International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 6, No. 2 (Nov-2019)

http://dx.doi.org/10.12785/ijcts/060202

Modelling Length of Stay in Hospitals using Multinomial Regression

Harini S[1], Subbiah M[2] and M.R. Srinivasan[1]

[1] *Department of Statistics, University of Madras, Chennai -5, India*
[2] *HCL Technologies, Chennai, India*

Received September 27, 2018 Revised June 10, 2019, Accepted October 18, 2019, Published November 1, 2019

Abstract: Hospital management is generally focused on studying the length of stay of patients since the measure has an impact on hospital resources. It is a challenging task for the hospital management to model the length of stay as they are asymmetric and heterogeneous in nature. Diabetes is a major health problem prevalent worldwide which leads to hospitalization over a time period. The present study deals with stay of diabetes patients classified as very short, short, medium and long duration of stay based on quantile classification rather than arbitrary approach. In this study, we have attempted to include an important covariate known as medical record since it assist in reducing the stay of a patient and can thereby accommodate more patients deserving treatment as inpatients. Based on the multiple levels of the response variable, we have considered fitting multinomial regression model for length of stay on diabetes. Further, this study has considered the validation of variable selection procedure for model fitting using subsampling approach. In conclusion, it has been identified that medical records is one of the important factor affecting the stay of patients and subsampling approach has been helpful in building the final model.

Keywords: Length of Stay, Medical Records, Multinomial Regression, Variable Selection, Subsampling

1. INTRODUCTION

Healthcare management is interested in studying the factors responsible for the duration of stay of patient in a hospital. With an increasing population, the hospital management and the administrators focus their studies on modelling the stay as they are the backbone for forecasting and making futuristic business decisions. These modelling and prediction of Length of Stay (LOS) helps the administrators for allocating the beds and utilizing the hospital resources. Efficiency of discharges are studied by Udayai *et al* [19], that the major role of the hospital management is to ensure satisfaction to the patients, bed availability and to maintain high standards of quality. Shukla *et al* [16] considered cross sectional study on insured patients and turnaround time of discharge process which are analyzed using t-test.

Modelling of LOS is important and is required for every hospital to allocate the resources (Gul *et al* [7]; Gardiner, [6]). Several studies discussed the nature of LOS as right skewed, asymmetric, and heterogeneous in nature (Harini *et al* [8]; Zenga *et al* [21]; Faddy *et al* [5]). It is also observed that stay might vary from patient to patient, like older age patient might have a longer stay when compared to younger age due to medical complexities (Jones *et al* [10]). Kembe *et al* [11] considered queuing model to determine the optimal number of beds for orthopedic clinic. Papi *et al* [14] discussed that estimating the stay of patient is challenging since the nature of LOS is asymmetric which makes the study difficult while fitting distributions. Singler *et al* [17] studied correlation of age of patients on LOS and admission rate in emergency department.

In this study, we have considered dataset from UCL data repository for modelling LOS (Beata *et al* [3]). The data deals with diabetes LOS representing 130 US hospitals with sample size of 9,548 patient encounters. Poisson regression model has been considered.by Carter *et al* [4] by treating LOS to be count in nature. Earlier studies considered the response variable to be continuous in nature (Verburg *et al* [20]; Austin [2]). In this study, we attempted an initial investigation with the hospital administrators to understand about the nature of response variable. From the investigation

E-mail address: harinivas11@yahoo.co.in, sisufive@gmail.com, mrsrin8@gmail.com

it has been observed that hospital management are keen in treating LOS to be categorical in nature like Long (L), Medium (M), Short(S), and Very Short (VS).

Tsai *et al* [18] has discussed that psychiatrists are able to predict LOS with accuracy for the patients they treat. Earlier study by Harini *et al* [9] considered fitting length of stay by multi stage classification of covariates using transformed gamma-Pareto distribution with the help of covariate such as Medical Records (MR), gender and age. However, the study has not been focused on modelling the length of stay in hospital. Hence, in this paper, we have considered MR as important predictor for modelling the length of stay. This predictor will be helpful for the hospital management to study and determine the stay of patients. We have also considered overall LOS model for comparative purpose.

Earlier studies classified the stays as long or short purely based on arbitrary approach (Meadows *et al* [12]), hence, this study attempts to classify the LOS by understanding the shape of the response variable. Hence, quantile approach of classification is considered for choosing the cut-off for stays since they are not affected by extreme observations. The cut off for the stays at multiple levels are confined with an initial investigation with the administrators. Since the nature of the response is at multiple levels, the most widely used Multinomial Regression modelling (Agresti [1]) is considered to study diabetes LOS.

This paper further considers three step procedure for multinomial regression modelling. As the first step variable selection is considered then Model fitting and Assessment of Model Fit are studied. Further, this study attempted validation using subsampling approach for variable selection step. This helps in deciding whether to include or exclude the insignificant covariates in the final model. The paper details the description of the diabetes dataset in Section 2 followed by the methodology in Section 3. The results of the analysis are discussed in Section 4 and general conclusion is detailed in Section 5.

2. DATA DESCRIPTION

Data from UCL data repository is considered in this study. This data involves the timeframe of 1999 to 2008, comprising of Midwest, Northeast, South and West USA. This contains both Type 1, Type 2 diabetes and do not involve patients who died or discharged to hospice. We have considered sixteen explanatory covariates with the response variable treated as "time in hospital" i.e. LOS which ranges from 1 to 14 days. Table 1 provides the details of variables considered in this study with its nature. The derived covariate Medical Records (MR) is obtained from the covariates number of inpatient and number of outpatient visits. The LOS dataset in this study considers 16 covariates, however the derived LOS dataset does not involve covariates such as glucose serum test, number of outpatient and number of inpatient visits. Since MR is derived from number of inpatient and outpatient visits, hence to avoid overfitting they are not considered.

