USING THE SAS® SYSTEM TO DEVELOP RISK PREDICTION MODELS FOR PATIENT RE-ADMISSION REDUCTION IN VAMCs

Issac Shams

Veteran Engineering Resource Center (VERC)-VA-CASE Healthcare Systems Engineering Research Group
Department of Industrial and Systems Engineering
Wayne State University

MISUG, October 2012
Overview

• Motivation and Statement

• Preliminary Analysis

• Modeling Strategy and details

• Result and outcomes
**Background and Motivation**

- 17 percent of Medicare patients discharged from the hospitals have a *readmission (rehospitalization)* within 30 days of discharge, accounting for $15 billion in spending (Medicare Payment Advisory Commission Report, 2007).

- Under Obama Care Rule (Patient Protection and Affordable Care Act or PPACA), about two-thirds of U.S. hospitals stand to be penalized for excess readmissions starting Oct. 1, 2012.

- Based on the formula developed by CMS (Centers for Medicare & Medicaid Services), this will account for 2,211 hospitals a cumulative $280 million cut in Medicare funds (1%) due to high rates of 30-day readmission after discharge. For hospitals that don’t improve, penalties will grow to a maximum of 2% for the 2014 program year and 3% for 2015 (American Medical News, Aug. 27, 2012).
DEFINITION AND CATEGORIZATION

• Readmission (in VAMCs):

The proportion of patients who were readmitted (for any cause) to the acute care wards of the hospital within 30 days following discharge from the hospital.

\[
\frac{\# \text{ 30 Day Readmits}}{\# \text{ Discharges in 30 Day Window}}
\]

• Some categories of potential readmission:

- Unavoidable
  - Inappropriate
  - Appropriate
- Avoidable
  - Inappropriate
  - Appropriate
  - Adverse Event
  - Poor Quality Care
Statements and Objectives

- Data were collected from John D. Dingell VAMC during September and October 2006 through 2011.

- There were 3108 records, 2449 patient id (subject), and 50 prognostic factors. Except for ‘DOB’, ‘admdate’ and ‘disdate’, all others were in nominal scale.

- Investigate the relationships between set of patient prognostic factors and his/her readmission risk.

- Utilize the predictive models to filter out significant variables for internally categorizing patients in terms of their hazard to rehospitalization and adjust the budget allocation accordingly.
**Preliminary Steps**

- Data collection and extraction using **SAS SQL**
- Data structure and exploration using **Base SAS**

```sas
PROC FREQ <DATA= >;
  TABLES requests < / chisq> ;
RUN;
```

**Results**: 94.72% male, 89.74% without insurance, 99.26% veterans, 96.59% inpatients, etc.

‘Admsource’ ➔ 93.85% Hospital, 6.15% NHCU

Tentative clusters of features:

1. Demographic  
2. Financial Situation  
3. War-Related  
4. Admission-Related  
5. Health Care Related
**Preliminary Steps (Cont’d)**

- Data Quality problem handling with **Base SAS**
  1. **Duplications**

```sas
DATA readmission;
MERGE radmsep06--radmoct11;
BY ssn;
run;
```

2. **Outliers** (for ‘los’ only)

```sas
PROC MEANS < DATA= > <min max median std p1 p5 p95 p99> ;
VAR variable(s) < >;
run;
```

3. **Missing Values** (again with **PROC FREQ**)

High-missing-value attributes:

‘agentl’(93.27%), race(52.70%), ‘pct’(24.55%), ‘providerpclass’(22.10%)
Data Pre-processing

• Data cleansing with Base SAS
  ‘viet’=YES and ‘vet’=NO ! — ‘opatientind’=YES and ‘pows’=YES !

• Feature creation with Base SAS

los = admdate – disdate;

disage = (disdate – dob)/365.25;

seq: if first.id then seq=1; else seq+1;

rtime:
data want (drop=_:);
set have;
set have (firstobs = 2 keep = admdate rename = (admdate = _admdate2))have(obs = 1 drop = _all_);
rtime=ifn(missing(_admdate2) or last.id,31,_admdate2-disdate);
run;

rstatus: if rtime=31 then rstatus=0; else rstatus=1;
Analysis Approaches

Since
1. The response variable ‘rtime’ is discrete and display timing of event (readmission),
2. There were lots of (near 88%) patients who had not re-admitted, and we did not want to discard all of them, and
3. There were so many repeated admissions for some specific patients; we used survival analysis in SAS/STAT.

