Forecasting Hospital Readmission of Diabetics
Project Report, STA303: Methods of Data Analysis II
Department of Statistics, University of Toronto

Huifeng Wu*
Instructor: George Stefan

August 30, 2020

1. Introduction

In 2014, Clore et al. collected a novel dataset describing a large volume of medical records from over 70,000 diabetics. Among extracted attributes, readmission with certain timestamps was used to measure health status after initial visits, as well as the cost of inpatient cure from patients. In clinical research, it is of great significance to accurately keep track of patients who are probabilistically subject to hospital readmission. Successful classification does not only prioritize patients at higher risk with more thorough diagnosis and considerate treatment, but it also presents advantages in cost reduction and budget planning for the healthcare system. Along the previous literature, this report proposes a significantly interpretive statistical model of logistic regression modeling the readmission rate of diabetics. The final model was derived through a comprehensive criterion covering variable selection, model assumption, diagnostics, and validation.

2. Methods

2.1 Choice of Methods

Hospital readmission was originally measured in three scales: readmission within 30 days, readmission in more than 30 days and no readmission. Late readmission (>30) highlights one's poor state of health, while early readmission (<30) could be a result of inappropriate treatment as the occurrence of medical incidents are inevitable. Nevertheless, they were converted to a generic level, and the variable was dichotomized to a binary response, with "readmission" and "no readmission".

A generalized linear model of logistic regression was used to fit the binary data. This approach requires the assumption that each observation is independent; duplicate observations were removed from the original dataset. Besides large sample needed and low multicollinearity between predictors, logistic regression assumes linearity between logit outcome and its predictors as follows,

$$log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = X_i \cdot \beta$$

where $\hat{p}_i$ is the predicted probability of hospital readmission for a patient, X_i denotes one's observed features, and β indicates the intercept estimate and coefficients estimates for their respective covariates.

2.2 Variable Selection

In addition to the response, the dataset contains demographic information of patients and their medical records such as IDs, laboratory tests, administered medications and length of initial stay. Rigorous data reconstruction was made towards the underlying dataset, reducing the number of variables from 49 to 14 for preliminary analysis (Appendix-A).

I eliminated variables of identification purpose, such as encounter and admission source ID. Variables with large proportion of missing values or ones with constant values were empirically excluded, since they did not properly explain the outcomes. Besides, imputation for missing values might even increase bias.

When dealing with variables with multiple levels, such as discharge disposition, A1C result, race, and age, they were preferably merged into fewer categories or transformed to continuous variable for convenience. For the set of 20 medications appeared in the dataset, a variable was created representing the number of medications prescribed. Meanwhile, the original variable “num_medications” was replaced for redundancy and miscount. Such data aggregation was also applied to take account of number of visits and procedures.

2.3 Model Diagnostics

The final dataset was split into train and test set by random selection, where the former was used to fit a full model, and the latter of 20,000 records evaluated prediction performance.

T-tests were preformed to model coefficients for regression significance. Using 5% significance level, it concluded whether individual predictor has a significant effect on patient’s readmission within the full model. Nevertheless, to resolve potential multicollinearity and investigate a set of significant covariates for model inclusion, a model selection procedure featuring AIC/BIC backward elimination and LASSO regression was performed. Deviance tests were conducted between each two candidate models. If null hypothesis is rejected, a more complicated model is preferred.

Once the final model was chosen, its goodness of fit was evaluated using Hosmer-Lemeshow test. Besides, the linearity between the logit value and important predictors was graphically examined using scatter plots. Deviance residual plots against linear predictor and fitted values assessed the quality of modal fitting, in which no obvious pattern was expected to be found. The variation inflation factors (VIF) were used to check the absence of multicollinearity; values exceeding 5 would typically compromise the assumption. If any violations above occur, different variable selection, model fitting and data transformation may be attempted. Followed by drawing the Receiver Operating Characteristic (ROC) curve on the test set, the greater area of under the curve (AUC) presents better discriminative ability in prediction.