Table 1. Nature of Predictors Considered in Modelling Diabetes Length of Stay

S.No	Variables	Nature
1	Gender	Categorical
2	Age Group	Categorical
3	Race	Categorical
4	Readmitted	Categorical
5	Number Lab Procedures	Continuous
6	Number of Procedures	Continuous
7	Number of Medications	Continuous
8	Number of Outpatient Visits	Continuous
9	Number of Emergency	Continuous
10	Number of Inpatient Visits	Continuous
11	Number of Diagnosis	Continuous
12	Glucose Serum Test	Categorical
13	HBA1c Result	Categorical
14	Insulin	Categorical
15	Change	Categorical

S.No	Variables	Nature
16	Diabetes Medication	Categorical
17	Time in Hospital(LOS)	Continuous
18	Medical Records	Categorical

The descriptive statistics such as mean and standard deviation of the LOS dataset are detailed in Table 2.

Table 2. Summary Measures of Predictors Considered in Modelling Diabetes Length of Stay Patients

Variables	Category	Mean±SD
Gender	Female	4.47 ± 2.84
	Male	4.28 ± 2.87
Age Group	10-40	3.66 ± 2.69
	40-70	4.21 ± 2.84
	70-100	4.62 ± 2.85
Race	AfricanAmerican	4.38 ± 2.90
	Caucasian	4.43 ± 2.86
	Hispanic	3.64 ± 2.42
	Others	4.04 ± 2.74
Readmitted	< 30	4.78 ± 2.88
	> 30	4.47 ± 2.87
	No	4.19 ± 2.82
Glucose Serum Test	0	4.19 ± 2.77
	< 200	4.20 ± 3.07
	200-300	6.01 ± 3.37
	None	4.55 ± 2.90
	Norm	3.72 ± 2.51
HBA1c Result	>8	4.69 ± 2.99
	None	4.34 ± 2.83
	Norm	4.41 ± 2.79
Insulin	Down	4.83 ± 3.00
	No	4.05 ± 2.72
	Steady	4.20 ± 2.73
	Up	5.10 ± 3.05
Change	Ch	4.69 ± 2.94
	No	4.06 ± 2.72
Diabetes Medication	No	4.08 ± 2.73
	Yes	4.47 ± 2.88
Medical Records	MRK	4.56±2.92
	MRU	4.19±2.78

3. STATISTICAL METHODOLOGY

The important step before fitting a model is selection of variable which involves backward elimination, forward elimination or stepwise elimination. For this study, the most commonly used backward elimination is considered. The advantage of backward elimination is that the decision maker has the opportunity to look at all the independent variables in the model before removing the variables that are not significant. Statistical significance should not be the sole criterion for inclusion of a term in a model. It is reasonable to include a variable that is central to the purpose of the study and report its estimated effect even if it is not statistically significant.

Further, the researcher can decide whether to include or exclude the results of the variable to the model or not. It is important that researcher has to understand the strengths of the covariates which might influence the study. Importance of the variable selected in the model should be verified and for this purpose, this study considers subsampling approach.

3.1 Subsampling Approach for Validating Variable Selection

This approach is implemented for validating the covariates which are excluded in the variable selection.

1. To attain the subsample, random samples are drawn with varying size (n=30%, 40%, 50%, 60% and 70%) from the dataset.
2. Draw each of the subsamples from the dataset and then proceed with backward elimination procedure to identify the non- significant variables.
3. The backward elimination procedure is repeated 500 times to find the number of times the variables are excluded in each of the subsampling approach.
4. The results obtained are validated with the excluded variables in the original model.

To decide whether to include or exclude the non-significant (excluded variables) variables from the model, Likelihood Ratio Test (LRT) is considered. It assists in identifying whether adding or excluding the non-significant variable has any improvement in the model. P-values helps to decide whether to reject or not to reject the model. Further, if the variables observed are significant when compared with the full model in likelihood ratio test, then the variables are retained else they would be removed and revised model will be fitted.

There are three components in a Generalized Linear Model – random component, linear predictor and link function. In GLM, response variable plays a vital role and depending upon the nature of the response variable, proper modelling for the dataset has to be chosen. They can be categorical or continuous in nature and few of the commonly used models are binary, multinomial, normal, and Poisson. In this study, the nature of response variable is at multiple levels hence multinomial regression is discussed. The estimates, diagnostic procedures are discussed in detail by Agresti [1]

The link functions describes the relation between linear predictor and mean of the distribution function. It is necessary to choose appropriate link functions since it helps in drawing inferences about the parameters (β). Therefore, logit, probit and Complementary Log Log (Clog) link function are considered in this paper. For this study, the most widely used multinomial regression is considered since the response variable has more than two categories (Agresti [1], Monyai *et al* [3]) .Once the model is fitted then assessment of the model is carried out using McFadden R squared test, Chi-squared test, AIC, BIC. These are helpful in assessing and choosing the best model. Based on least AIC, BIC values, best model is chosen and to confirm whether the predictors are significant, Chi-Squared test is helpful.

4. ANALYSIS AND RESULTS

As discussed in the methodology, we consider classification by understanding the shape of the response variable, histogram is plotted which shows that diabetes LOS is right skewed in nature.

