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

Braydon Schaible
Georgia Southern University, bs05313@georgiasouthern.edu

Follow this and additional works at: https://digitalcommons.georgiasouthern.edu/research_symposium
Part of the Epidemiology Commons

Recommended Citation
https://digitalcommons.georgiasouthern.edu/research_symposium/2016/2016/6

This presentation (open access) is brought to you for free and open access by the Conferences & Events at Digital Commons@Georgia Southern. It has been accepted for inclusion in Georgia Southern University Research Symposium by an authorized administrator of Digital Commons@Georgia Southern. For more information, please contact digitalcommons@georgiasouthern.edu.
Analysis of Relapse in Leukemia Patients With Missing Data Using the EM-Algorithm

Braydon Schaible, MPH, Lili Yu, PhD
Biostatistics

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% 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.

RESULTS

Results from the EM algorithm for full model

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

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Intercept</th>
<th>Z1</th>
<th>Z2</th>
<th>Z3</th>
<th>Z4</th>
<th>Z5</th>
<th>Z6</th>
<th>Z7</th>
<th>Z8</th>
<th>Z10</th>
<th>Cut Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Estimates</td>
<td>0.1348</td>
<td>0.0904</td>
<td>0.0129</td>
<td>0.2599</td>
<td>0.3265</td>
<td>0.0562</td>
<td>0.0771</td>
<td>0.0604</td>
<td>0.0824</td>
<td>0.068</td>
<td>N/A</td>
</tr>
<tr>
<td>1</td>
<td>-1.8791</td>
<td>0.0324</td>
<td>0.0438</td>
<td>0.3992</td>
<td>1.1949</td>
<td>0.3009</td>
<td>0.0116</td>
<td>0.0015</td>
<td>1.3556</td>
<td>0.0465</td>
<td>0.4395</td>
</tr>
<tr>
<td>2</td>
<td>-1.3135</td>
<td>0.0267</td>
<td>0.0409</td>
<td>0.0654</td>
<td>1.5456</td>
<td>1.4874</td>
<td>0.0573</td>
<td>0.0005</td>
<td>1.4355</td>
<td>0.0222</td>
<td>0.4372</td>
</tr>
<tr>
<td>3</td>
<td>-1.3135</td>
<td>0.0267</td>
<td>0.0409</td>
<td>0.0654</td>
<td>1.5456</td>
<td>1.4874</td>
<td>0.0573</td>
<td>0.0005</td>
<td>1.4355</td>
<td>0.0222</td>
<td>0.4372</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates - Model Containing All Independent Variables

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>z-value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-1.3315</td>
<td>0.1842</td>
<td>0.0001</td>
<td>0.2597</td>
</tr>
<tr>
<td>Z1</td>
<td>1</td>
<td>0.5727</td>
<td>0.4125</td>
<td>0.0015</td>
<td>0.1099</td>
</tr>
<tr>
<td>Z2</td>
<td>1</td>
<td>-0.0449</td>
<td>0.0128</td>
<td>0.0015</td>
<td>0.1194</td>
</tr>
<tr>
<td>Z3</td>
<td>1</td>
<td>-0.0848</td>
<td>0.2081</td>
<td>0.0015</td>
<td>0.1734</td>
</tr>
<tr>
<td>Z4</td>
<td>1</td>
<td>1.5464</td>
<td>0.2497</td>
<td>0.0015</td>
<td>0.0755</td>
</tr>
<tr>
<td>Z5</td>
<td>1</td>
<td>0.7515</td>
<td>0.5119</td>
<td>0.0015</td>
<td>0.1194</td>
</tr>
<tr>
<td>Z6</td>
<td>1</td>
<td>0.5397</td>
<td>0.4424</td>
<td>0.0015</td>
<td>0.1734</td>
</tr>
<tr>
<td>Z7</td>
<td>1</td>
<td>0.0599</td>
<td>0.0044</td>
<td>0.0015</td>
<td>0.7063</td>
</tr>
<tr>
<td>Z8</td>
<td>1</td>
<td>1.3752</td>
<td>0.0445</td>
<td>0.0015</td>
<td>0.0755</td>
</tr>
<tr>
<td>Z10</td>
<td>1</td>
<td>0.0222</td>
<td>0.0462</td>
<td>0.0015</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

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


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 (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).

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).

Weaknesses:
- The sample size is small.
- The method is new and original.

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 was used 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.

Sample Size:
- The sample size is small.

Statistical Methodology:
- 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.