3. Results

3.1 Data Description

From an exploratory perspective, the final dataset comprises of 69569 observations from distinct patients, where the response readmission is of imbalance distribution. The average readmission rate is 40.2%. Figure 1 illustrates histogram plots with numerical predictors, where most of them show skewed distribution in both “readmission” and “no readmission”. Age approximates a normal distribution, but spikes are evident on left end. There are high leverage points detected for number of diagnosis, visits, and medications. For categorical variables, Table 1 summaries the number of observations with respective levels and their proportional readmission rates, as well as p-values from chi-square tests for independence. In brief, all of them are significantly correlated with the response.

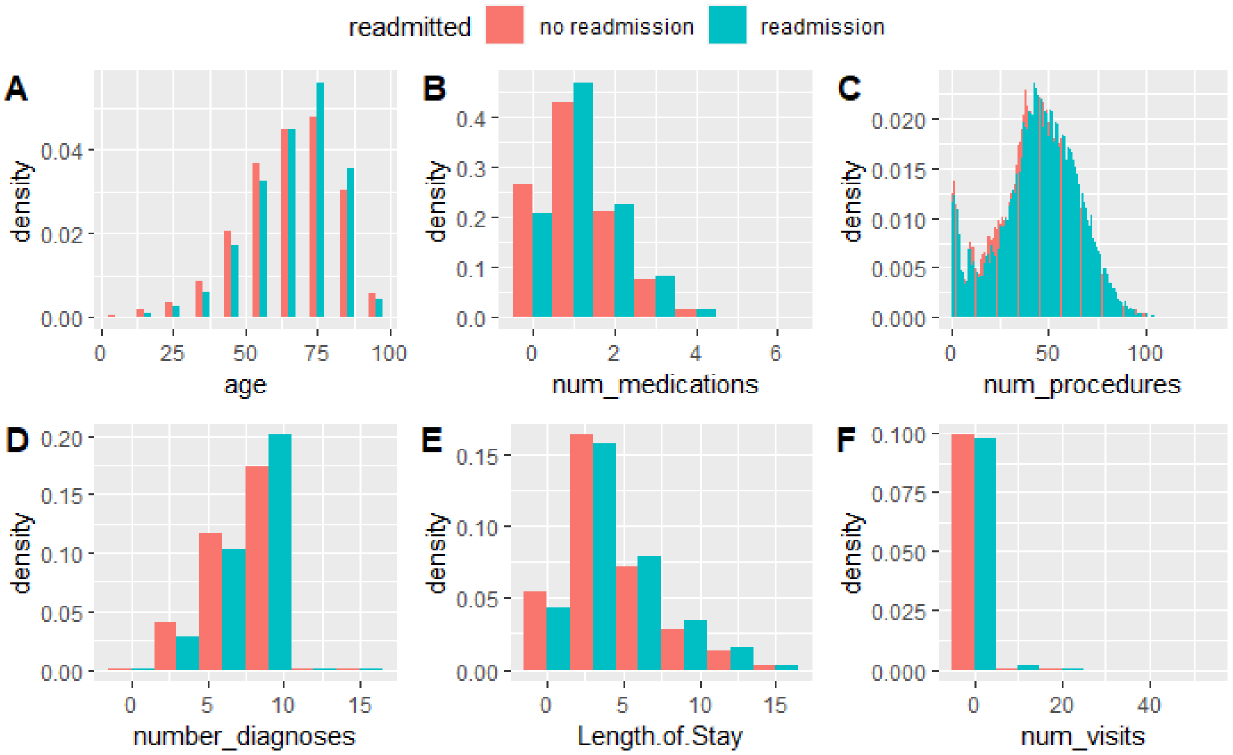


Figure 1. Histogram of numerical covariates, with respect to readmission events

Variable	Level	Count	Readmission rate	p-value
race	Caucasian	53491	40.98%	≈0
	Other	16078	37.54%	
gender	Female	37044	40.96%	≈0
	Male	32525	39.30%	
discharge_disposition_id	Home	43031	39.08%	≈0
	Other	26538	41.97%	
A1Cresult	High	8886	39.38%	≈0
	None	56966	40.57%	
	Norm	3717	36.19%	
change	No	31096	42.34%	≈0
	Yes	38473	38.45%	
diabetesMed	No	16828	34.54%	≈0
	Yes	52741	41.99%	

