STA 304

Final Project Report

Prediction on Readmission to Hospital using a Popular Diabetes Dataset

Yijia Liu

1. Introduction

Diabetes is a life long condition where your body does not produce enough insulin or your body cannot use the insulin it has effectively. It can increase the risk of high blood pressure, narrowing of the arteries (atherosclerosis), coronary artery disease and stroke. Thus, it is important to identify the patients with worse outcomes.

In general, it is hard to measure directly. However, readmissions can provide a low cost, stable estimate. If the readmissions are none, which means the treatment is great; else if the readmissions are less than 30 days, which means the treatment may not be appropriate; else if the readmissions are more than 30 days, which means the treatment is not very good, but the reason could be the patients.

2. Methods section

2.1 Constructing Regression Model

Model1: GLM with Binomial Response

When the response variable has only two outcomes, which follows a binomial distribution Bin (m_i, π_i), we can use

$$\mathrm{P}(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

to express. We further assume that the Y_i are independent. The individual trials that compose Y_i are subject to the same q predictors $(x_{i1}, \dots, x_{iq})$.

As in the binary case, we construct a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$

We can use a logistic link function $\eta_i = log(\pi_i /(1 - \pi_i))$ and the log-likelihood is given by

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \eta_i - m_i \log(1 + e^{\eta_i}) + \log \binom{m_i}{y_i} \right]$$

The deviance is given by

$$D = 2 \sum_{i=1}^{n} \left[\frac{y_i \log(y_i)}{\hat{y}_i} + \frac{(m_i - y_i) \log(m_i - y_i)}{m_i - \hat{y}_i} \right]$$

Model2: Generalized Linear Mixed Models (GLMMs)

After considering random effects, we can use generalized linear mixed models. The response follows an exponential family distribution,

$$f(y_i \mid \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi)\right\}$$

Let $\mathbb{E}(Y_i) = \mu_i$ and we connect it to the linear predictor η_i using the link function g by $\eta_i = g(\mu_i)$. If the random effect γ have distribution $h(\gamma \mid V)$ for parameters V. The fixed effects are β. Condition on the random effects γ, we have $\theta_i = x_i^\top + z_i^\top \gamma$. Then the likelihood function is

$$L(\beta, \phi, V \mid y) = \prod_{i=1}^{n} \int f(y_i \mid \beta, \phi, \gamma) h(\gamma \mid V) d\gamma$$

Where $\gamma \sim N(0, D)$

2.2 Variable Selection

Some variables in original model are redundant or there exist Multicollinearity. So I develop some variable selection method:

(a) AIC and BIC Criterion

AIC and BIC are Information criteria methods used to assess model fit while penalizing the number of estimated parameters. Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following.

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

The formula for (BIC) is similar to the formula for AIC, but with a different penalty for the number of parameters.

$$\text{BIC} = \ln(n)k - 2\ln(\text{L})$$

(b) Stepwise selection

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure.

Forward selection: Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one.

Backward selection: The backward selection model starts with all candidate variables in the model. At each step, the variable that is the least significant is removed. We applied backward selection model in my project.

2.3 Model Validation/ Dignostics

After selecting variables, we need to check the following assumptions:

(a) Normality of residuals; (QQ-plot)

(b) Homoscedasticity; (Fitted Values v.s. Residuals)

(c) Independence (Ljung-Box test)

3. Results section

3.1 Description of Data

The data contains 101766 observations from 71518 patients. It includes over 50 features representing patient and hospital outcomes, where some are numerical and some are categorical. To clean the data and keep the most important variables, we will do the following steps:

1) Transform the response variable “readmitted” to be binary: combine “”less than 30 (total 11357)” and “over 30 (total 35545)” to “1”, which means the patient does have readmission; change “no” to “0 (54864)”, means there is no readmission. This transformation also helps keep the data balanced.
2) Remove some seriously imbalanced or NA variables: Since for “weights”, “payer_code” and “medical_specialty” contain almost all NAs, we will remove them. Also, “examide”, “metformin_rosiglitazone” and “citoglipton” are removed due to the seriously imbalanced.
3) Remove other rows which contains NAs.
4) Correlation Matrix: For all rest numerical variables, we can construct the correlation matrix as following:

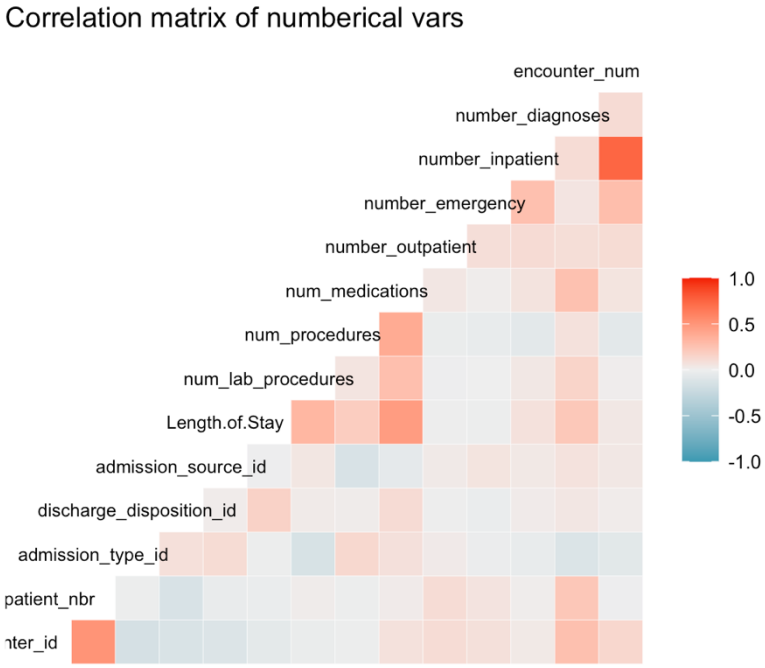

Figure 1: Correlation Matrix

From the figure above, we can find “num_medications” has relatively strong relationship with several variables and “number_inpatient” has strong relationship with “encounter_num”.

3.2 Process of Obtaining Final Model

(a) **Split the data to train set and test set**: I create a test dataset that contains a random selection of 20000 patients by using 'patient nbr' variable. The rest are set as train set.

(b) **Remove other variables which cannot be used as covariates**: I continue to remove "encounter_id", "patient_nbr", "admission_source_id", "encounter_num" before fitting models.

(c) **Fitting GLM with Binomial Response:** Since the response variable "readmitted" is a binary variable, we can let $Y_i \sim \text{Bernoulli}(\pi_i)$, where π_i is the probability of readmission. And then we fit a GLM as $\text{logit}(\pi_i) = X_i\beta + e_i$, where $e_i \sim N(0, \sigma^2)$.

The fitted formula for model1 is:
formula = readmitted ~ race + gender + age + num_lab_procedures + num_procedures + num_medications + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + A1Cresult + metformin + glipizide + pioglitazone + rosiglitazone + acarbose + insulin + change + diabetesMed + Length.of.Stay (totally 21 variables)

(d) **Backward Stepwise Selection:** From the regression results above, I find half of the covariates are not significant. So I decide to use backward stepwise selection method to find more proper and simpler model but with reasonable explanation.

The final result is:
readmitted ~ race + gender + age + num_lab_procedures + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + A1Cresult + metformin + glipizide + rosiglitazone + acarbose + insulin + diabetesMed + Length.of.Stay (totally 18 variables)

(e) **Generalized Linear Mixed Model**: After selecting variables using Stepwise, we still find some of the variables that are not significant. Since one patient may have more than once encounters, the generalized linear mixed model condition on "patient_nbr" maybe a better choice.
We can let $Y_{ij} \mid U_i \sim \text{Bernoulli}\ (\pi_{it})$ where π_{it} is the probability of

readmission. And then we fit a GLMM as $\text{logit}(\pi_{it}) = X_{it}\beta + U_i$.
The fitted formula for model1 is:
Formula: readmitted ~ (1 | patient_nbr) + race + gender + age + num_lab_procedures + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + A1Cresult + metformin + glipizide + rosiglitazone + acarbose + insulin + diabetesMed + Length.of.Stay

From the regression results, we continue to remove the variables are not significant, the final model is:
final_formula = readmitted ~ (1 | patient_nbr) + race + gender + age + num_procedures + number_outpatient + number_emergency + number_diagnoses + insulin + diabetesMed + Length.of.Stay

3.3 Goodness of Final Model

(a) Normality of residuals; (QQ-plot)

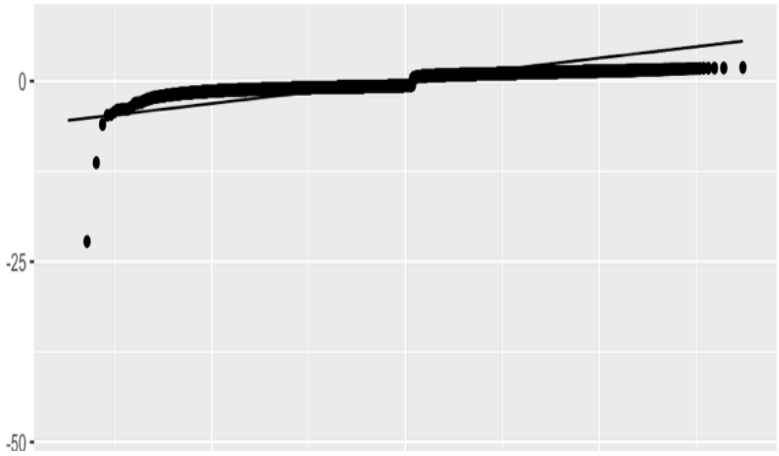

Figure 2: QQ-Plot

From the QQ plot above, we can find most of the points are on the straight line, which means the residuals follow normal distribution approximately.

