Predicting In-Hospital-Death and Mortality Percentage Using Logistic Regression

Steven L Hamilton[1], James R Hamilton[2]

[1]The University of Oklahoma, Oklahoma City, USA
[2]Independent, Alamogordo, USA

Abstract

Logistic regression is an appropriate analysis technique for this CinC Challenge problem. Derived variables from provided patient data records are screened for significance by linear stepwise regression. Screened derived variables and corresponding patient outcome data serve respectively as the predictor and response variables for logistic regression analysis. Each of the two CinC Challenge events use separate logistic regression models, and include limited investigation of non-linear effects. Short descriptions of excursions from the logistic regression approach summarize the scope of the effort.

1. Introduction

Logistic regression is a common analysis technique for situations with binary outcome data. The CinC Challenge problem poses such a situation. Logistic regression also presents a straightforward method to produce predictions of "in-hospital-death", the Event 1 score metric, and "*mortality percentage* (risk)", used in the Hosmer-Lemeshow test statistic for Event 2 scoring. Based on these observations, logistic regression is appropriate [1].

A three phase approach follows. The first phase is selection of derived variables based on the set-a patient data. In the second phase, compute logistic regression models using those derived variables as independent variables and the set-a "In-hospital death (0: survivor, or 1: died in-hospital)" outcome-related descriptor as the dependent variable. Third, apply and refine the logistic regression model(s) to produce Events 1 and 2 scores.

2. Phase 1: derived variables

Descriptions of the seven derived variables used for this work follow, where $Y_i = (Y_0, Y_1, \ldots, Y_j)$ represents the observations for a variable (also referred to as a "parameter" in the CinC Challenge literature) taken from a data file. The observations span times $t_i = (t_0, t_1, \ldots, t_j)$.

First value = Y_0 (1)

Average, $\bar{Y} = \dfrac{\sum_{i=0}^{j} Y_i}{j}$ (2)

Minimum value = $\min(Y_i)$ (3)

Maximum value = $\max(Y_i)$ (4)

Total time = $(t_j - t_0)$. (5)

First difference = $\dfrac{Y_j - Y_0}{t_j - t_0}$ (6)

Last value = Y_j (7)

3. Phase 2: regression methodology

This effort uses the 40 variables described in the CinC Challenge literature (37 time series variables plus gender, height, weight). For each of those 40 variables, the 7 derived variables described above serve as the data for regression analysis. Thus, $40 \cdot 7 = 280$ derived data values represent the content for each patient file. With 4000 set-a patient files, the inputs for regression analysis consist of a 4000 by 280 matrix of predictor variable values, and a 4000 by 1 column vector of outcome values (0 or 1, taken from the set-a Outcome data file).

There are complications with this scheme to determine 7 derived variables for each of the 40 data variables. As expected, the data does not reflect a disciplined data collection policy conducted in accordance with an appropriate experimental design. The derived variables must accommodate this fact. Also, producing 7 derived variables is not viable for some of the patient data variables (e.g. Weight, Age, Gender, and Height). Many occurrences of single entries for a specific variable in a patient data file preclude any of the derived variables requiring multiple observations and/or time stamps. Some data values are clearly wrong (e.g. Age=200, two different-valued data observations at the same time stamp). The policy in these cases assigned zero values to the problematic derived variable element. For example, when a single value for a variable occurs in a patient data file, the 7 element vector for that variable is [(the single observation value) 0 0 0 0 0 0]. The result is inconsistent

population of the 40 7-element spans of each 280 element row, and a relatively sparse, potentially singular matrix situation for which meaningful regression analysis is difficult. After processing all 4000 data files, these zero-value spans within the rows of the 4000 by 280 matrix are populated with average values from the corresponding column according to one of three policies: 1) overall average of meaningful values from the column, 2) average for cases with "in-hospital death = 1", and 3) average for cases with "in-hospital death = 0". After experimentation with each of the three policies, policy 2) serves for the final submission, and is reflected in Figures 1 and 2 results for data set-a.

The MATLAB® "stepwisefit" function identifies the subset of the 280 columns with significant explanatory power. There was no investigation of the Outcome predictions using the linear regression model produced by "stepwisefit". Only the "stepwisefit" screening feature of identifying significant variables was of interest.

