On Effective Prediction of Preventable Hospital Readmission

Ching-Wei Cheng and Lingsong Zhang
Department of Statistics

Motivation

- National 30-day readmission rate averaged 19% in 2007–2011 (MMRR, 2013)
- Centers for Medicare & Medicaid Services (CMS) charged a total of 2,213 hospitals with high readmission rates about $280 million in readmission penalties in 2013
- Reducing preventable readmission is crucial in healthcare engineering
- To build an efficient predictive model to identify high-risk patients for readmission
  - Which covariates are more important than others?  ➔ Dominance analysis
  - How often shall we update the predictive models for future risks?  ➔ Area under the curve (AUC) of the receiver operating characteristic (ROC) curve

Data

- A healthcare system with 11 hospitals at Florida; 8 of them are included in analysis (excluding children, women and long-term care hospitals)
- 7.5 years long; 462,878 inpatient records
- 30-day readmissio ns as response:
  - \( r_{30} = \mathbb{I}\{\text{time since previous discharge} \leq 30\} \)
- Covariates of interest:

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>length_of_stay</td>
<td>numeric</td>
</tr>
<tr>
<td>admission_type</td>
<td>factor</td>
</tr>
<tr>
<td>discharge_disposition</td>
<td>factor</td>
</tr>
<tr>
<td>age</td>
<td>numeric</td>
</tr>
<tr>
<td>race</td>
<td>factor</td>
</tr>
<tr>
<td>sex</td>
<td>factor</td>
</tr>
<tr>
<td>language</td>
<td>factor</td>
</tr>
<tr>
<td>payer_class</td>
<td>factor</td>
</tr>
<tr>
<td>marital_status</td>
<td>factor</td>
</tr>
<tr>
<td>disease_severity_index</td>
<td>factor</td>
</tr>
<tr>
<td>marital_status</td>
<td>factor</td>
</tr>
<tr>
<td>race</td>
<td>factor</td>
</tr>
<tr>
<td>length_of_stay</td>
<td>numeric</td>
</tr>
</tbody>
</table>

- Logistic regression is applied for predictive modeling
- The data is partitioned as monthly segments
  - First-year (first 12 segments) as training data
  - The rest monthly segments as test data

Relative Importance in the Explanatory Variables

**Dominance analysis for linear regression model**

\[ \hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \]

(Azen and Budescu, 2003, 2006)

- The model performance can be measured by the portion of variation explained by the model

\[ R^2 = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}} \]

- General dominance determines relative importance is based on the average contribution in \( R^2 \) observed by adding a predictor to all possible subsets of the remaining predictors, i.e.

\[ \text{dom}(X_i) = \frac{1}{|V|} \sum_{V \subseteq X_i \cup X_j \in V} (R_i^2 - R_j^2), \]

where \( V \) is the power set of \( \{X_i, i = 1, \ldots, p\} \) and \( |V| \) is the cardinality of the set \( V \).

**Dominance analysis for logistic regression model**

\[ \logit(p(X)) = \log \left( \frac{\hat{p}(X)}{1 - \hat{p}(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \]

(Azen and Traxel, 2009)

- The former \( R^2 \) measure is NOT appropriate here
  - For logistic regression models, define
  
  \[ R^2 = 1 - \frac{\log(L_0)}{\log(L_M)} \]
  
  - \( L_0 \) denotes the likelihood of the null (intercept only) model
  - \( L_M \) denotes the likelihood of the fitted (intercept and predictors) model

**Evaluation of Predictive Modeling**

- **AUC statistic of the ROC curve** is applied to evaluate the predictive efficacy of the logistic model
- **Bootstrap tests**
- **Concluding Remarks**
- **Future Work**