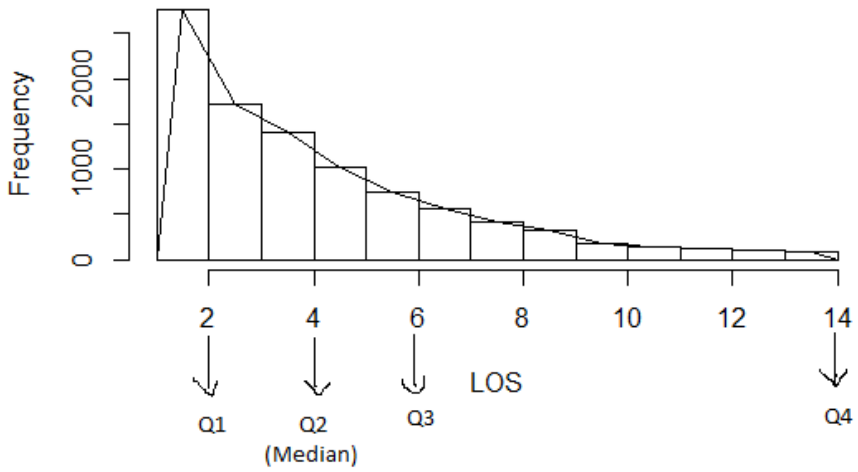

Figure 1. Histogram of Diabetes Length of Stay in Hospital

From Figure 1, the shape of the LOS is observed to be right skewed and asymmetric in nature. It can be observed that mean and median are 4.38 and 4.00 respectively. Quantile measure is considered as the cut-off for classifying the stays – very short (1-2), short (3-4), medium (5-8), long (9-14). The results of the classifications are also confined with initial investigation carried out with the hospital administrators and management.

As discussed in methodology, based on backward elimination method, it is observed that 1) Gender Male-L, 2) Age 40-70-L, 3) Age 40-70-M, 4) Age 40-70-S, 5) Race –Caucassian –S, 6) Race – Hispanic –L, 7) Race - Hispanic-M, 8) Race – Hispanic -S, 9) Race - Other-L, 10) Race - Other-M, 11) Race - Other-S, 12) Read >30 - L, 13) No. of Emergency – L, 14) No. of Emergency – M, 15) HBA1 None - L, 16) HBA1 None - S, 17) HBA1 Norm - L, 18) HBA1 Norm - M, 19) HBA1 Norm - S, 20) Insulin No - L, 21) Insulin No -M, 22) Insulin No - S, 23) Insulin Steady - M, 24) Insulin Steady - S, 25) Change No - L, 26) Change No - M, 27) Change No - S, 28) DiabMed Yes-S are observed to be non-significant for derived LOS. Similarly, all the variables as mentioned above are observed to be non-significant in Overall LOS except number of emergency. The variable other than above which are excluded in derived model are Readmitted No – L, Number of Outpatient – L, Number of Outpatient – S, Number of Inpatient – S, Glucose Serum 0 – L, Glucose Serum 0 - S, Glucose Serum 200-300 – L, Glucose Serum 200-300 – S, Glucose Serum None – L, Glucose Serum None - M, Glucose Serum None - S , Glucose Serum Norm – L, Glucose Serum Norm - M, and Glucose Serum Norm - S. To validate the results, subsampling approach for n = 30%, 40%, 50%, 60% and 70% are carried out.

For each model, variable selection using backward regression approach is carried out 500 times for varying n. The percentage of times each of the excluded variable in original model getting excluded in subsampling(varying n) for derived and overall LOS dataset are depicted in Figure 2. It can be observed that the variables excluded in original model are 85% of the times excluded in the subsampling model for both derived and overall LOS model. It is also observed that variable which were included in the original model are included in the subsampling approach.

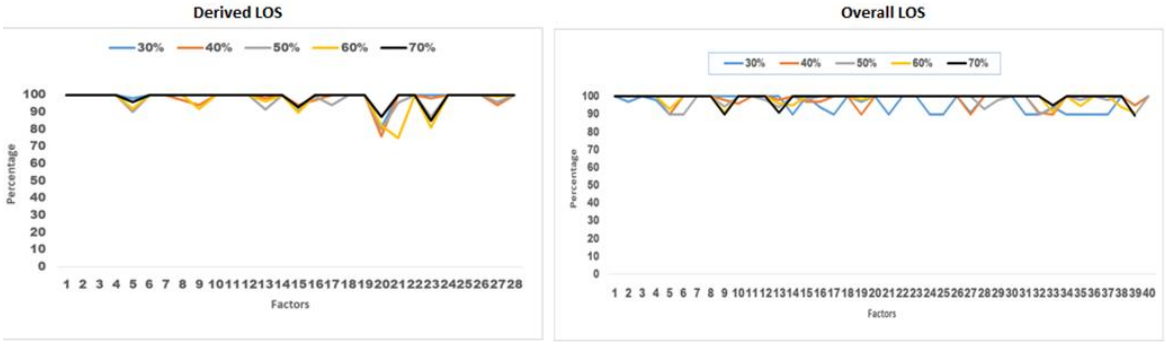

Figure 2. Validation for Variable Selection Using Subsampling Approach in Multinomial Regression Model for Diabetes Length of Stay Dataset.

Table 3 and 4 discusses the multinomial regression Full model fit with the details of estimates SE, p value, 95% confidence interval for derived and Overall LOS models respectively. Table 5 provides the details of assessment of goodness of fit for the full model. It can be observed in full model from Table 3 and Table 4 that variables which are excluded in variable selection are found to be not significant in the Full model for both derived and overall LOS model. In case of assessment of fit, from Table 5 it can be observed that derived LOS model is better since it as a least AIC and BIC when compared to overall LOS. Even though predictors are found to be significant it needs to be decided whether the model might improve after excluding the non-significant variables. As discussed in the methodology, we compare full model with the individual models by dropping the insignificant variables for both derived and overall LOS datasets for the categorical variables. Further, as discussed in methodology, we test whether the significant of the variables in the model using LRT. The result shows that model without gender, HBA1c, Change are observed to be non-significant and model without age group, race, readmitted, DiabMed yes are significant.