‘rtime’ is a response: [0,31]
‘rstatus’ is 0-1 censoring ind. (right censored, Singly Type I)
Origin of time → ‘disdate’

- Relevant procedures in SAS:
  lifetest, lifereg, phreg, nlmixed
Nonlinear Mixed Effect Model aka Hierarchical Nonlinear Model

- Widely used in pharmacokinetics (Row; 1997), HIV Dynamics (Wu and Ding; 1999), Forestry (Fang and Bailey; 2001) and etc.

- Response evolves over *time, within* individual profiles of *repeated events* and these do *vary* in the cohort.

- *Inter-patient variations* of re-admission process and how these *vary* across subjects (unique SSN) can, *simultaneously*, be elucidated.

- Both *fixed* and *random* effects can enter *nonlinearly*, and response (given the random effects) can have lots of conditional distributions.

- Inference is based on *marginal likelihood* and maximizing its approximation over the random effects.
**MODELING FRAMEWORK**

- $y_i$: observed data vector for subject $i = 1, \ldots, s$
- $u_i$: latent random vector for within-subject covariance

  $y_i$ and $u_i$ are independent across $i$

- There is an appropriate model linking $y_i$ and $u_i$ which leads to joint density function

  $$p(y_i|X_i, \phi, u_i) q(u_i|\xi)$$

- $X_i$: matrix of observed explanatory variables
- $\phi$ and $\xi$ are vectors of unknown parameters

Let $\theta = [\phi, \xi]$, then inference on $\theta$ is based on the marginal likelihood function

$$m(\theta) = \prod_{i=1}^{s} \int p(y_i|X_i, \phi, u_i) q(u_i|\xi) \, d\, u_i$$

Then $f(\theta) = -\log m(\theta)$ is minimized over $\theta$ numerically in order to estimate $\theta$, and the inverse Hessian matrix at the estimates provides an approximate variance-covariance matrix for the estimate of $\theta$. 
Hazard for the $j$th re-admission for $i$th patient at time $t$ is governed by 

$$\log h_{ij}(t) = \alpha(t) + \beta X_{ij} + \varepsilon_i$$

$\varepsilon_i$ represents unobserved heterogeneity, and is subscripted by $i$ not by $j$. It has, here, a normal density with a mean of 0 and a variance $\sigma^2$.

When events are repeated, such models are not highly sensitive to the choice of a distribution for $\varepsilon$ (Kelin, 1992; McGilchrist, 1993).

$\alpha(t)$ represents the baseline hazard function, describes how the risk of re-admission changes over time at baseline levels of covariates ($\beta = 0$).

$\beta$ represents the effect parameters, uncover how the re-admission risk varies in response to explanatory covariates.
SETTIGNs AND IMPLEMENTATIONS

NLMIXED procedure

- It has MLE of non-linear mixed models.
- In MODEL statement, there are bunches of conditional distribution to choose or you can define a general log likelihood function.
- It’s better to have some initial guesses for parameter estimates in PARMS.
- It has different integral approximation methods as well as lots of optimization techniques for dealing with minimizing the marginal log-likelihood function.

How to find out $\alpha(t)$?

1. Non-parametric: KDE with bootstrap, NPEB (Robbins, Herbert; 1956)
2. Parametric:

   - Exponential: $ \log h(t) = \mu + \beta X$  
   - Gompertz: $ \log h(t) = \mu + \alpha t + \beta X$  
   - Weibull: $ \log h(t) = \mu + \alpha \log t + \beta X$
How to implement Frailty model into HNLMMs framework? (Allison, 2010)

Let \( t_i \) be the event (or censoring) time, and \( \delta_i \) be the censoring indicator

Define \( \lambda_i = \exp(\beta X_i + \varepsilon_i) \)

Based on the Weibull baseline hazard, the single patient’s contribution to the log-likelihood function (condition on \( \varepsilon \)) is:

\[
\log L_i = -\lambda_i t_i^{(\alpha+1)} + \delta_i \{\log(\alpha + 1) + \alpha \log t_i + \log \lambda_i\}
\]

and \( \alpha = 0 \) for the exponential distribution.

In SAS we have:

\[
\text{ll} = - \text{lambda} \times \text{nrtime}\^{(\alpha+1)} + \text{rstatus} \times (\log(\alpha + 1) + \alpha \times \log(\text{nrtime}) + \log(\text{lambda}));
\]

\textbf{MODEL} \text{ nrtime} \sim \text{GENERAL(ll)};
**DIFFICULTIES AND SOLUTIONS**

**D1.** Near all covariates were categorical and in nominal scale, and proc nlmixed, until now, does not have CLASS statement.

