STA303 Final Report

Predicting Diabetic Patients' Hospital Readmission with Statistical Model

Wei Cui 1004536479

August 30, 2020

1 Introduction

Diabetes is a widely distributed chronic disease that is characterised by blood glucose levels abnormalities due to insulin-related problems. The prevalence of diabetes is growing most rapidly in low- and middle-income countries [1].

Readmission to hospital is defined by the period a patient is taking before returning to the hospital. Readmission is considered a indicator of hospital success in terms of efficiency and a way of reducing healthcare costs. Hospitals are financially penalised if the permitted 30-day readmissions rate is exceeded. The United States Medicare Payment Advisory Commission evaluated that preventing 10% of readmissions will save more than $1 billion for Medicare in the US [2].

Therefore, in order to not only improve the quality of healthcare but also reduce the medical expenses on readmission, it's important for us to create interventions to provide additional assistance to diabetic patients with increased risk of readmission. Thus, our goal of the study focuses on coming up with a statistical model to predict if a patient with diabetes will be readmitted to the hospital within 30 days.

2 Methods Section

(i) Choice of Methods

We choose **logistic regression (GLM)** to be our model. **Logistic regression** is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). After processing the diabetes dataset, we dichotomize our response variable **readmission** as **no readmission** and **readmission**, which makes our response variable become binary and satisfy the requirement of the **logistic regression (GLM)**. **Logistic regression**, like all regression analyses, is a predictive analysis and is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Since we want to analyze the relationship between **readmission** and different covariates and to predict diabetic patients' hospital readmission, it's appropriate to use **logistic regression (GLM)** as our statistical model.

(ii) Variable Selection

Firstly, In data preparation and preprocessing, for those identification variables, we directly drop them as specified in our handout. If we find that there are variables containing large proportion of missing values, then we decide to remove them. There are also some variables with only 1 factor/value meaning that these variables have zero variance and we should not take them as predictors. Finally, to reduce number of covariates, we combine and recode several variables into few more meaningful and important new variables.

Secondly, after finishing data preparation and preprocessing, we perform model selection procedure based on **stepwise methods** (both **AIC** and **BIC**). We notice that the model selected by **AIC** and **BIC** is different, they have a different set of covariates, so we use `drop1` command to test for each different covariate's significance to decide whether we keep it or not.

(iii) Model Violations/Diagnostics

We perform model violation checks, diagnostics and how each will be handled as follow:

- We plot the deviance residuals versus the fitted values and the linear predictor. If the plot reveals curvature, it suggests that one or more important covariates do not influence the log-odds of success in a linear fashion. Thus we will do some reasonable transformation on those covariates.

- The `QQ-plot` allows us to check if the standardized residuals follow a $\boldsymbol{N}(0, 1)$. We will display a `QQ-plot` and if there are some departures from the diagonal line, it is fine. The reason is simply that the deviance residuals are significantly non-normal, which happens often in logistic regression.

- Detection of outliers and influential cases and corresponding treatment is very crucial. We will detect unusual observations by examining the leverages and use a `half-normal` plot for this. If there are some particularly extreme points in the plot, we should drop them.

3 Results Section

(i) Description of Data

The `diabetes` dataset contains 101,766 encounters, and there are 48 features describing the diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information [3].

The original dataset contains incomplete, redundant, and noisy information as expected in any real-world data. Especially, `weight` has approximate 98% missing values and is considered to be too sparse, so it isn't included in further analysis.

Hemoglobin A1c (`HbA1c`) is an important measure of glucose control, which is widely applied to measure performance of diabetes care [4] [5]. Therefore, `A1Cresult` is one of the important variables that we need to pay attention to. `A1Cresult` is a nominal variable and greater than 8%, `">7"` if the result was greater than 7% but less than 8%, `"normal"` if the result was less than 7%, and `"none"` if not measured.