Table 3. Multinomial Regression Full Model for Diabetes Length of Stay (Derived Model)

					95% CI	
Predictors		**Estimate**	**SE**	**p value**	**LL**	**UL**
Gender Male						
	L	-0.0446	0.0873	0.6098	-0.2157	0.1266
	M	-0.1520	0.0616	0.3556	-0.2728	-0.0312
	S	0.1281	0.0554	0.3875	-0.2367	-0.0195
Age 40-70						
	L	-0.0895	0.2195	0.6834	-0.5198	0.3408
	M	0.1391	0.1505	0.3556	-0.1560	0.4341
	S	0.1081	0.1251	0.3875	-0.1371	0.3532
Age 70-100						
	L	0.6106	0.2232	0.0001	0.1731	1.0481
	M	0.7782	0.1540	0.0001	0.4760	1.0797
	S	0.4964	0.1292	0.0001	0.2431	0.7496
Race Caucasian						
	L	-0.4678	0.1338	0.0004	-0.7301	-0.2056
	M	-0.2851	0.0967	0.0032	-0.4747	-0.0955
	S	-0.1649	0.0863	0.0565	-0.3341	0.0043
Race Hispanic						
	L	-0.4031	0.3448	0.2424	-1.0789	0.2727
	M	-0.0252	0.2133	0.9058	-0.4434	0.3929
	S	-0.0694	0.1851	0.7078	-0.4321	0.2933
Race Other						
	L	-0.5292	0.2657	0.0464	-1.0500	-0.0084
	M	-0.1245	0.1771	0.4819	-0.4716	0.2225
	S	-0.0637	0.1558	0.6825	-0.3690	0.2415
Readmitted > 30						
	L	-0.1951	0.1353	0.1494	-0.4603	0.0702
	M	-0.3681	0.0985	0.0002	-0.5611	-0.1751
	S	-0.1993	0.0922	0.0306	-0.3800	-0.0186
Readmitted No						
	L	-0.3250	0.1403	0.0001	-0.6000	-0.0500
	M	-0.4667	0.1011	0.0001	-0.6648	-0.2686
	S	-0.2900	0.0942	0.0001	-0.4746	-0.1055
Number of Lab procedures						
	L	0.0259	0.0024	0.0001	0.0211	0.0307
	M	0.0236	0.0016	0.0001	0.0204	0.0268
	S	0.0118	0.0014	0.0001	0.0090	0.0146
Number of Procedures						
	L	0.2878	0.0285	0.0001	0.2320	0.3436
	M	0.1703	0.0231	0.0001	0.1250	0.2156
	S	0.0891	0.0220	0.0001	0.0460	0.1322
Number of medications						
	L	0.2201	0.0079	0.0001	0.2046	0.2356
	M	0.1573	0.0063	0.0001	0.1449	0.1698
	S	0.0891	0.0059	0.0001	0.0775	0.1008

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
Number of emergency						
	L	-0.0601	0.0433	0.1605	-0.1455	0.0241
	M	-0.0400	0.0282	0.1572	-0.0953	0.0154
	S	-0.0653	0.0263	0.0130	-0.1168	-0.0137
Number of diagnosis						
	L	0.2162	0.0347	0.0001	0.1481	0.2843
	M	0.1634	0.0206	0.0001	0.1231	0.2037
	S	0.0827	0.0168	0.0001	0.0499	0.1156
HBA1c None						
	L	0.2060	0.1339	0.1241	-0.0565	0.4685
	M	0.2332	0.1000	0.0197	0.0372	0.4291
	S	0.0898	0.0914	0.3259	-0.0893	0.2688
HBA1c Norm						
	L	-0.2176	0.2316	0.3476	-0.6716	0.2364
	M	-0.0893	0.1670	0.5927	-0.4165	0.2379
	S	0.0763	0.1500	0.6111	-0.2178	0.3704
Insulin No						
	L	-0.3320	0.1712	0.0524	-0.6676	0.0035
	M	-0.1929	0.1196	0.1069	-0.4274	0.0416
	S	0.0185	0.1070	0.8624	-0.1911	0.2282
Insulin Steady						
	L	-0.3960	0.1579	0.0121	-0.7054	-0.0866
	M	-0.2094	0.1131	0.0641	-0.4310	0.0122
	S	0.0592	0.1024	0.5632	-0.1415	0.2598
Insulin Up						
	L	0.4675	0.1446	0.0012	0.1841	0.7508
	M	0.3166	0.1125	0.0641	0.0960	0.5372
	S	0.2327	0.1057	0.5632	0.0254	0.4399
Change No						
	L	0.1056	0.1374	0.4596	-0.1676	0.3708
	M	-0.0086	0.0932	0.9256	-0.1889	0.1717
	S	-0.1531	0.0801	0.0560	-0.3101	0.0039
Diabmed Yes						
	L	-0.3433	0.1477	0.0201	-0.6327	-0.0538
	M	-0.3044	0.0972	0.0017	-0.4949	-0.1139
	S	-0.0628	0.0862	0.4664	-0.2316	0.1061
MRU						
	L	0.5690	0.1012	0.0001	0.3707	0.7673
	M	0.5467	0.0712	0.0001	0.4072	0.6862
	S	0.2435	0.0642	0.0001	0.1176	0.3694
Intercept						
	L	-7.4728	0.4445	0.0001	-8.3458	-6.6028
	M	-4.6136	0.2858	0.0001	-5.1737	-4.0534
	S	-2.1624	0.2409	0.0001	-2.6346	-1.6902

Table 4. Multinomial Regression Full Model for Diabetes Length of Stay (Overall Model)

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
Gender Male						
	L	-0.0422	0.0875	0.6297	-0.2136	0.1292
	M	-0.1535	0.0617	0.0129	-0.2745	-0.0325
	S	-0.1283	0.0554	0.0206	-0.2370	-0.0197
Age 40-70						
	L	-0.0435	0.2202	0.8436	-0.4751	0.3882