Next, the "*n*" columns identified by "stepwisefit" serve as input for the MATLAB® logistic regression "mnrfit" function. The predictor variable matrix is 4000 by *n*, and the response variable remains the 4000 by 1 column vector of 0 or 1 outcome values. The "mnrfit" function produces the logistic regression model coefficients, and the "mnrval" function uses those coefficients to estimate the probability of a "0" outcome (survival), or a "1" outcome (in-hospital death) for a given patient data file. Without further manipulation, the probability of a "1" outcome is the CinC Challenge Event 2 scoring "*mortality percentage* (risk)" value used in the Hosmer-Lemeshow H statistic. Figure 2 shows H statistic results for the logistic regression model used for the final CinC Challenge entry.

Arriving at the Event 1 in-hospital mortality prediction (0: survival, or 1: In-hospital death) for each patient is more complicated than Event 2 scoring. Inspection of the Sensitivity (Se) and positive predictivity ($^{+}$P) terms reveals a tradeoff between false negative (FN) and false positive (FP) predictions. As stated above, the output from "mnrval" is two probabilities: 1) the probability of a "0" outcome, and 2) the probability of a "1" outcome. The key to making an Event 1 in-hospital mortality prediction lies in deciding a threshold value at which either of the two probabilities will indicate a "0" or a "1" outcome. Since the sum of both probabilities is 1, this is equivalent to deciding the outcome prediction based on a ratio of the two probabilities. This work uses such a ratio approach to determine the outcome prediction, and investigate the tradeoff between Se and $^{+}$P. Figure 1 illustrates the process, with the solid gray line showing the score 1 value, at various probability ratios. As with Figure 2, the results presented in Figure 1 reflect the final configuration of the logistic regression model results for data set-a.

4. Phase 3: apply and refine logistic regression model(s)

The regression coefficients comprise the model for the CinC Challenge entry. Model implementation consists of incorporating regression coefficients and derived-data processing code into the "physionet2012.m" script. Separate models are used for each of Event 1 and Event 2.

Logistic regression model refinement included non-linear transformations for the derived variables. This involved copying the original predictor matrix, applying transforms of interest, combining the original and transformed matrices, then using "stepwisefit", "mnrfit", and "mnrval" as described before. Four transformations investigated a limited set on non-linear possibilities. Equations (8) through (11) show the four transformations. Also, several ranges of negative and positive fractional exponent transformations were investigated.

$$Y_{\text{added}} = Y^{2}_{\text{original}} \quad (8)$$

$$Y_{\text{added}} = Y^{-2}_{\text{original}} \quad (9)$$

$$Y_{\text{added}} = \ln(Y)_{\text{original}} \quad (10)$$

$$Y_{\text{added}} = \ln(1/Y)_{\text{original}} \quad (11)$$

The final submission for the CinC Challenge used only fractional exponential transformations to augment the predictor matrix. The fractional exponential policy produced the best set-a score 1 results, with set-a scores correlated with but consistently well below set-b scores for all entries. Several of the non-linear transformation policies produced excellent score 2 results (H < 4) for set-a data, but consistently non-competitive score 2 results for set-b data. Poor comparisons between set-a and set-b scores leads to suspicion that the two data sets may be from different populations, or the 4000 sample size is too small. Otherwise, intuition is that a model developed from a randomly drawn 4000-member sample should more effectively predict outcomes for another random 4000-member sample from the same population. With only ten set-b comparison opportunities available, it was not possible to conduct enough experiments to tailor models to the set-b data. Even if this had been possible it was unknown if the same "bait and switch" situation would occur for set-c. Ultimately unable to achieve a meaningful correlation between set-a and set-b scores 2, attempts to improve score 2 results were abandoned and the final submission selection was based on a policy producing the best score 1 and a good (but not the best observed) score 2. Congratulations to those with the talent and insight to effectively address this situation without relying on luck. Table 1 shows the entire array of derived variables considered, with the "**X**"s marking the derived variables used in the final model for Event 1, with "*"s identifying exponential transformed variables.

