Gustavo Fernandez

8/12/2020

Statistical Methods for the Analysis of Length of Hospital Stay of 70,000 Clinical Database Diabetic Patients Records

1.Introduction

The Length of Hospital Stay is an important indicator of the efficiency of hospital management. Reduction in the number of inpatient days results in decreased risk of infection and medication side effects, improvement in the quality of treatment, and increased hospital profit with more efficient bed management. Therefore, assessing the length of stay (LOS) is especially important in organizing hospital services and health system.

The variable of interest Length of Hospital Stay is a count data. In statistics, count data refer to observations that have only nonnegative integer values ranging from zero to infinity, but they are always limited to some lesser distinct value-generally the maximum value of the count data being modeled. When the data being modeled consist of large number of distinct values, even if they are positive integers many statisticians prefer to model the counts as if they were continuous.

The aim of this study is to analyze a real-world data set containing 70,000 diabetic patients records and using the time in hospital as the response, implementing several methods such as Multivariate Linear Regression, Poisson Regression, and K-Nearest Neighbors (KNN), compare their performance and the usefulness of each method in predicting a patient expected length of stay as well as to identify predictors of LOS among diabetic patients. The variables used in the comparison process include, The Akaike Information Criterion (AIC) and the Root Mean Square Error (RMSE).

2.Literature Review

Many research has been conducted to predict a patient expected length of stay as well as to identify predictors of LOS among patients who have the same diagnosis-related group classification. However, most research on LOS have not been subjected to well-design modeling. Most studies have been conducted implementing linear, log-normal and logistic

regression approaches. LOS distribution is frequently positive skewed. The fact that LOS data is frequently positive skewed with a mode near 0 and heavy tails is a very important issue. Statistical methods are vulnerable to heteroscedastic variation and biased estimates of coefficients of predictors; surprisingly most research analyses do not check for skewness before embarking on their research, or the techniques used to deal with the problem either still violates fundamental model assumptions or result in loss of important information. For instance, (Basques, 2014) treated LOS as a continuous variable, and multivariate linear regression was performed with LOS as the continuous outcome variable to analyze the effect of each predictor variable on LOS and no assessment of distribution of the outcome variable LOS was mentioned in the analysis.

(H.H Dasenbrok, 2015) despite mentioning the positive skew distribution of LOS in the data, evaluated LOS dichotomously on a binomial split of the data and employed multivariate logistic regression to determine independent predictors of extended hospitalization days. In this study the researchers also, analyzed LOS continuously and multivariate linear regression was employed. Logarithmic transformations of the outcome variable are often used with ordinary least squares regression. The weakness of this is that many of this transformation fail to produce a successful approximation. For instance, (Smith,2008) applied a linear regression model to logged LOS. In this study, the log transformation was not successful at producing an approximate normal distribution. Failing to mention the result of the transformation is another issue, for instance (Husted, Holm & Jacobsen, 2009) used multivariate logistic regression analyses in order to identify significant parameters influencing LOS. However, the researcher did not mention anything about the distribution of the outcome variable. Another weakness of this method is that given the fact that LOS data is frequently positives skewed log-LOS is harder to interpret and might not be useful for policy making or simple enough to understand for the general population and since log-models are about geometric, not arithmetic, means, and retransformation are complicated due to heteroscedasticity (Manning WG,1998). For instance, (Husted, Holm & Jacobsen, 2009) by employing logistic regression found that for each year of age, the probability of staying more than 3 days increased by 2.4% per year of increasing age or 27 % per decade. Also found that women had almost 40% greater probability of staying more than 3 days than men. This findings in terms of probability, although important does not help in predicting the LOS of a single patient.

3.Methods

3.1 Preliminary Analysis and the Final Dataset.

The original dataset contained 50 features describing the diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter. The full list of features and their description is provided in Table 1.
The original dataset contains incomplete and redundant information as expected in any real-world data. There were several features that could not be considered in this analysis due to the extremely high percentage of missing values. These features were weight wich has (96.9%) of values missing and medical specialty (49.1%). Weight would have been an interesting feature to include but due to such an extremely percentage of missing values (96.9%) it was not included

in further analysis. Similarly, Medical specialty was excluded due to the high percentage of missing values. For the 2.2% of missing values for the variable race it was decided to remove those missing observation for further analysis.