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
	M	0.1886	0.1514	0.2127	-0.1081	0.4854
	S	0.1230	0.1256	0.3274	-0.1231	0.3690
Age 70-100						
	L	0.6701	0.2244	0.0028	0.2302	1.1099
	M	0.8368	0.1552	0.0000	0.5327	1.1409
	S	0.5135	0.1299	0.0001	0.2588	0.7681
Race Caucasian						
	L	-0.4677	0.1342	0.0005	-0.7307	-0.2047
	M	-0.2797	0.0971	0.0040	-0.4700	-0.0894
	S	-0.1639	0.0865	0.0581	-0.3334	0.0056
Race Hispanic						
	L	-0.3931	0.3448	0.2542	-1.0688	0.2826
	M	-0.0157	0.2138	0.9414	-0.4347	0.4033
	S	-0.0701	0.1852	0.7052	-0.4331	0.2930
Race Other						
	L	-0.5196	0.2660	0.0508	-1.0410	0.0018
	M	-0.1177	0.1773	0.5069	-0.4652	0.2298
	S	-0.0612	0.1558	0.6943	-0.3665	0.2441
Readmitted > 30						
	L	-0.1536	0.1363	0.2598	-0.4208	0.1136
	M	-0.3287	0.0994	0.0009	-0.5235	-0.1340
	S	-0.1841	0.0929	0.0475	-0.3662	-0.0020
Readmitted No						
	L	-0.2649	0.1421	0.0622	-0.5434	0.0135
	M	-0.4161	0.1025	0.0000	-0.6170	-0.2152
	S	-0.2713	0.0953	0.0044	-0.4581	-0.0845
Number of Lab procedures						
	L	0.0260	0.0025	0.0000	0.0212	0.0309
	M	0.0238	0.0017	0.0000	0.0205	0.0271
	S	0.0118	0.0014	0.0000	0.0089	0.0146
Number of Procedures						
	L	0.2910	0.0285	0.0000	0.2352	0.3469
	M	0.1728	0.0231	0.0000	0.1274	0.2182
	S	0.0898	0.0220	0.0000	0.0467	0.1330
Number of medications						
	L	0.2185	0.0079	0.0000	0.2030	0.2340
	M	0.1558	0.0064	0.0000	0.1434	0.1683
	S	0.0884	0.0060	0.0000	0.0767	0.1001
Number of Outpatients						
	L	-0.0317	0.0209	0.1287	-0.0727	0.0092
	M	-0.0386	0.0162	0.0171	-0.0703	-0.0069
	S	-0.0012	0.0137	0.9292	-0.0280	0.0255
Number of emergency						
	L	-0.0906	0.0450	0.0441	-0.1789	-0.0024
	M	-0.0582	0.0294	0.0478	-0.1158	-0.0006
	S	-0.0702	0.0268	0.0089	-0.1228	-0.0176
Number of Inpatient						
	L	0.0947	0.0343	0.0058	0.0274	0.1619
	M	0.0753	0.0255	0.0031	0.0254	0.1252
	S	0.0254	0.0239	0.2884	-0.0215	0.0723
Number of diagnosis						
	L	0.2230	0.0348	0.0000	0.1547	0.2913
	M	0.1702	0.0207	0.0000	0.1297	0.2107
	S	0.0840	0.0168	0.0000	0.0511	0.1170
Glucose Serum 0						
	L	0.7288	0.5280	0.1675	-0.3060	1.7635
	M	0.7708	0.3800	0.0425	0.0260	1.5156
	S	0.2366	0.3144	0.4516	-0.3795	0.8528

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
Glucose Serum 200-300						
	L	1.2145	0.6228	0.0512	-0.0061	2.4352
	M	1.0089	0.4812	0.0360	0.0657	1.9520
	S	0.1704	0.4421	0.6998	-0.6960	1.0369
Glucose Serum None						
	L	0.0300	0.5205	0.9541	-0.9902	1.0501
	M	0.1556	0.3752	0.6784	-0.5798	0.8910
	S	-0.0566	0.3101	0.8551	-0.6644	0.5511
Glucose Serum Norm						
	L	-0.1241	0.6919	0.8576	-1.4802	1.2320
	M	0.4157	0.4471	0.3525	-0.4606	1.2921
	S	-0.0154	0.3745	0.9672	-0.7495	0.7187
HBA1c None						
	L	0.1977	0.1341	0.1405	-0.0652	0.4605
	M	0.2238	0.1001	0.0255	0.0275	0.4200
	S	0.0879	0.0914	0.3365	-0.0913	0.2671
HBA1c Norm						
	L	-0.1924	0.2319	0.4068	-0.6469	0.2621
	M	-0.0763	0.1672	0.6482	-0.4041	0.2515
	S	0.0821	0.1501	0.5842	-0.2121	0.3764
Insulin No						
	L	-0.3090	0.1720	0.0724	-0.6460	0.0281
	M	-0.1712	0.1202	0.1544	-0.4068	0.0644
	S	0.0262	0.1074	0.8070	-0.1842	0.2367
Insulin Steady						
	L	-0.3739	0.1582	0.0181	-0.6840	-0.0638
	M	-0.1932	0.1135	0.0886	-0.4156	0.0292
	S	0.0645	0.1027	0.5296	-0.1367	0.2657
Insulin Up						
	L	0.4859	0.1448	0.0008	0.2020	0.7697
	M	0.3376	0.1128	0.0028	0.1165	0.5588
	S	0.2368	0.1059	0.0253	0.0293	0.4444
Change No						
	L	0.0869	0.1378	0.5284	-0.1832	0.3569
	M	-0.0188	0.0922	0.8384	-0.1995	0.1619
	S	-0.1539	0.0802	0.0548	-0.3110	0.0032
Diabmed Yes						
	L	-0.3165	0.1482	0.0327	-0.6070	-0.0261
	M	-0.2821	0.0975	0.0038	-0.4732	-0.0909
	S	-0.0560	0.0863	0.5164	-0.2252	0.1132
Intercept						
	L	-7.7998	0.6807	0.0000	-9.1339	-6.4657
	M	-5.0013	0.4701	0.0000	-5.9226	-4.0799
	S	-2.1973	0.3907	0.0000	-2.9630	-1.4317