Table 1. Summary of categorical variables including levels, counts and proportional readmission rate and p-values from chi-square tests

3.2 Process of Obtaining Final Model

A full model was initially fitted with the train set. For “Other” type for discharge disposition (p=0.11), number of medications (p=0.053), change of diabetic medication (p=0.371) and “None” type for A1C result (p=0.0877), their p-values are greater than 5% significance level. Thus, individual one has no significant effects on patient readmission in this model and may be excluded later.

Table 2 describes the coefficients estimates and related statistics for the full model and ones selected by different approaches. Methods of LASSO and AIC share very similar results, so AIC’s was used to compare later. BIC indicated the simplest one with 5 covariates. To compare these candidate models, separate deviance tests were preformed. The result (p≈0) indicates an significant improvement from model 3 to model 2, but not statistical better model fitting (p=0.372) when comparing model 2 and 1. Even though the coefficients estimates seem consistent across all models, model 2 (AIC/LASSO) was chosen as the final model:

$$\begin{aligned}\log\left(\frac{\hat{\mathrm{p}}_{\mathrm{i}}}{1-\hat{\mathrm{p}}_{\mathrm{i}}}\right) = {} & 1.7873 - 0.0522 * \mathrm{raceOther} - 0.049 * \mathrm{genderMale} + 0.0054 * \mathrm{Age} \\ & - 0.0326 * \mathrm{discharge\_disposition\_idOther} + 0.0220 * \mathrm{Length.of.Stay} \\ & + 0.0018 * \mathrm{num\_procedures} - 0.0245 * \mathrm{num\_medications} + 0.0490 \\ & * \mathrm{A1CresultNone} - 0.1317 * \mathrm{A1CresultNorm} + 0.3471 * \mathrm{diabetesMedYes} \\ & + \ 0.1996 * \mathrm{num\_visits}\end{aligned}$$

Method	**Covariates**	**coefficients estimates**	**p-value**
Model 1 (Full)	(Intercept)	-1.7634	0.0000
	raceOther	-0.0525	0.0209
	genderMale	-0.0496	0.0083
	age	0.0055	0.0000
	discharge_disposition_idOther	-0.0332	0.1107
	Length.of.Stay	0.0219	0.0000
	num_procedures	0.0018	0.0006
	num_medications	-0.0340	0.0529
	number_diagnoses	0.0716	0.0000
	A1CresultNone	0.0500	0.0878
	A1CresultNorm	-0.1308	0.0074
	changeNo	-0.0244	0.3720
	diabetesMedYes	0.3477	0.0000
	num_visits	0.1967	0.0000
Model 2 (AIC/LASSO)	(Intercept)	1.7873	0.0000
	raceOther	-0.0522	0.0215
	genderMale	-0.0497	0.0082
	age	0.0054	0.0000
	discharge_disposition_idOther	-0.0326	0.1165
	Length.of.Stay	0.0220	0.0000
	num_procedures	0.0018	0.0006
	num_medications	-0.0245	0.0794
	number_diagnoses	0.0718	0.0000
	A1CresultNone	0.0490	0.0940
	A1CresultNorm	-0.1317	0.0070
	diabetesMedYes	0.3471	0.0000
	num_visits	0.1968	0.0000
Model 3 (BIC)	(Intercept)	-1.7498	0.0000
	age	0.0056	0.0000
	Length.of.Stay	0.0232	0.0000
	number_diagnoses	0.0730	0.0000
	diabetesMedYes	0.3087	0.0000
	num_visits	0.1996	0.0000

Table 2. Summary table for model selected by AIC/BIC and full model

3.3 Goodness of Final Model

To check model linearity, Appendix B is a faucet of scatter plots between numerical covariates and logit outcome; most of which present signs of linear relationships, so no additional transformations were made. For nominal variables, it is difficult to visualize their dependence with readmission, but chi-square tests have previously validated them and all show significance in Table 1. Appendix C confirmed no signs of heavy multicollinearity by taking advantage of model selection; all VIF values are less than 5.

