Adjusted survival curves by using inverse probability of treatment weighting: the comparison of three adapted log-rank tests

Florent Le Borgne \textsuperscript{1,2} & Yohann Foucher \textsuperscript{1}

\textsuperscript{1} EA4275-SPHERE - ITUN, University of Nantes, France
\textsuperscript{2} IDBC/A2com, Pace, France
florent.le-borgne@etu.univ-nantes.fr

27 August 2014
Introduction

Context: Observational study in presence of survival data.

- The causality evaluation between the exposure and the time-to-event requires adjustment.
  - Kaplan-Meier estimator inadequate
  - Multivariate (Cox) model suitable but loss of information:
    Result summarized in a single HR: no graphical representation of a possible evolution over time of the HR

Solution

Adjusted survival curves using the method IPTW (Inverse Probability of Treatment Weighting) based on propensity scores
• The log-rank test = standard test for comparing two survival curves.
• Three versions adapted to the adjusted Kaplan-Meier estimator.
• Other methods based on propensity scores exist (stratification, matching, IPTW).

Objectives of our simulation study

- Evaluate the performances of the adjusted log-rank test compared to the Cox model in terms of type I and II error rates ⇒ Is it necessary to use a multivariate Cox model?
- Choose the most powerful among the three
Notations

- $n = \text{sample size}$
- $T_i = \text{participating time} (i = 1, \ldots, n)$
- $\delta_i = \text{censoring indicator} (\delta_i = 0 \text{ if } T_i \text{ is a right censoring and } \delta_i = 1 \text{ otherwise})$
- $X_i = \text{explanatory variable representing the interest exposure factor composed of } K \text{ groups}$
- $D_k = \text{number of different times for which events are observed in the group } k$, we then have at time $t_j$ ($j = 1, \ldots, D_k$):
  - $d_{jk} = \sum_{i: t_i = t_j} \delta_i I(X_i = k)$: number of subjects in group $k$ undergoing the event in time $t_j$
  - $Y_{jk} = \sum_{i: t_i \geq t_j} I(X_i = k)$: number of subjects in group $k$ at risk at time $t_j$
  - Let $d_j = \sum_{k=1}^{K} d_{jk}$ and $Y_j = \sum_{k=1}^{K} Y_{jk}$
The IPTW weighting

- The IPTW method proposes to correct the contribution of each individual by a weight $w_{ik} = 1/p_{ik}$.

  where $p_{ik} = P(X_i = k | Z_i)$

  and $Z_i$ the vector of potential confounding factors

- The weighted number of events and individuals at risk can then be obtained:

  - $d_{jk}^w = \sum_{i: t_i = t_j} w_{ik} \delta_i I(X_i = k)$
  
  - $Y_{jk}^w = \sum_{i: t_i \geq t_j} w_{ik} I(X_i = k)$

  - $d_j^w = \sum_{k=1}^K d_{jk}^w$ et $Y_j^w = \sum_{k=1}^K Y_{jk}^w$

- Let us consider now only two groups, noted $X = 0$ et $X = 1$. 
Xu and al. proposed an adjusted log-rank test equivalent to the standard one by simply replacing:

- The observed numbers of events by the weighted ones
- The numbers of individuals at risk by the weighted ones

The resulting statistic is:  \( \frac{G^w}{\sqrt{\text{Var}(G^w)}} \) where:

- \( D \) is the number of different times for which events are observed regardless of group
- \( G^w = \sum_{j=1}^{D} d^w_j - Y^w_{j1}(\frac{d^w_j}{Y^w_j}) \)
- \( \text{Var}(G^w) = \sum_{j=1}^{D} \left\{ \frac{Y^w_{j0} Y^w_{j1} d^w_j (Y^w_j - d^w_j)}{(Y^w_j)^2 (Y^w_j - 1)} \right\} \)

A second variant of the adjusted log-rank test is given by Sugihara.