Table 5. Assessment of Fit for Multinomial Regression Full Model for Diabetes Length of Stay Dataset

Models	Measures	Values
Derived	McFadeen R^2	0.1318
	χ^2	3672
	(p-value)	(0.0001)
	AIC	**21867.65**
	BIC	**22351.48**
Overall	McFadeen R^2	0.1302
	χ^2	3687
	(p-value)	(<0.0001)
	AIC	21878.96
	BIC	22448.25

Based on LR test, it can be observed Model excluding gender, Model excluding variable race, HBA1C, and Change are statistically insignificant which infers that model containing these variables does not provide any significant improvement therefore, these variables are removed from the model and revised model is fitted for derived LOS dataset. In the overall LOS model it is observed with similar LRT results as in derived model with an addition of variable glucose serum to be significant. Therefore, we have excluded the categorical variables such as gender, race, HBA1C, and change; continuous variable number of emergency from the derived and overall LOS model. Number of outpatient visits is not significant in overall model hence it is excluded. The refined model are discussed in Table 6 and 7 which provides the details of estimates, standard error, p values, 95% CI for multinomial for derived and overall LOS dataset. Following are observed from Table 6, 7, and 8.

It is observed that age 40-70, Insulin No, Insulin Steady – S, DiabMed Yes – S are not significant in the refined model for derived and overall LOS model. However, we have retained them in this refined model because other level of categorical variable such as age 70-100, DiabMed, Insulin are significant. MRU is observed to be significant, this shows that they are one of the important factor for longer duration of stay in hospital. The relative risk for longer stay for MRU patient is observed to 1.76 times high in L, 1.72 times in M and 1.27 times in S when compared to very short stay.

Even though age 40-70 is insignificant, direction of the estimates are found to be appropriate in the case of derived and overall LOS model. It can be observed that relative risk of higher stays decreases for the age 40-70. For age 40-70 (S) we have exp (0.1351) which is 1.144; age 40-70 (M) is 1.149; age 40-70 (L) is 0.90. Similar kind of behaviour can also be observed in 95% confidence interval for Upper Limit (UL) and Lower Limit (LL). The least standard error can be observed in number of lab procedures (S) and highest standard error can be observed in age 70-100(L). From Table 8, it can be observed that derived LOS is better model when compared to overall LOS model since least value is observed in AIC and BIC. The predictors are found to be significant for both the models since chi-squared statistic is observed to be significant.

Table 6. Multinomial Regression Final Model for Diabetes Length of Stay (Derived Model)

					95% CI	
Predictors		**Estimate**	**SE**	**p value**	**LL**	**UL**
Age 40-70						
	L	-0.1054	0.2195	0.6834	-0.5198	0.3408
	M	0.1391	0.1505	0.3556	-0.1560	0.4341
	S	0.1351	0.1251	0.3875	-0.1371	0.3532
Age 70-100						
	L	0.6106	0.2232	0.0001	0.1731	1.0481
	M	0.7782	0.1540	0.0001	0.4760	1.0797
	S	0.4964	0.1292	0.0001	0.2431	0.7496
Number of Lab procedures						
	L	0.0259	0.0024	0.0001	0.0211	0.0307
	M	0.0236	0.0016	0.0001	0.0204	0.0268
	S	0.0118	0.0014	0.0001	0.0090	0.0146
Number of Procedures						
	L	0.2878	0.0285	0.0001	0.2320	0.3436
	M	0.1703	0.0231	0.0001	0.1250	0.2156
	S	0.0891	0.0220	0.0001	0.0460	0.1322
Number of medications						
	L	0.2201	0.0079	0.0001	0.2046	0.2356
	M	0.1573	0.0063	0.0001	0.1449	0.1698
	S	0.0891	0.0059	0.0001	0.0775	0.1008
Number of diagnosis						
	L	0.2162	0.0347	0.0001	0.1481	0.2843
	M	0.1634	0.0206	0.0001	0.1231	0.2037
	S	0.0827	0.0168	0.0001	0.0499	0.1156
Insulin No						
	L	-0.2869	0.1471	0.0511	-0.5753	0.0014
	M	-0.2082	0.1196	0.0458	-0.4274	-0.0004
	S	-0.0800	0.1070	0.3963	-0.1911	0.2282
Insulin Steady						
	L	-0.3373	0.1579	0.0157	-0.7054	-0.0866
	M	-0.2080	0.1131	0.0395	-0.4067	-0.0100
	S	-0.0170	0.1024	0.5632	-0.1415	0.2598

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
Insulin Up						
	L	0.4675	0.1446	0.0023	0.1841	0.7508
	M	0.3166	0.1125	0.0077	0.0960	0.5372
	S	0.2327	0.1057	0.0359	0.0254	0.4399
DiabmedYes						
	L	-0.3843	0.1477	0.0058	-0.6327	-0.1108
	M	-0.3191	0.0972	0.0007	-0.4949	-0.1316
	S	-0.0010	0.0862	0.9061	-0.2316	0.1061
MRU						
	L	0.5690	0.1012	0.0001	0.3707	0.7673
	M	0.5467	0.0712	0.0001	0.4072	0.6862
	S	0.2435	0.0642	0.0001	0.1176	0.3694
Intercept						
	L	-7.4728	0.4445	0.0001	-8.3458	-6.6028
	M	-4.6136	0.2858	0.0001	-5.1737	-4.0534
	S	-2.1624	0.2409	0.0001	-2.6346	-1.6902