Alongside 95% confidence intervals involving the binomial variance, Figure 2.1 illustrates no consistent deviation between observed proportion and predicted probability.

No patterns were found in deviance residual plots in Figure 2.2 and 2.3; residuals are distributed evenly. The half-normal plot also detects no outlying points. However, when formalizing the Hosmer-Lemeshow test, I concluded the model lacked fitting (p≈0).

In terms of predictive power, the final model yielded a discriminative AUC value of 61% on test set in Figure 3. This implies in 61% of the times, it can correctly discriminate whether a patient experiences readmission or not.

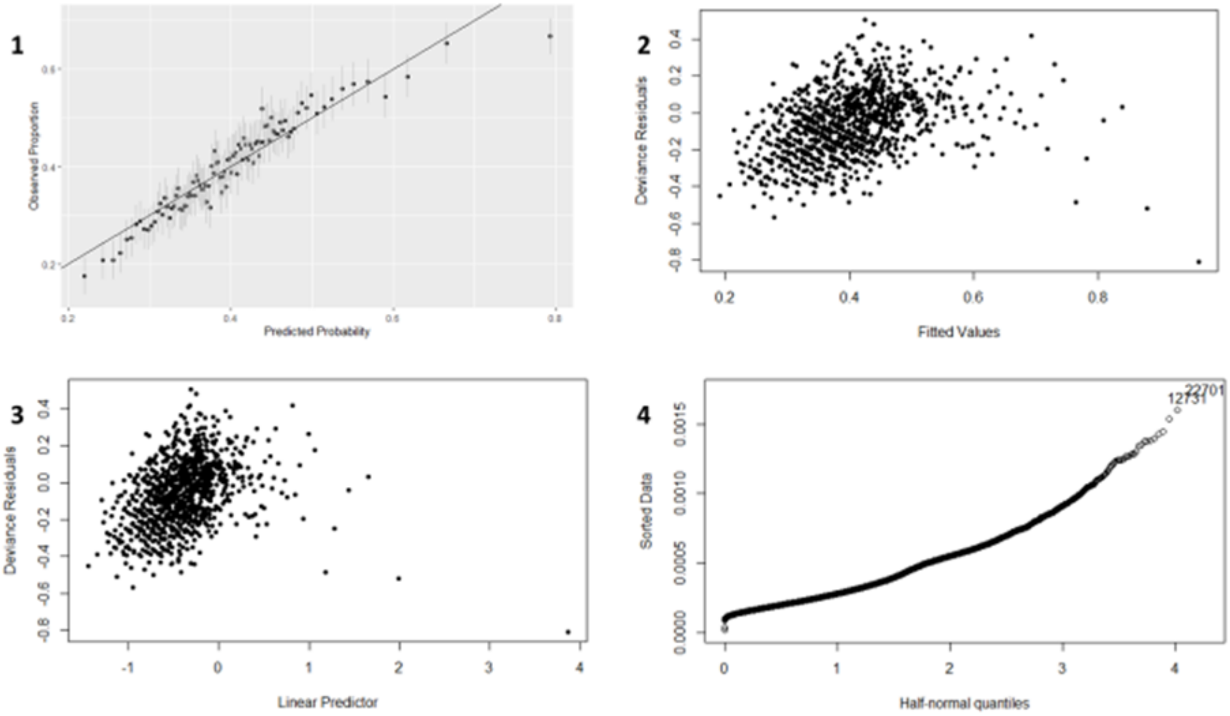


Figure 2. Model diagnosis: (1) observed proportion versus predicted probability, (2)-(3) deviance residual plots against linear predictor and fitted values and (4) a half normal plot

