effect	9110.175	9269.195	0.4482	0.4593

Without provider effect, the RMSE from Linear Regression and Logistic Regression is mostly weaker than the model with provider effect in the performance of the training dataset and test dataset, which once again affirmed the importance of provide effect.

Results

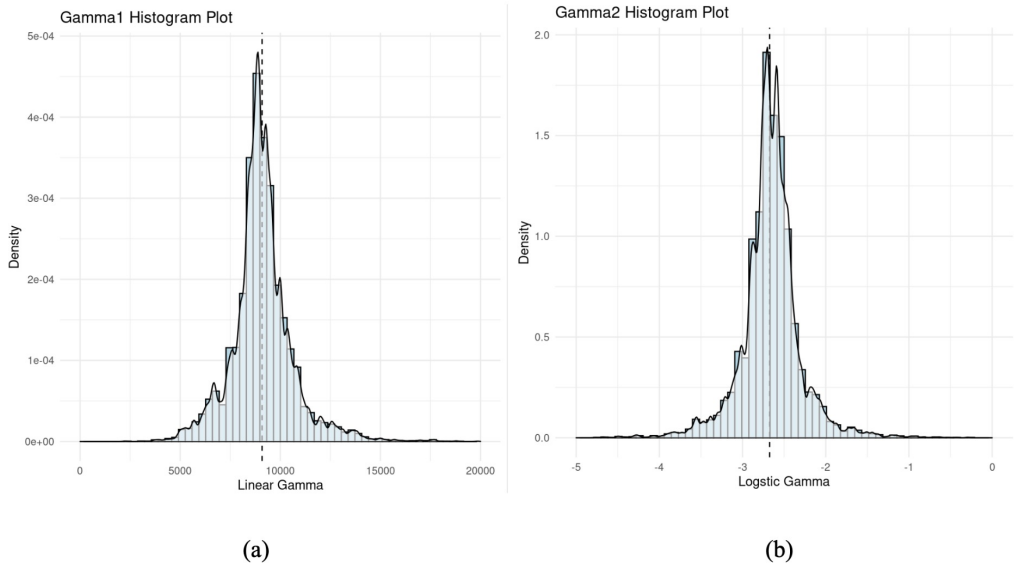

Figure 2: Histogram of Regression Gamma (a) Intercept of Different Provider Institution by Linear Regression; (b) Intercept of Different Provider Institution by Logistic Regression.

The range of gamma is very large, the minimum value is 2243, and the hospital code is 4500UA, which requires the least premium; the maximum value is 27839, and the hospital code is 0500MQ, which requires the most premium. This fully demonstrates the necessity of obtaining different intercepts according to different provider Institutions, and then performing regression.

Table 2: Beta value for linear and logistic model

Variables	Beta for linear model	Beta for logistic model
Sex	69.4099	0.0142
Race	48.8480	0.0123
End stage renal disease Indicator	850.4910	0.2089
Alzheimer	-723.5360	0.3488
Heart Failure	165.3763	0.5272
Chronic Kidney Disease	1409.0601	0.4740
Cancer	317.7563	0.2901
Chronic Obstructive Pulmonary Disease	-344.9652	0.5665
Depression	-711.7975	0.3179
Diabetes	-573.7083	0.4252
Ischemic Heart Disease	461.0560	0.5752
Osteoporosis	-233.0594	0.1107
rheumatoid arthritis and osteoarthritis (RA/OA)	-25.0572	0.2161
Stroke/transient Ischemic Attack	351.9300	0.4113
Age	1.8905	-0.0011

Chronic Kidney Disease has the largest beta for linear model, which requires the most expenditure, and

the most reimbursed by medical insurance. Some other diseases, such as end stage renal disease and heart disease, can also cost more money, while diseases such as depression and Alzheimer dramatically lead to less claim amount. When it comes to beta for logistic regression, we find that all the diseases lead to higher probability of readmission and reclaim of Medicare. Among them, heart failure and chronic obstructive pulmonary disease are the most dangerous ones. However, for age, we find that older patients tend to have less probability of reclaim, this might because they have higher death rate compared to younger group.

The smaller the gamma value, the greater the probability and the lowest disease recurrence rate, which means the better level of the hospital. The larger the beta value, the higher the probability of disease recurrence and the lower the level of the hospital. From the results, it can be concluded that the hospital code-named 0506CA has the highest technical level, while the hospital code-named 2313NU has a lower level.

Discussion & Future Study

Before conducting our project, we thought that providers effect should be a significant factor when it comes to Medicare claim. Our result somewhat prove that if only considering the training error, however, when it comes to test error, we found that our results didn't show too much advantage compared to model without considering providers effect. After carefully analysis, we think this might be caused by our restriction of fixed effect. A more proper and rigorous way is to consider the providers' effect as random effect, which is also presented in He, Kalbfleisch, Li & Li (2013). This can be the future direction of our study. What's more, our project is also limited by the risk factors that we used. Most of the risk factors are categorical variables, as the data source could not provide us with enough continuous variables. We doubt that this might affect our model performance. Finding better relevant data source is another direction of our future study.

Contribution:

Xiangeng Fang contributed on data preprocessing, model construction and fitting. He also wrote most of code. Xinjue Li showed his effort on background and data visualization, and Yixuan Zeng worked on model evaluation and results.

Reference

1. He, K., Kalbfleisch, J. D., Li, Y., & Li, Y. (2013). Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. Lifetime Data Analysis, 19(4), 490-512.
2. Hossain, M. E., Khan, A., Moni, M. A., & Uddin, S. (2019). Use of electronic health data for disease prediction: A comprehensive literature review. IEEE/ACM transactions on computational biology and bioinformatics, 18(2), 745-758.
3. Xie, Y., Schreier, G., Chang, D. C., Neubauer, S., Liu, Y., Redmond, S. J., & Lovell, N. H. (2015). Predicting days in hospital using health insurance claims. IEEE journal of biomedical and health informatics, 19(4), 1224-1233.