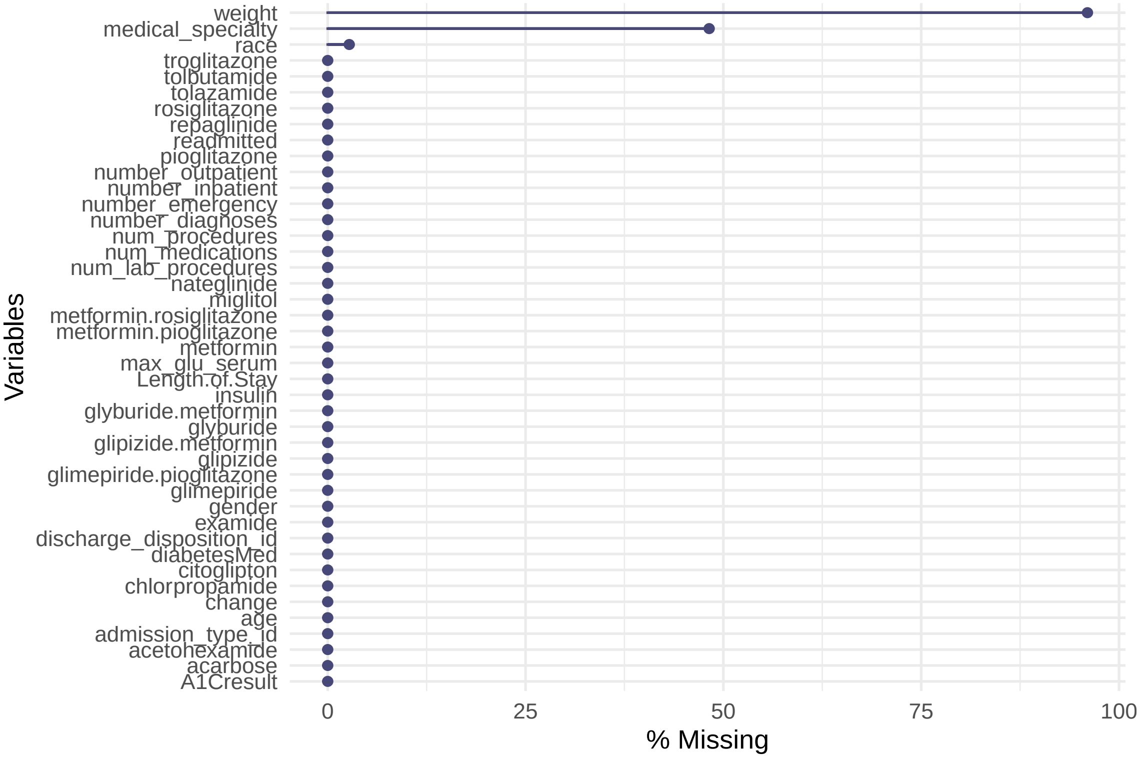

Figure 1: Missing percentage of each variables in the `diabetes` dataset

(ii) Process of Obtaining Final Model

In the data preparation and preprocessing section, we take actions on those variables as follow:

Table 1: Variable Selection Part 1: Data preprocessing

Variables	Modifications & Results
`encounter_num`	Since we remove all duplicate patients' encounters. For variable `encounter_num`, it only has 1 value, i.e. 1. Therefore, this variable has zero variance and thus we drop this variable.
`encounter_id`	identification variable, thus we drop it
`admission_source_id`	identification variable, thus we drop it
`patient_nbr`	identification variable, thus we drop it
`payer_code`	identification variable, thus we drop it
`weight`	Variable `weight` contains approximate 98% of the missing values, thus we decide to drop this variable.
`examide`	Examide only has 1 value, so we will not use this variable as predictor
`citoglipton`	citoglipton only has 1 value, so we will not use this variable as predictor
`21 remaining features for medications`	We remove these drug variables and based on these variables, we add 2 additional features: *num_of_med* and *num_of_changes*
`num_lab_procedures`	regroup *num_lab_procedures* and *num_procedures* into a single variable *num_procedures*
`num_procedures`	regroup *num_lab_procedures* and *num_procedures* into a single variable *num_procedures*
`number_outpatient`	Regroup and recode variables *number_outpatient*, *number_emergency* and *number_inpatient* into a new variable called *num_visits*
`number_emergency`	Regroup and recode variables *number_outpatient*, *number_emergency* and *number_inpatient* into a new variable called *num_visits*
`number_inpatient`	Regroup and recode variables *number_outpatient*, *number_emergency* and *number_inpatient* into a new variable called *num_visits*
`A1Cresult`	Combine with variable `change` and recode to `HbA1c`
`change`	Combine with variable `A1Cresult` and recode to `HbA1c`

And the results of `drop1` command:

Table 2: Variable Selection Part 2: `results of drop1 commands`

Variables	P-value from the `drop1` command
`admission_type_id`	0.03577
`max_glu_serum`	0.07253
`HbA1c`	0.01622
`medical_specialty`	1.501e-05
`num_procedures`	5.867e-12
`num_of_med`	0.01821
`num_of_changes`	1.716e-06

Combine the results of model selection procedure based on **stepwise methods** and above `drop1` commands . We decide to choose the following covariates: (1) `age`, (2) `admission_type_id` (3) `Discharge`, (4) `Length.of.Stay`, (5) `medical_specialty`, (6) `num_procedures`, (7) `number_diagnoses`, (8) `diabetesMed`, (9) `num_of_med`, (10) `num_of_changes`, (11) `num_visits`, (12) `HbA1c` to fit our final **logistic regression** model.