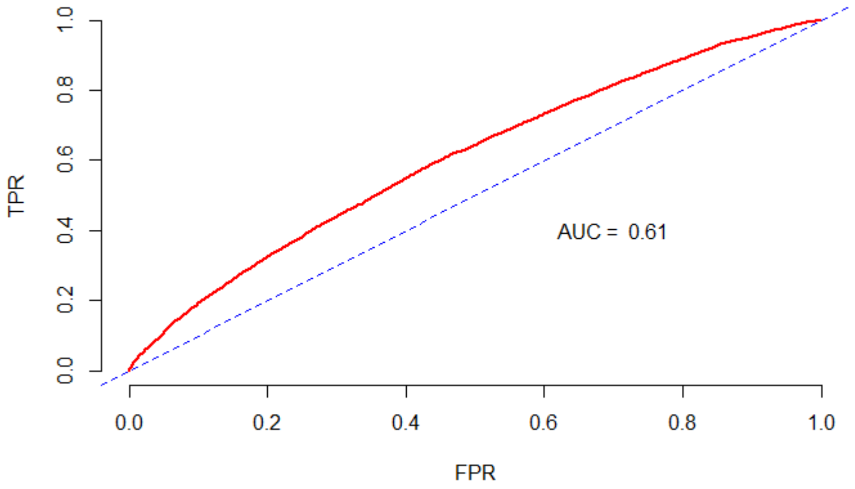


Figure 3. ROC curve on test data set

4. Discussion

4.1 Final Model Interpretation and Importance

To interpret, the intercept of 1.7873 is the log-odds of readmission contributed when considering reference group, a Caucasian female patient who is sent home after initial visit, and he/she has high A1C result and no diabetic medication prescribed. Moreover, the increase in log-odds of readmission for one more visit to hospital is 0.1968, and the increase in log-odds of readmission is 0.3471 if one takes diabetic medicine, while keeping other variables constant. Similar interpretation goes beyond other predictors.

The final model is interpretive and useful to explain the readmission events in the dataset, which does not only include the scenario when diabetics develop worse medical outcomes, but also inappropriate treatment from initial visits. With this model, it will reduce financial burden of healthcare system in better budget management, and practitioners will pay attention to those patients at higher risks and make intervention if possible.