Table 7. Multinomial Regression Final Model for Diabetes Length of Stay (Overall Model)

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
Age 40-70						
	L	-0.0816	0.2191	0.7095	-0.5112	0.3479
	M	0.1696	0.1504	0.2594	-0.1251	0.4643
	S	0.1268	0.1248	0.3098	-0.1179	0.3715
Age 70-100						
	L	0.6022	0.2217	0.0066	0.1678	1.0367
	M	0.8087	0.1530	0.0000	0.5088	1.1087
	S	0.5091	0.1284	0.0001	0.2575	0.7606
Readmitted > 30						
	L	-0.1647	0.1360	0.2258	-0.4312	0.1018
	M	-0.3349	0.0991	0.0007	-0.5292	-0.1407
	S	-0.1809	0.0927	0.0510	-0.3626	0.0008
Readmitted No						
	L	-0.2657	0.1416	0.0606	-0.5433	0.0119
	M	-0.4175	0.1021	0.0000	-0.6177	-0.2173
	S	-0.2692	0.0951	0.0046	-0.4555	-0.0828
Number of Lab procedures						
	L	0.0249	0.0024	0.0000	0.0212	0.0309
	M	0.0226	0.0016	0.0000	0.0205	0.0271
	S	0.0115	0.0014	0.0000	0.0089	0.0146
Number of Procedures						
	L	0.2944	0.0284	0.0000	0.2352	0.3469
	M	0.1730	0.0231	0.0000	0.1274	0.2182
	S	0.0882	0.0220	0.0001	0.0467	0.1330
Number of medications						
	L	0.2167	0.0078	0.0000	0.2030	0.2340
	M	0.1551	0.0063	0.0000	0.1434	0.1683
	S	0.0892	0.0059	0.0000	0.0767	0.1001
Number of emergency						
	L	-0.0818	0.0447	0.0669	-0.1693	0.0057
	M	-0.0525	0.0291	0.0707	-0.1095	0.0044
	S	-0.0642	0.0266	0.0156	-0.1163	-0.0122
Number of Inpatient						
	L	0.1030	0.0341	0.0025	0.0362	0.1698
	M	0.0807	0.0253	0.0015	0.0310	0.1303
	S	0.0255	0.0239	0.2869	-0.0214	0.0723

Predictors		Estimate	SE	p value	95% CI LL	95% CI UL
Number of diagnosis						
	L	0.2174	0.0346	0.0000	0.1547	0.2913
	M	0.1630	0.0205	0.0000	0.1297	0.2107
	S	0.0797	0.0166	0.0000	0.0511	0.1170
Glucose Serum 0						
	L	0.7844	0.5261	0.1359	-0.2467	1.8156
	M	0.8392	0.3776	0.0263	0.0991	1.5793
	S	0.2415	0.3124	0.4393	-0.3707	0.8537
Glucose Serum 200-300						
	L	1.2042	0.6223	0.0530	-0.0156	2.4239
	M	0.9947	0.4799	0.0382	0.0540	1.9353
	S	0.1597	0.4409	0.7171	-0.7044	1.0239
Glucose Serum None						
	L	0.0533	0.5204	0.9184	-0.9902	1.0501
	M	0.1585	0.3745	0.6721	-0.5798	0.8910
	S	-0.0490	0.3096	0.8742	-0.6644	0.5511
Glucose Serum Norm						
	L	-0.1308	0.6917	0.8500	-1.4864	1.2248
	M	0.3944	0.4462	0.3768	-0.4801	1.2689
	S	-0.0333	0.3741	0.9291	-0.7665	0.6999
Insulin No						
	L	-0.2699	0.1478	0.0679	-0.5595	0.0198
	M	-0.1882	0.1047	0.0722	-0.3934	0.0170
	S	-0.0715	0.0949	0.4508	-0.2575	0.1144
Insulin Steady						
	L	-0.3171	0.1400	0.0236	-0.5915	-0.0426
	M	-0.1880	0.1016	0.0642	-0.3871	0.0111
	S	-0.0088	0.0928	0.9243	-0.1907	0.1731
Insulin Up						
	L	0.4518	0.1442	0.0017	0.1692	0.7344
	M	0.3139	0.1124	0.0052	0.0936	0.5341
	S	0.2274	0.1056	0.0312	0.0205	0.4344
Diabmed Yes						
	L	-0.3613	0.1404	0.0100	-0.6364	-0.0862
	M	-0.2947	0.0922	0.0014	-0.4753	-0.1141
	S	-0.0119	0.0815	0.8837	-0.1716	0.1478
Intercept						
	L	-7.9422	0.6574	0.0000	-9.2306	-6.6537
	M	-5.0616	0.4516	0.0000	-5.9468	-4.1764
	S	-2.3443	0.3727	0.0000	-3.0748	-1.6137

Table 8. Assessment of Fit for Multinomial Regression Model for Diabetes Length of Stay

Models	Measures	Values
Derived	McFadeen R^2	0.1320
	χ^2	3165
	(p-value)	< 0.0001
	AIC	**21734.65**
	BIC	**22235.75**
Overall	McFadeen R^2	0.1297
	χ^2	3186
	(p-value)	< 0.0001
	AIC	21873.73
	BIC	22282.08

We also conduct modelling for classification of short, medium, and long. However, it has been observed with a similar variable selection as in the case of very short/short/medium/long. Also, the variables which are significant and non-significant in this model remains the same with the very short/short/medium/long.

5. CONCLUSION

The study on LOS is important for hospitals because they help in understanding the utilization of resources. The main objective of this work is to model the diabetes length of stay in a hospital which will help in decision making and prediction for the hospital administrators. LOS are mostly treated to be continuous in nature however, this study has considered them to be categorical in nature which is also the interest of hospital administrators. In this study, we have attempted classifying the length of stays as very short, short, medium, and long using quantile classification rather than an arbitrary approach.