Table 3: Coefficients of noninteraction terms estimated from the final logistic regression model.

		Estimate	P-value
	(Intercept)	-3.623e+00	<2e-16
age	age[30, 60)	reference	
	age[60, 100)	1.993e-01	2.00e-07
	age<30	7.227e-02	0.558492
admission_type_id	1	reference	
	2	4.220e-02	0.359790
	3	3.507e-02	0.456236
	4	-1.012e+01	0.933601
	5	-1.935e-01	0.021701
	6	6.983e-02	0.284110
	7	-1.034e+01	0.914596
	8	2.603e-02	0.923976
Discharge	Home	reference	
	Home with home health service	2.283e-01	1.02e-05

Table 3: Coefficients of noninteraction terms estimated from the final logistic regression model.

		Estimate	P-value
	Other	7.573e-01	<2e-16
	SNF	5.550e-01	<2e-16
Length.of.Stay		2.337e-02	6.46e-05
medical_specialty	Emergency	reference	
	General Practice	2.649e-01	0.003909
	Internal Medicine	2.716e-01	0.000464
	Other	1.852e-02	0.868622
	Surgery	6.415e-02	0.489483
	Missing	1.949e-01	0.008972
num_procedures		1.810e-03	0.052961
number_diagnoses		4.524e-02	1.14e-06
diabetesMed	diabetesMedNo	reference	
	diabetesMedYes	3.351e-01	1.14e-09
num_of_med		-7.339e-02	0.003286
num_of_changes		7.657e-02	0.033274
num_visits		9.134e-02	<2e-16
HbA1c	HbA1c1	reference	
	HbA1c2	-1.243e-01	0.031373
	HbA1c3	1.893e-01	0.066331
	HbA1c4	-1.011e-01	0.176514

(ii) Goodness of Final Model

The **logistic regression** method assumes that:

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
- logistic regression requires the observations to be independent of each other.
- There is no influential values (extreme values or outliers) in the continuous predictors

To perform model violation checks and diagnostics, we do the following:

- We dichotomize our response variable **readmission** as **no readmission** and **readmission**.
- We remove duplicate patients' encounter and only take the first observation to avoid bias and ensure independence of observations
- We detect unusual observations by examining the leverages and use a `half-normal` plot for this.

Beside that, we construct a **binned residual plot** and observe an even variation as the linear predictor and fitted values vary, thus the plots do not detect any inadequacies in the model. And we also plot the binned residuals against the `HbA1c` predictor. We display a `QQ plot` of the residuals, just like we would for linear models. However, there is no reason to expect these residuals to be normally distributed, so this is fine. From the `half-normal` plot, there are two outlying points but given the relatively large size of the dataset and the fact that these points are not particularly extreme, there is no need to be concerned. Finally, we perform the cross-validation process on our test set. The Bias-corrected line doesn't fit very well to the right tail of the diagonal line.

We also plot **the observed proportions against the predicted probabilities**. Although there is some variation, there is no consistent deviation from what is expected. We have computed approximate 95% confidence intervals using the binomial variance. The line passes through most of these intervals confirming that the variation from the expected is not excessive.