4.2 Limitations

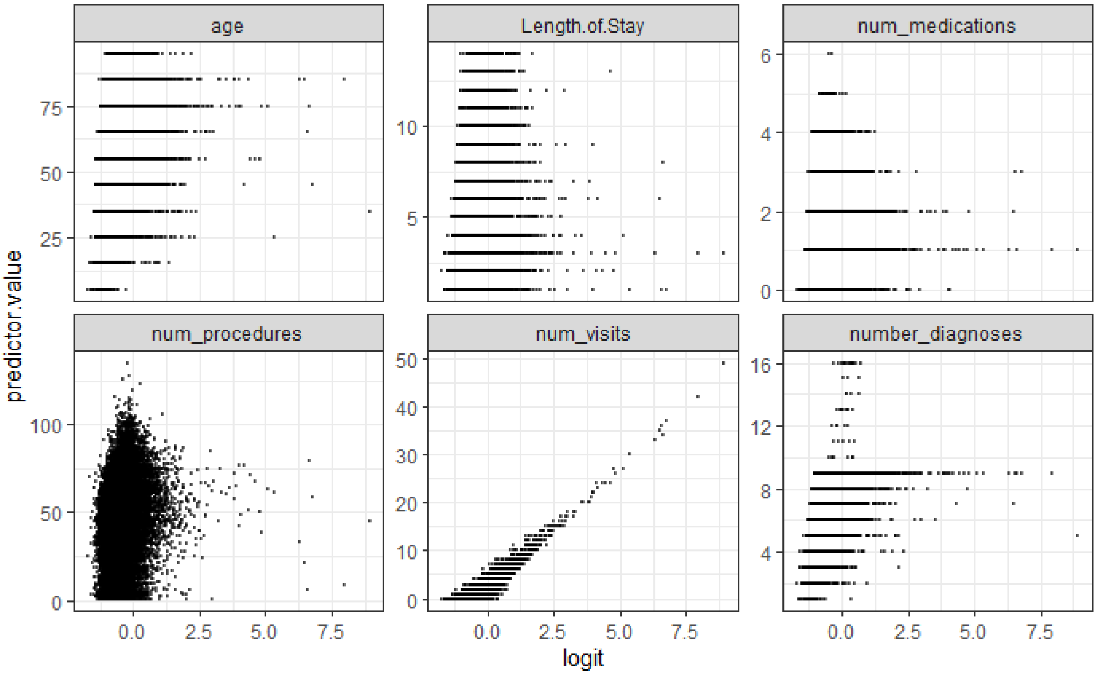
Given the longitudinal data, original observations are dependent, yet only patients' first records were used in model fitting, and they may be less representative for readmission events. A considerable portion of data were removed directly. Moreover, the model may be inadequate to predict on new data from the same patient since random effects are not considered.

The final model lacks fitting, which may be caused by data processing. For instance, readmission reflects upon either the poor health state or inadequate care from last visit; it is difficult to conclude the exact reason. These subcategories may lead to different regression models in fact, but they were merged in this report. Besides, no interaction effects were contained in modelling, which could be problematic of capturing correct effects. In contrast, there may exist confounding variables as well, which biases the estimation. Above problems discount the usefulness of final model. To tackle this classification problem, GLMM, neural network and other techniques are worth of investigation. Instead of a single test set, 5-folds cross validation may better evaluate models.

Appendix

Original	New	Transformation	Reason
gender (male, female, unknown/invalid)	gender (male, female)	removed unknown or invalid observations	invalid data
race (African American, Asian, Caucasian, Hispanic, Other, NA's)	race (Caucasian, Other)	dichotomization	Caucasian (75%) and low percentages for others
age (nominal)	age (continuous)	used midpoints of original intervals	convenience
A1Cresult (None, Normal, >7, >8)	A1Cresult (None, Normal, High)	few categories	convenience
discharge disposition (28 levels)	discharge disposition (Home, Other)	few categories	Home (59%)
20 medications	num_medications	sum of medications prescribed	too many dummy variables if fitted
number of procedures number of lab procedures	number of procedures	Sum of procedures and lab test procedures	convenience
number of inpatient visits number of outpatient visits number of emergency visits	number of visits	sum of these visits	Convenience
weight examide citoglipton max_glu_serum admission_type_id medical_specialty encounter_id admission_source_id payer_code encounter_num	null	removed	too many missing values, constant value variable or identification purpose

Appendix A. Summary of restructured covariates in the final dataset



Appendix B. Scatter plots using the final model between numerical predictors and the predicted log outcomes

Variable	VIF	Df
race	1.0425	1
gender	1.0130	1
age	1.1795	1
discharge_disposition_id	1.1769	1
Length.of.Stay	1.2665	1
num_procedures	1.2253	1
num_medications	1.9654	1
number_diagnoses	1.1319	1
A1Cresult	1.1133	2
diabetesMed	1.9567	1
num_visits	1.0150	1

Appendix C. VIF values on the final model

Reference

1. Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J., & Clore, J. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014.
2. Schreiber-Gregory, D., Jackson, H.M., & Bader, K.S. (2018). Logistic and Linear Regression Assumptions: Violation Recognition and Control.
3. http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/#logistic-regression-diagnostics
4. https://clinical.diabetesjournals.org/content/24/1/9.full-text.pdf