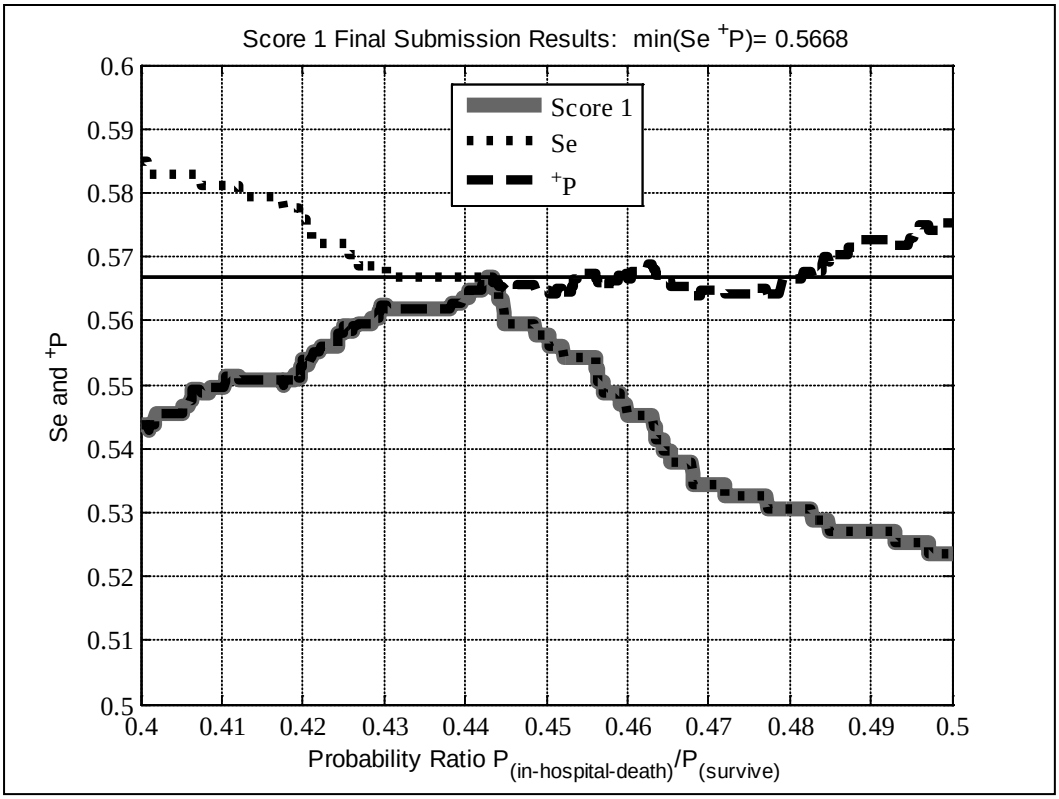

Figure 1. Score 1 Results For Data set-a.