Table 1. List of Features and their description in the initial dataset

Name	Type	Description and values	%missing
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, Other	2.2%
Weight	Numeric	Weight in pounds	96.9%
Medical specialty	Nominal	Values: cardiology, internal medicine, family\general practice, and surgeon	49.1%
Gender	Nominal	Values: male, female and unknown	0%
Age	Nominal	Group of 10-year interval: code 1 for [0,10), 2 for [10,20) … 10 for [90,100)	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Length of Hospital Stay	Numeric	Integer number of days between admission and discharge	0%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatients visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	was 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.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%

The preliminary dataset contained multiple inpatient visits for some patients and the observations could not be considered as statistically independent. We thus used only one encounter per patient; in particular, we considered only the first encounter for each patient as the primary admission. Our final data set for the analysis contains 69,570 patient records with 18 features and the variable of interest is Length of Hospital Stay. Each encounter corresponds to a unique patient diagnosed with diabetes, although the primary diagnosis may be different.

A training data set was build containing 75% of the final data set. The remaining observations were assigned to the testing data set.

3.2 Statistical Methods.

Multivariate linear Regression

Stepwise backward variable selection was implemented to select the optimal number of variables for our multivariate linear regression model. Then in the subset model selection, the best model was selected based on best Akaike Information Criterion (AIC). After fitting the model with the training data set. Highly influential observations with cooks' distance > $(\frac{4}{Sample\ size})$ were removed from the training set and the model was re-fitted. The coefficient estimates and p-values from the final multiple linear regression model are provided in table 3.

Model Assumptions

1. Observations are independent.
2. There must be a linear relationship between the outcome variable and the independent variables.
3. Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
4. No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
5. Homoscedasticity–This assumption states that the variance of error terms is similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

Diagnostic plots are provided in the appendix under model diagnostic. As well as the results of the Variance inflation Factor.

One of the points about statistical modeling rarely discussed is the relationship of the data to a probability distribution. All parametric statistical models are based on an underlying probability distribution. The linear regression models are based on the Gaussian or normal probability distribution. When we are attempting to estimate a least squares regression model, we are estimating the parameters of the underlying probability distribution that characterize the data.

The response variable that we are trying to model looks more like it follows a Poisson distribution than a Normal distribution see figure 1.

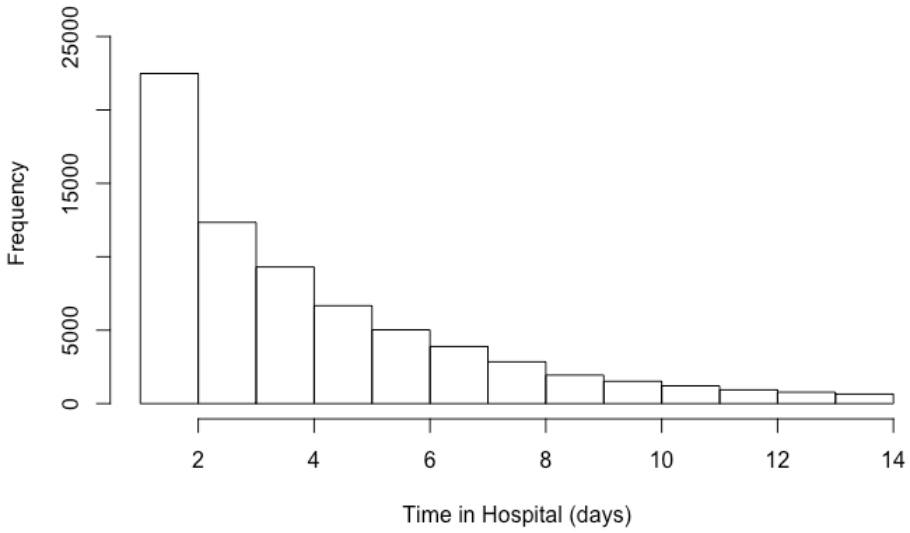

Figure 1. Distribution of Length of Hospital Stay

Poisson Regression

Poisson regression is fundamental to the modeling of count data. It was the first model specifically used to model counts, and it still stands at the base of the many types of count models available to analyst. However, the Poisson distributional assumption of equidispersion, is typically not met and becomes unsatisfactory the use of Poisson regression on real world data. Even though some adjustments can be made to remedy the problem of overdispersion often this is not possible.