**S1.** Generate a non-singular one with PROC LOGISTIC.

```plaintext
PROC LOGISTIC data=< > outdesign=< > outdesignonly;
CLASS admsource(ref='NHCU') mar(ref='NEVER MARRIED') userenrollee (ref='NO')
pheartind(ref='YES');
```

**D2.** Overspecification: ‘pct’ had 54 values, ‘DRG’ had 87 values, ‘pdiagnosis’ had 65 values, etc. In a fully specified survival model, there should be 10-15 outcomes (readmission) per degree of freedom (Harrell, 2010). So we needed more than 2000 readmissions in the data set but we had only 372.

**S2.** Combine levels of ‘pdiagnosis’ to a higher level and transform ‘DRG’ to Major Disease Classification ‘mdc’, based on ICD09 index.

- **DRG:** 570-585 and 592-607 belongs to MDC09, then new levels are 25.
- **‘pdiagnosis’:** 715.96, 541, 250.12, etc. are associated with ‘DISEASES OF THE BLOOD AND BLOOD-FORMING ORGANS’, resulting in only 15 new levels.
**Difficulties and Solutions (Cont’d)**

**D3.** Model-selection methods (like Forward Selection, Backward Elimination, etc.) are not provided in proc nlmixed.

**S3.** Since there were no continuous attributes, Correlation Feature Selection (CFS) and minimum-redundancy-maximum-relevance (mRMR) techniques were used to shrink the pool of attributes to less than 10.

**A.** Explore class variables interactions and continuous-by-class interactions

```plaintext
PROC FREQ <data= >;
tables (admsource ward--elig enrollp enrolls)*(pct--pdiagnosis providerpclass) / chisq;
run;

PROC GLM <data= >;
class mtest;
model disage=mtest/solution;
run;
```

**B.** Drop redundant effects that carries the same meaning and functionality

‘enrollp’ is determined by ‘elig’; ‘enrolls’ is digged out from ‘mtest’ and ‘elig’
DIFFICULTIES AND SOLUTIONS (CONT’D)

D4. The final Hessian matrix is full rank but has at least one negative eigenvalue. **Second-order optimality** condition violated.

S4. There are some suggestions provided ([SAS Global Forum 2012; Paper 332-2012](#)):  
1. Rescale your data so the values are on a similar scale

\[
\text{slos} = (\text{los} - 7.653) / 2.8814713; \\
\text{nrttime} = \text{rtime} / 10;
\]

2. Reparameterize the model. Use a standard deviation rather than the variance.

\[
\text{RANDOM} \ e \sim \text{normal}(0, \text{sd}\ast\text{sd}) \ \text{SUBJECT}=\text{id};
\]

D5. Find the most appropriate functional form of each effect.

S5. Did some plotting efforts and best-guess approaches. For example ‘sloglos’ is used instead of ‘los’.
## Results

### The NL MIXED Procedure

**Specifications**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>WORK.X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>nrtime</td>
</tr>
<tr>
<td>Distribution for Dependent Variable</td>
<td>General</td>
</tr>
<tr>
<td>Random Effects</td>
<td>e</td>
</tr>
<tr>
<td>Distribution for Random Effects</td>
<td>Normal</td>
</tr>
<tr>
<td>Subject Variable</td>
<td>id</td>
</tr>
<tr>
<td>Optimization Technique</td>
<td>Dual Quasi-Newton</td>
</tr>
<tr>
<td>Integration Method</td>
<td>Adaptive Gaussian Quadrature</td>
</tr>
</tbody>
</table>

**Note:** GCONV convergence criterion satisfied

### Fit Statistics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>2973.9</td>
</tr>
<tr>
<td>AIC (smaller is better)</td>
<td>2987.9</td>
</tr>
<tr>
<td>AICC (smaller is better)</td>
<td>2987.9</td>
</tr>
<tr>
<td>BIC (smaller is better)</td>
<td>3028.5</td>
</tr>
</tbody>
</table>
**RESULTS**

