Apr 16th, 10:45 AM - 12:00 PM

Analysis of Relapse in Leukemia Patients With Missing Data Using an Extension of the EM Algorithm

Braydon Schaible
Georgia Southern University, bs05313@georgiasouthern.edu

Analysis of Relapse in Leukemia Patients With Missing Data Using the EM-Algorithm

Braydon Schaible, MPH, Lili Yu, PhD
Biostatistics

GEORGIA SOUTHERN UNIVERSITY
JIANN-PING HSU COLLEGE OF PUBLIC HEALTH

INTRODUCTION

1. Leukemia is a type of cancer. In 2014, over 18,000 people were diagnosed with AML and approximately 6,000 people were diagnosed with ALL (Leukemia & Lymphoma Society, 2015). The overall survival rates over a 5 year period for patients diagnosed with ALL is 70% and 25.4% for patients diagnosed with AML (Leukemia & Lymphoma Society, 2015). Relapse rates for AML vary from 33% to 78% depending on the patients risk classification (good, intermediate, poor) (Grimwade et al, 1998). For children originally diagnosed with ALL who have achieved complete remission, there is a 15%-20% chance of relapse (Dana-Farber Boston Children's, 2015).
2. Leukemia has high relapse rates, which will reduce the overall survival time for patients and lower their quality of life. Therefore, it is important to do research on the relapse of the Leukemia patients.
3. We will analyze a dataset that includes information for 137 bone marrow transplant patients with leukemia. From the statistical analysis we will find significant factors that can affect the relapse, in order to find ways for clinicians to make better treatment plans.

PURPOSE

➢The purpose of this study was to find significant factors that can affect the relapse of Leukemia patients. The dataset contained 54 observations which had missing response variable (relapse variable) data. Therefore, we propose a new statistical methodology to impute the missing binary response variable, in which we assume the missing data mechanism for this dataset is assumed to be missing at random (MAR).

METHODS

1. We use the Expectation-Maximization (EM) algorithm. In the expectation step, we use the mean predicted probability as the cut-point to impute the missing binary response variable. In the maximization step, we use logistic regression to find the parameter estimates of the coefficients and then the significant factors. There are three steps to the EM logistic regression model.
(1) The first step, the initial step, requires us to obtain initial parameter values. These values are obtained by running a logistic regression model, using only the observations where there is no missing data, containing all variables of possible interest.
(2) Next, the probability of relapse is calculated, based on the model obtained in the initial step, for all observations (including those with missing relapse indicator variable). Then, based on the cut-point determined by calculating the mean of the predicted probabilities, the data for the observations with missing relapse indicator variable are imputed in the expectation step.
(3) If the probability of relapse is greater than the specified cut-point, then the imputed value is 1 (relapse), otherwise the imputed value is 0 (no relapse). A logistic regression is then run on the imputed data which returns a set of parameter values which are then used in the next iteration of the algorithm (this is the maximization step) (Anderson & Hardin, 2014). The algorithm iterates between step (2) and step (3) until the parameter values converge.

2. We do model selection to select the significant factors that can affect the relapse of Leukemia patients. We use backward elimination to delete non-significant factors one by one.

RESULTS

Results from the EM algorithm for full model

Leukemia Relapse EM-Algorithm Parameter Estimates – Model Containing All Independent Variables

Iteration	Intercept	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z10	Cut Point
Initial Estimates	-0.1468	-0.0046	-0.0129	-0.2509	0.3626	0.6262	0.0371	-0.0004	0.8024	0.069	N/A
1	-0.8481	0.0079	-0.0334	-0.2243	0.8747	1.3752	-0.0691	-0.001	1.5843	0.1185	0.4845
2	-1.0739	0.0242	-0.0426	-0.1596	0.9302	1.4069	-0.01198	-0.001	1.5266	0.0067	0.4599
3	-1.2335	0.0267	-0.0469	-0.0686	1.0346	1.4674	0.0357	-0.0009	1.4358	0.0222	0.4672
4	-1.2335	0.0267	-0.0469	-0.0686	1.0346	1.4674	0.0357	-0.0009	1.4358	0.0222	0.4599

Analysis of Maximum Likelihood Estimates – Model Containing All Independent Variables