Model Assumptions

1. The distribution is discrete with a single parameter, the mean (λ)
2. The response are non-negative integers
3. Observations are independent of one another
4. No observed counts have substantially more or less than what is expected based on the mean of the empirical distribution
5. The mean and variance of the model are identical
6. The Pearson Chi2 dispersion statistic has a value approximately 1. A value of 1. Results when the observed and predicted variances of the response are the same.

For the Poisson model is particularly important to have an understanding of the response variable in our case the Length of hospital stay. Some data description is necessary to have an understanding of the response farther than the shape of its distribution shown in figure 1. The

mean Length of Hospital Stay (in days) is 4.3 days and the variance (standard deviation square) is 8.7 days, the variance exceeds the mean in this case. Therefore, some violation of assumption number 5 occurs, however, the apparent overdispersion is not as extreme one. Based on a Poisson distribution with mean of 4.3 days we would expect a relative frequency of days spent in hospital for the case of 1 day to be 0.0583 or 5.83 % and based on the results of the data provided in table 2. We see that the relative occurrence of 1 day in hospital is 15.06%, Therefore, some violation of the assumption that indicates that for the poison regression no observed counts have substantially more or less than what is expected based on the mean of the empirical distribution occurs in our data.

Table 2. Distribution of Counts Data Set vs Poisson

Length of Stay	Relative Frequency	Poisson ($\lambda = 4.3$)
1	15.06%	5.83%
2	17.25%	12.54%
3	17.74%	17.98%
4	13.37%	19.33%
5	9.60%	16.62%
6	7.21%	11.91%
7	5.62%	7.32%
8	4.17%	3.91%
9	2.74%	1.88%
10	2.12%	0.80%
11	1.72%	0.32%
12	1.32%	0.11%
13	1.15%	0.04%
14	0.92%	0.01%

Despite some assumption being not completely satisfied, a Poisson model was built, stepwise backward variable selection was implemented to select the optimal number of variables for the Poisson Regression Model. Then in the subset model selection, the best model was selected based on best Akaike Information Criterion (AIC). After fitting the model with the training data set. Highly influential observations with cooks' distance > ($\frac{4}{Sample\ size}$) were removed from the training set and the model was re-fitted. The coefficient estimates and p-values from the final Poisson regression model are provided in table 4.

Each method we have seen so far has been parametric. For instances the beta coefficients in multiple linear regression are the parameters of the model, wich we estimated by fitting the

model. k-nearest neighbors (KNN) have no such parameters. Instead, it has a tuning parameter, k. This is a parameter which determines how the model is trained, instead of a parameter that is learned through training.

K-Nearest Neighbors (KNN) was trained using the training data set mentioned earlier and a K=255 was chosen to predict the response length of hospital stay of the test data set.
The attrtibutes used are age, admission type id, number of lab procedures, number of procedures, number of medications, number of outpatient visits, number of emergency visits, number of inpatient visits, number of diagnosis, max glucose levels, A1Cresult, Change of medication, Diabetes Medication.

Results

As can be seen in table 3 many of the features were highly significant predictor of length of hospital stay in the multi variate regression. The predictors Glucose serum test>300 A1Cresult>8, A1Cresult-None and A1Cresult-Norm were found not significant. Those features under the VIF test resulted in positive for multicollinearity. Therefore, might be indicating the reason why are not significant in the model.