*PARMS*  

\[
\begin{align*}
\text{b0} &= 0 \\
\text{badmsource} &= 0 \\
\text{bseqadmsource} &= 0 \\
\text{bseqsloglos} &= 0 \\
\text{b1maruser} &= 0 \\
\text{b2maruser} &= 0 \\
\text{sd} &= 1 \\
\end{align*}
\]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>St. Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Lower</th>
<th>Upper</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>b0</td>
<td>-8.7573</td>
<td>0.4963</td>
<td>2443</td>
<td>-17.65</td>
<td>&lt;.0001</td>
<td>-9.7304</td>
<td>-7.7841</td>
<td>0.000423</td>
</tr>
<tr>
<td>badmsource</td>
<td>4.7175</td>
<td>0.4359</td>
<td>2443</td>
<td>10.82</td>
<td>&lt;.0001</td>
<td>3.8628</td>
<td>5.5723</td>
<td>0.000473</td>
</tr>
<tr>
<td>bseqadmsource</td>
<td>-1.1979</td>
<td>0.1176</td>
<td>2443</td>
<td>-10.19</td>
<td>&lt;.0001</td>
<td>-1.4284</td>
<td>-0.9673</td>
<td>0.000622</td>
</tr>
<tr>
<td>bseqsloglos</td>
<td>-0.1795</td>
<td>0.0533</td>
<td>2443</td>
<td>-3.37</td>
<td>0.0008</td>
<td>-0.284</td>
<td>-0.07495</td>
<td>-0.00064</td>
</tr>
<tr>
<td>b1maruser</td>
<td>-0.1581</td>
<td>0.1464</td>
<td>2443</td>
<td>-1.08</td>
<td>0.2805</td>
<td>-0.4452</td>
<td>0.1291</td>
<td>-0.0001</td>
</tr>
<tr>
<td>b2maruser</td>
<td>0.3321</td>
<td>0.1229</td>
<td>2443</td>
<td>2.7</td>
<td>0.0069</td>
<td>0.0911</td>
<td>0.573</td>
<td>-0.00037</td>
</tr>
<tr>
<td>sd</td>
<td>2.7614</td>
<td>0.1794</td>
<td>2443</td>
<td>15.39</td>
<td>&lt;.0001</td>
<td>2.4096</td>
<td>3.1132</td>
<td>0.000738</td>
</tr>
</tbody>
</table>

✓ ‘sd’→ There is definitely an unobserved heterogeneity across patients, or there is undoubtedly high dependence among repeated re-admissions.

✓ ‘badmsource’→ Hazard of re-admission, controlling for other covariates, for those admitted in ‘Hospital’ is only about 4.7% of the hazard for those admitted in ‘NHCU’
**Results (Cont’d)**

### Empirical Correlation Matrix of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>b0</th>
<th>badmsource</th>
<th>bseqadmsource</th>
<th>bseqsloglos</th>
<th>b1maruser</th>
<th>b2maruser</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>b0</td>
<td>1</td>
<td>-0.884</td>
<td>0.3763</td>
<td>0.08008</td>
<td>0.0191</td>
<td>-0.09193</td>
<td>-0.627</td>
</tr>
<tr>
<td>badmsource</td>
<td>-0.884</td>
<td>1</td>
<td>-0.512</td>
<td>-0.06851</td>
<td>-0.00529</td>
<td>0.0136</td>
<td>0.3243</td>
</tr>
<tr>
<td>bseqadmsource</td>
<td>0.3763</td>
<td>-0.512</td>
<td>1</td>
<td>0.1232</td>
<td>0.0138</td>
<td>-0.01987</td>
<td>-0.431</td>
</tr>
<tr>
<td>bseqsloglos</td>
<td>0.08008</td>
<td>-0.06851</td>
<td>0.1232</td>
<td>1</td>
<td>0.02835</td>
<td>0.009294</td>
<td>-0.0876</td>
</tr>
<tr>
<td>b1maruser</td>
<td>0.0191</td>
<td>-0.00529</td>
<td>0.0138</td>
<td>0.02835</td>
<td>1</td>
<td>-0.3486</td>
<td>0.00092</td>
</tr>
<tr>
<td>b2maruser</td>
<td>0.09193</td>
<td>0.0136</td>
<td>-0.01987</td>
<td>0.009294</td>
<td>-0.3486</td>
<td>1</td>
<td>0.00578</td>
</tr>
<tr>
<td>sd</td>
<td>-0.627</td>
<td>0.3243</td>
<td>-0.431</td>
<td>-0.08761</td>
<td>0.000919</td>
<td>0.005782</td>
<td>1</td>
</tr>
</tbody>
</table>

### Contrast (Delta Method; Cox 1998)

<table>
<thead>
<tr>
<th>Label</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>mar*userenrollee</td>
<td>2</td>
<td>2443</td>
<td>4.14</td>
<td>0.0160</td>
</tr>
</tbody>
</table>
How well did the model work?

- Generalized R-Squared (Cox and Snell; 1989)

\[ R^2 = 1 - \exp\left(\frac{-G^2}{n}\right) \]

\( G^2 \) is the likelihood ratio chi-square statistics for null hypothesis that all covariates are set to zero.

For our modeling, it was 79.79%!

- Model evaluation and robustness

7-Fold cross-validation
THANKS FOR YOUR PATIENCE

Questions
Comments