(b) Homoscedasticity; (Fitted Values v.s. Residuals)

From the figure below, we can find the points have some trends even if they are on the two sides around zero. Thus, it is not very confident to state the variance is constant, which also means there may not exist homoscedasticity.

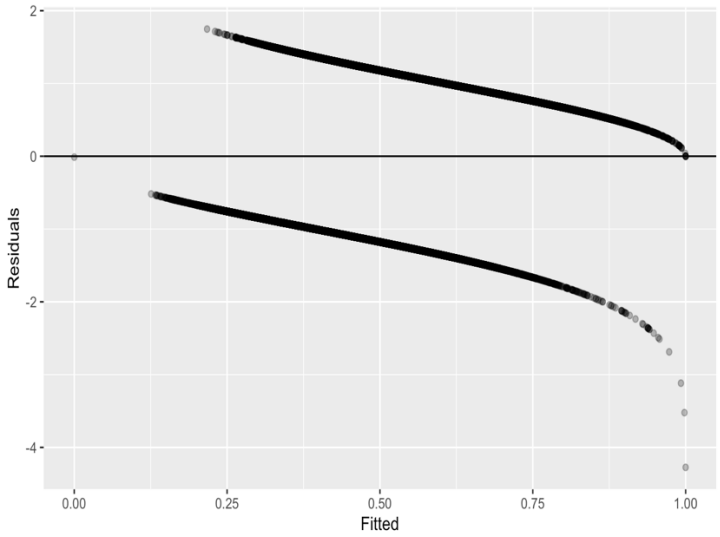

Figure 3: Fitted Values v.s. Residuals

(c) Independence; (Ljung-Box test)

From the R output, we can find the p-value of Ljung-Box test is 7.03e-08, which means we have to reject null hypothesis and not all of the variables are independent.

4. Discussion section

4.1 Final Model Interpretation and Importance

1) Finally, we build the GLMM model condition on "patient_nbr", totally 10 covariates, which is:
final_formula = readmitted ~ (1 | patient_nbr) + race + gender + age + num_procedures + number_outpatient + number_emergency + number_diagnoses + insulin + diabetesMed + Length.of.Stay

2) After fitting this model on train set, we use it on test set. The AUC is 0.8854 and the ROC curve is as below. From the accuracy, we can find the predicted results are acceptable and even good. Thus, the model maybe a suitable model for the data.

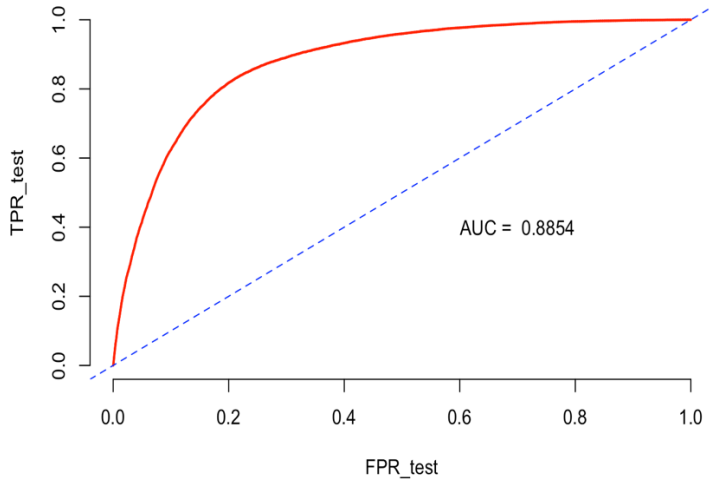

Figure 4: ROC for Test Set

4.2 Limitations of Analysis

There are some problems about stepwise selection method:

(a) It yields R-squared values that are badly biased to be high.
(b) The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
(c) It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.
(d) It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.
(e) The stepwise selection allows us to think too much about statistical model but not the original problem.

In addition, I didn't consider the interaction terms which may have better results. Also, it is only a multiple linear model. The fitted efficiency may be more accurate if we add non-linearity elements in our model.