Table 3. Coefficients terms estimated Linear Regression

Coefficients	Estimate	P Value
(Intercept)	0.5122	<0.001
RaceAsian	-0.1117	<0.001
Race-Caucasian	-0.0878	<0.001
Race-Hispanic	-0.0522	<0.001
Race-Other	-0.0572	<0.001
Gender-Male	-0.0271	<0.001
Age	0.0443	<0.001
Admission Type ID	-0.0197	<0.001
Number of lab procedures	0.0074	<0.001
Number of procedures	0.0088	<0.001
Number of medications	0.0254	<0.001
Number of outpatient visits	-0.0184	<0.001
Number of emergency visits	-0.0229	<0.001
Number of inpatient visits	0.0362	<0.001
Number of diagnoses	0.0379	<0.001
Glucose serum test>300	0.0322	0.2019
Glucose serum test- None	-0.2966	<0.001
Glucose serum -Norm	-0.0757	<0.001
A1Cresult>8	0.0309	0.0107
A1Cresult-None	-0.0094	0.3625
A1Cresult-Norm	0.0158	0.2297
Change of Medications NO	-0.0179	<0.001
Diabetes Medicacation YES	-0.0168	0.0052

In the Poisson regression model similarly to the multiple linear regression model many of the features were found significant contributor of length of hospital stay see table 4. Glucose serum test>300, A1Cresult>8, A1Cresult-None, A1Cresult-Norm were found not significant.

Interestingly, the coefficients estimate from the Poisson vs Linear Regression changes slightly but not too drastically, for example the coefficient for race Caucasian in the linear model is -0.0878 while in the poisson is -0.0727, the estimates for gender male in the linear is - 0.0271while in the poisson is -0.0322.

Table 4. Coefficients Estimates Poisson Regression

Coefficients	Estimate	P Value
(Intercept)	0.3437	<0.001
RaceAsian	-0.0729	0.07723
Race-Caucasian	-0.0727	<0.001
Race-Hispanic	-0.0747	<0.001
Race-Other	-0.0918	<0.001
Gender-Male	-0.0322	<0.001
Age	0.0456	<0.001
Admission Type ID	-0.0322	<0.001
Number of lab procedures	0.0081	<0.001
Number of procedures	0.0038	<0.001
Number of medications	0.0298	<0.001
Number of outpatient visits	-0.0253	<0.001
Number of emergency visits	-0.0353	<0.001
Number of inpatient visits	0.0246	<0.001
Number of diagnoses	0.0379	<0.001
Glucose serum test>300	0.0509	0.1356
Glucose serum test- None	-0.3182	<0.001
Glucose serum -Norm	-0.0112	<0.001
A1Cresult>8	0.0584	0.0107
A1Cresult-None	0.0464	0.3625
A1Cresult-Norm	0.0301	0.2297
Change of Medications NO	-0.0218	<0.001
Diabetes Medicacation YES	-0.0275	<0.001

The KNN performed better in predicting the test response than the multivariate linear regression and the Poisson regression model. KNN provided a residual mean square error rmse=2.426. While the multiple linear regression model provided a rmse=2.483. The poisson regression provided the worst rmse=2.518. However, when comparing the two regression models the Poisson regression model resulted in a better AIC score AIC=190861 vs Linear Regression AIC=212321. Therefore, when modeling this type of data, the Poisson would be a better model than the Multiple Linear. The result is summarized in table 5.

Table 5. Performance Comparison of the Methods

	AIC	RMSE
Multivariate Linear Regression Model	212321	2.483
Poisson Regression Model	190861	2.518
	KNN	2.426

Discussion

When modeling count data is important to have an understanding of the response variable and to check for overdispersion and other key assumptions of the Poisson. Even in the situation when the data present some small departures from the assumptions of the Poisson regression model, Poisson regression might be better model than the multiple linear regression to make inference and to predict the length of hospital stay of a single patient. KNN method performed better in predicting the test response. However, if our goal is to determine the factors that influence the length of hospital stay from available features or to determine the marginal effects then KNN would not be useful. The implementation of a Poisson regression might be more appropriate than Linear Regression.

Limitation

When fitting the KNN, normalizing the training data can improve its accuracy dramatically, in this analysis the training data was not normalized to train the KNN. Also, the selection if K was based on empirical rules it was not subject of an analysis to determine the most optimal K. Interaction effects were not tested in the regression models. Test for overdispersion were not performed when building the Poisson regression. However, it was stated that the data showed apparent overdispersion table 2.