Parameter	DF	Estimate	Standard Error	P-Value
Intercept	1	-1.2335	0.8142	0.129781
Z1	1	0.0267	0.0329	0.416916
Z2	1	-0.0469	0.0305	0.123748
Z3	1	-0.0686	0.4201	0.870345
Z4	1	1.0346	0.4324	0.016726
Z5	1	1.4674	0.4599	0.001418
Z6	1	0.0357	0.4424	0.935747
Z7	1	-0.0009	0.0008	0.238047
Z8	1	1.4358	0.4454	0.001267
Z10	1	0.0222	0.4802	0.963084

Results from model selection

Iteration	Intercept	Z4	Z5	Z8	Cut Point
Initial Estimates	-0.875	0.3796	0.5808	0.8528	N/A
1	-1.9965	1.0190	1.3956	1.6251	0.484453
2	-2.114	1.1285	1.4661	1.5257	0.4744526
3	-2.114	1.1285	1.4661	1.5257	0.4671533

Model: Relapse=int + Z4 + Z5 + Z8
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	P-Value
Intercept	1	-2.114	0.4591	4.14E-06
Z4	1	1.1285	0.4229	0.00762
Z5	1	1.4661	0.4003	0.00025
Z8	1	1.5257	0.4332	0.000429

PUBLIC HEALTH SIGNIFICANCE

➢This is a new, original method that can be extended to analyzing other disease datasets where there are missing data for a categorical response variable and the missing data mechanism is assumed to be MAR.
➢In this case, it was found that Donor Sex, Patient CMV Status, and FAB Grade together create a significant model for predicting relapse in leukemia patients with AML or ALL.
➢Knowing about the significant predictors of leukemia relapse, medical professionals can create better, more knowledgeable treatment plans for different patients.
➢The results of this study can be used to develop health education programs and interventions targeting physicians to increase their awareness of the possible factors affecting leukemia relapse.
➢Policy officials should be aware of some of the possible factors contributing towards leukemia relapse and construct their policies accordingly.

STRENGTHS & LIMITATIONS

Limitations:
➢Data was not collected regarding the type of leukemia (AML vs ALL), it is only known that the data contains information from patients diagnosed with AML or ALL. There are very different relapse rates for the two types of acute leukemia and this is not taken into account.
➢The method used was only applied to those patients diagnosed with AML or ALL that had already received a bone marrow transplant.
➢The method is applicable only to datasets where the missing data mechanism is missing at random (MAR) and the missing data is in a categorical response variable.
➢The sample size is small.
Strengths:
➢This method adds to an already successful method applied by Anderson and Hardin (Anderson & Hardin, 2014) by selecting an optimal cut point at each iteration, rather than using the same cut point throughout the process.
➢The method used in this study incorporates the use of a likelihood method for imputation of missing data where the missing data mechanism is MAR.
➢This method showed faster convergence than using just the EM algorithm alone on this specific dataset (4 iterations for this method compared to 5 iterations using the same cut point at each iteration).

CONCLUSIONS

➢This study used the EM algorithm, along with selecting the cut-point based on the mean of the predicted probabilities at each iteration, to impute missing data for the relapse indicator variable. Based on the final model, it was found that the variables Z4 (Donor Sex: 1-Male, 0-Female), Z5 (Patient CMV Status: 1-CMV Positive, 0-CMV Negative), and Z8 (FAB Grade: 1-FAB Grade 4 or 5 & AML, 0-Otherwise) are significant factors when predicting leukemia relapse. This method can be applied to many other studies on diseases other than leukemia as long as the missing data mechanism is MAR.

REFERENCES

➢Disease Information & Support, Facts & Statistics. 2015, March 18. The Leukemia & Lymphoma Society. Retrieved from http://www.lls.org/diseaseinformation/getinformationsupport/factsstatistics/.
➢Leukemia Information. 2015, March 18. Cancer Treatment Centers of America. Retrieved from http://www.cancercenter.com/leukemia/learning/.
➢Relapsed Acute Lymphoblastic Leukemia. 2015, March 20. Dana-Farber Boston Children's. Retrieved from http://www.danafarberbostonchildrens.org/conditions/leukemia-and-lymphoma/relapsed-acute-lymphoblastic-leukemia.aspx.
➢Grimwade, David et al. "The Importance of Diagnostic Cytogenetics on Outcome in AML: Analysis of 1,612 Patients Entered Into the MRC AML 10 Trial." *Blood* 92.No. 7 (October1) (1998): 2322-333.
➢Anderson, Billie, and J. Michael Hardin. "Modified Logistic Regression Using the EM Algorithm for Reject Inference." *International Journal of Data Analysis Techniques and Strategies* 5.4 (2014): 359-73. Print.