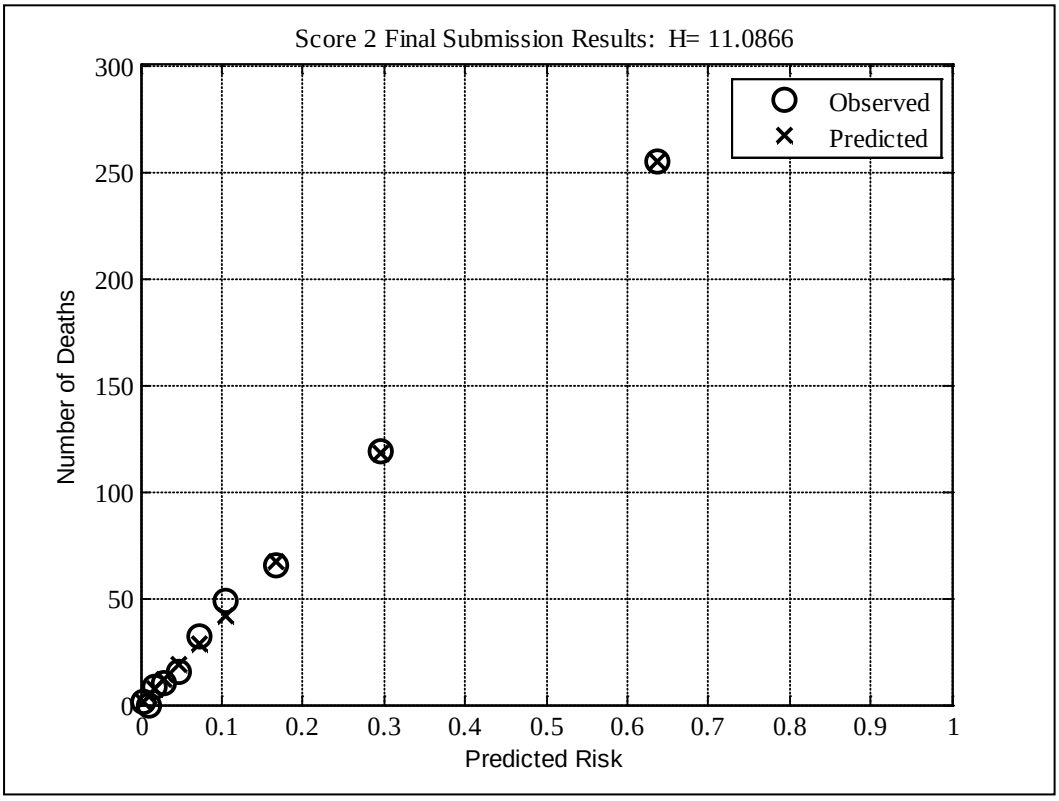

Figure 2. Score 2 Results For Data set-a.

Table 1. Derived variables.

"**X**" means variable is in the final model	Derived Variable Short Description See Section 2, Eqns (1) thru (7)						
	first	avg	min	max	T	diff	last
Albumin				**X***		**X**	
ALP	**X**			**X**			
ALT							
AST			*				
Bilirubin	**X***					**X**	
BUN	**X**						**X**
HCT							
HR				**X**	**X**		**X***
K							
Lactate		**X**		*		*	
Mg							
MAP							
PaCO2		**X**	**X**				**X***
PaO2							
pH		**X**				**X**	
Platelets						**X**	
RespRate	*						
Cholesterol							
Creatinine	*			*			
DiasABP							
FiO2						*	
GCS				**X***		**X***	*
Glucose							**X**
HCO3							
MechVent					**X**		
Na						*	
NIDiasABP							
NIMAP							**X***
NISysABP							
SaO2				*			
SysABP	**X***			**X***			
Temp		**X**			**X**	*	
TroponinI							
TroponinT							
Urine	*			**X***			
WBC							**X***
Weight		(no derived variables by Equations (2) through (7) for these parameters)					
Age	**X***						
Gender							
Height							

– Every derived variable table element shown was screened for significance by "stepwisefit".
– Table elements with "**X**" were basic, non-transformed logistic regression model predictors.
– Table elements with "*" were exponential transformed variables used as predictors.

5. Excursions

Other derived variables besides Equations (1)-(7) were investigated, including a SOFA-Score-based approach as outlined in [2]. The data was not well-suited to produce a consistent stream of SOFA scores for each patient data file, requiring many assumptions and data manipulations similar to those described previously to populate extensive zero-span regions of the predictor variable matrix. In the end, this issue rendered the SOFA score approach actually less effective for the purposes of the two CinC Challenge scoring metrics.

The idea of "predicting" patient mortality, given that the full ICU history for each patient is available suggests the following perspective. By the time the complete data set is collected, the outcome is known since the data stream ends with a "0" or a "1" outcome. A more interesting "prediction" problem may be to make the same "0" or "1" prediction based on a more restricted data situation of, for example, only the five general descriptors and the "first" values as described in Equation (1). This excursion, including non-linear transformations described earlier, yielded score 1 = 0.4116, and score 2 = 8.843.

Finally, other analysis techniques were attempted with results inferior to the logistic regression approach. Two-group discriminant analysis using the MATLAB® "classify" function was not effective because of too much overlap between the data describing the "0" outcome group and the "1" outcome group, even when using a discriminant function that employed the prior probabilities associated with the set-a "survivor" and "in-hospital-death" groupings. Two hypothesis testing approaches using each of the MATLAB® Kolmogorov-Smirnov (KS) "kstest" and "kstest2" functions suffered from too little data for each parameter for each patient data file to test against the "survivor" or "in-hospital-death" populations defined with the set-a data. Logistic regression proved superior to all of these approaches.

References

[1] Hosmer DW, Lemeshow S. Applied Logistic Regression, Second Edition, New York, John Wiley & Sons, Inc. 2000:1-7.

[2] Ferreira FL, Bota DP, Bross A, Melot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients.

Addresses for correspondence.

Steven L. Hamilton
711 Stanton L Young Blvd, Rm 524
Oklahoma City, OK 73104
steven.l.hamilton@gmail.com

Jim Hamilton
PO Box 3
High Rolls, NM 88325
jimhamilton@q.com