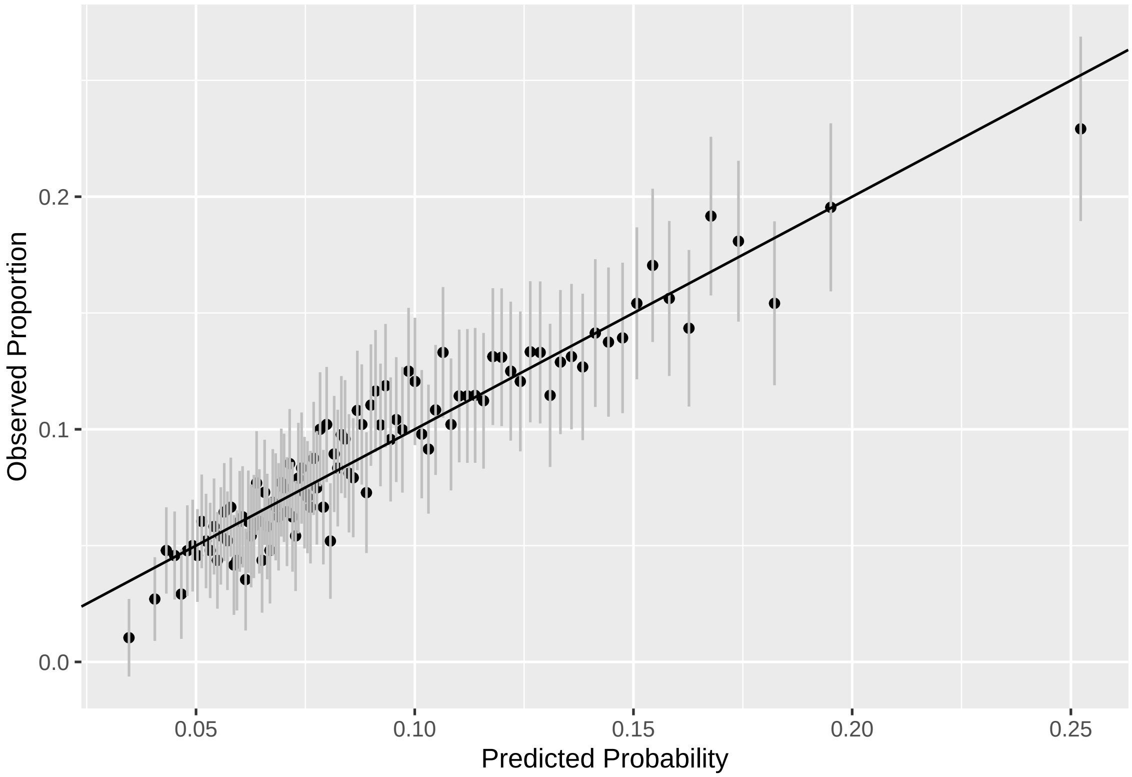

Figure 2: The observed proportions against the predicted probabilities

We perform a **Hosmer-Lemeshow test** and the p-value is given by 0.3782252. Since the p-value is quite large, we detect no lack of fit.

We also plot an **ROC curve** for both our train and test data. The **AUC** value is 0.63 for both of them, meaning that our logistic regression model can correctly discriminate **readmission** and **no readmission** 63% of the time.

4 Discussion Section

(i) Final Model Interpretation and Importance

In conclusion, we see that `age`, `number_diagnoses`, `diabetesMed`, `num_of_med`, `num_visits`, `Discharge`, `Length.of.Stay` are some relatively very significant covariates for predicting `readmission`. Our analysis showed that the profile of readmission differed significantly in patients within different `age` levels. The odds of `readmission` is expected to increase by a factor of $exp(1.993e-01)$ for people with `age` within `[60, 100)` as compared to those with `age[30, 60)`. And also the odds of `readmission` increase by a factor of $exp(9.134e-02)$ with each additional visit in `num_visits`. Therefore, `age`, `number_diagnoses`, `diabetesMed`, `num_of_med`, `num_visits`, `Discharge`, `Length.of.Stay` are relative useful predictors of `readmission` rates which may prove useful in developing strategies for reducing readmission rates and treatment costs for diabetic patients.

(ii) Limitations of Analysis

In our data analysis, we fit a **logistic regression** (GLM) model which assumes that observations are independent. However, To satisfy this assumption, we remove duplicate patient encounters, which lets us ignore that there may also be random variability across each encounter of those patients.

References

[1] World Health Organisation. (2016). *Global Report on Diabetes.*
[2] Medicare Payment Advisory Commission. (2007). $Report\ to\ the\ Congress\ promoting\ greater\ efficiency\ in\ Medicare.\ Washington,\ DC.$
[3] Beata Strack, 1 Jonathan P. DeShazo, 2 Chris Gennings, 3 Juan L. Olmo, 4 Sebastian Ventura, 4 Krzysztof J. Cios, 1 , 5 and John N. Clore 6 ,* (2014) Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.
[4] Bergenstal RM, Fahrbach JL, Iorga SR, Fan Y, Foster SA. Preadmission glycemic control and changes to diabetes mellitus treatment regimen after hospitalization. Endocrine Practice. 2012;18(3):371375.
[5] Baldwin D, Villanueva G, McNutt R, Bhatnagar S. Eliminating inpatient sliding-scale insulin: a reeducation project with medical house staff. Diabetes Care. 2005;28(5):10081011.