The study has highlighted the importance of maintaining medical records as it helps largely in the treatment of patients faster and whereby the length of stay gets reduced. The result identified the impact of medical records showing that MRU have a significant effect on longer stay when compared to MRK patients with a shorter stay. Similarly, it has been observed that the age of a patient has a bearing on the length of stay in hospital. Further, we attempted to validate the variable selection procedure using subsampling approach with varied sample sizes which helped in building the final model.

The present study has proposed the methods for building a multinomial regression model, with additional scope for including more covariates on classifying the stay and its impact on the model. However, this study is also restricted to a stay of 1 to 14 days for diabetes, and the same can be extended to different specialties with varying length of stays. The study has paved the way for similar kind of modelling in different specialties and its impact on LOS. The study underlines the importance of maintaining medical records of patients as it helps in reducing the LOS and can thereby accommodate more patients deserving treatment as inpatients. In general, modeling LOS will be a supportive tool in the optimization of resources for proper health management.

REFERENCES

[1] Agresti A. Foundations of linear and generalized linear models. Wiley Series in Probability and Statistics. 2015.

[2] Austin P.C. A comparison of statistical modelling strategies for analyzing length of stay after CABG surgery. Health Services and Outcome Research Methodology, Volume 3, pp 107-133, 2003.

[3] Beata, S., Deshazo J.P, Gennings. C, Olmo. J.L , Ventura. S, Cios. K. J, and Clore J.N.. "Impact of HBA1c measurement on hospital readmissions rates: analysis of 70,000 clinical database patient records." BioMed Research International, pp 1–11, 2004.

[4] Carter E. M, Potts H. W. Predicting length of stay from an electronic patient record system: A primary total knee replacement example. BMC Medical Informatics and Decision Making. Volume 14- 1, pp 14-26, 2015.

[5] Faddy M, Graves N, Pettitt A. Modelling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. Value in Health. Volume 12-2, pp 309-314, 2009.

[6] Gardiner J. C (2012). Modelling heavy-tailed distributions in healthcare utilization by parametric and Bayesian methods. SAS Glob Forum. pp 1-15, 2012.

[7] Gul M, Guneri A.F. Forecasting patient length of stay in an emergency department by artificial neural networks. Journal of Aeronautics and Space Technologies. Volume 8-2, pp 1-15, 2015.

[8] Harini S, Subbiah M, Srinivasan M R. Fitting length of stay in hospitals using transformed distributions. Communications in Statistics : Case Studies, Data Analysis and Applications. Volume 4 - 1, pp 1-8, 2018.

[9] Harini S, Subbiah M, Srinivasan M R. Fitting of Length of Stay by Multi Stage Classification of Covariates using Transformed Gamma-Pareto Distribution. Journal of Indian Society for Probability and Statistics, in Press.

[10] Jones B, Mcclean S, Stanford D . Modelling mortality and discharge of hospitalized stroke patients using a phase-type recovery model. Healthcare Management Science. Pp 1-19, 2018.

[11] Kembe K.M., Agada P.O., Owuna D. A Queing Model for Hospital Bed Occupancy Management: A Case Study. International Journal of Computational and Theortical Statistics. Volume 1-1 , pp 13-28, 2014.

[12] Meadows K, Gibbens R, Vuylsteke A. Prediction of Patient Length of Stay on the Intensive Care Unit following Cardiac Surgery: A Logistic Regression Analysis Based on the Cardiac Operative Mortality Risk Calculator, EuroSCORE. Journal of Cardiothoracic Vascular Anesthesia, in Press.

[13] Monyai S, Lesaoana M, Darikwa T, Nyamugure P. Application of multinomial logistic regression to educational factors of the 2009 General Household Survey in South Africa 4763. Journal of Applied Statistics, Volume 43 – 1; pp 128-139, 2016.

[14] Papi M, Pontecorvi L, Setola R. A new model for the length of stay of hospital patients. Health Care Management Science. Volume 19 – 1, pp 58-65, 2016.

[15] Shimizutani S, Hiroyuki Y, Haruko Noguchi, Yuichiro Masuda & Masafumi Kuzuya. Exploring the causal relationship between hospital length of stay and re-hospitalization among Japanese AMI patients. Applied Economics, Volume 47 - 22, pp 2307-2325, 2015.

[16] Shukla K, Upadhyay S. Predictive Modelling for Turn Around Time (TAT) of Discharge Process for Insured Patients in a Corporate Hospital of Pune City. Journal of Health Management. Volume 20 - 1, pp 55-63, 2018.

[17] Singler K, Bail HJ, Christ M, Weis P, Sieber C, Heppner HJ, et al.. Correlation of patients age on length of stay and admission rate in a German emergency department]. Dtsch Med Wochenschr. Volume 138 -30, pp 1503–1508, 2013.

[18] Tsai PJ, Chen P, Chen Y, Song Y, Lin H M, Lin F M, et al. Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. Journal of Healthcare Engineering. Article ID 7035463, pp 1-11, 2016.

[19] Udayai, K., & Kumar, P. (2012). Implementing six sigma to improve hospital discharge process. International Journal of Pharmaceutical Research and Science, Volume 3 - 11, pp 4528–4532, 2012.

[20] Verburg I WM , De Keizer NF, De Jonge E, Peek N . Comparison of regression methods for modeling intensive care length of stay. PLoS One. Volume 9-10, pp 1-11. 2014.

[21] Zenga M, Marshall A H, Giordano S. Modelling Students' Length of Stay at University Using Coxian Phase-type Distributions. International Journal of Statistics and Probability. Volume 2 - 1, pp 73 – 89, 2013