- Differs from the first by the formula of the variance used.

\[ G^w = \sum_{j=1}^{D} d_j^w - Y_{j1}^w \left( \frac{d_j^w}{Y_j^w} \right) \]

\[ \text{Var}(G'^w) = \sum_{j=1}^{D} \left\{ \frac{d_j(Y_j - d_j)}{Y_j(Y_j - 1)} \sum_{i=1}^{Y_j} \left[ \left( \frac{Y_{j0}^w}{Y_j^w} \right)^2 w_i^2 X_i + \left( \frac{Y_{j1}^w}{Y_j^w} \right)^2 w_i^2 (1 - X_i) \right] \right\} \]

Xie and Liu proposed another adaptation of the log-rank test by adjusting the weights of each individual over the time.

- At time $t_j \ (j = 1, \ldots, D_k)$, the weight for an individual $i$ in the group $k$ is reassigned as:
  \[
  w'_{ijk} = w_{ik} \cdot \frac{Y_{jk}}{Y^w_{jk}}
  \]

- The weighted number of events and at risk subjects becomes:
  \[
  d^w_{jk} = \sum_{i: t_i = t_j} w'_{ijk} \delta_i I(X_i = k)
  \]
  and
  \[
  Y^w_{jk} = \sum_{i: t_i \geq t_j} w'_{ijk} I(X_i = k)
  \]

• Same formulas as those proposed by Sugihara but with different weights:

\[ G_{w'} = \sum_{j=1}^{D} d_{j1}^{w'} - Y_{j1}^{w'} \left( \frac{d_{j}^{w'}}{Y_{j}^{w'}} \right) \]

\[ \text{Var}(G_{w'}) = \sum_{j=1}^{D} \left\{ \frac{d_{j}(Y_{j} - d_{j})}{Y_{j}(Y_{j} - 1)} \sum_{i=1}^{Y_{j}} \left[ \left( \frac{Y_{j0}^{w'}}{Y_{j}^{w'}} \right)^{2} w_{ij}'^{2} X_{i} + \left( \frac{Y_{j1}^{w'}}{Y_{j}^{w'}} \right)^{2} w_{ij}'^{2}(1 - X_{i}) \right] \right\} \]

Two other models based on propensity scores

- Weighted univariate Cox model proposed by Cole et Hernán (2004)
  - Exposure : only variable in the model
  - Weighted by the weights $w_{ik}$


- Matching on the logit of the propensity score
  - Matching 1:1 without replacement with the nearest neighbor
  - Caliper equal to 0.2 of the standard deviation
  - Stratified log-rank test

  Rosenbaum PR. and Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* (1985)
Simulation study

- Simulations limited to 5 variables:
  - 1 binary exposure
  - 4 confounders

- Performances of the different models were compared for different:
  - Right-censoring rates (0.30 and 0.68)
  - Sample sizes (100, 250, 500 and 1500)
  - Percentages of exposed subjects (5%, 20% and 40%)
  - Coefficient $\beta_X$ associated with the exposure variable under interest (0, 0.250, 0.365, 0.500)

- When $\beta_X = 0$ we calculated the percentage of rejection of the null hypothesis (type I error rate).

- When $\beta_X \neq 0$ we calculated the percentage of non-rejection of the null hypothesis (type II error rate).
Simulation Results

<table>
<thead>
<tr>
<th>$\beta_X$</th>
<th>$n$</th>
<th>Multivariate Cox</th>
<th>$X_u$</th>
<th>Sugihara</th>
<th>$X_{ie}$</th>
<th>Weighted Cox</th>
<th>Caliper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>100</td>
<td>5.69</td>
<td>19.53</td>
<td>6.94</td>
<td>6.12</td>
<td>10.92</td>
<td>4.70</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>5.31</td>
<td>23.42</td>
<td>6.58</td>
<td>6.27</td>
<td>8.80</td>
<td>4.81</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>4.59</td>
<td>25.39</td>
<td>5.73</td>
<td>5.46</td>
<td>7.10</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>4.85</td>
<td>27.58</td>
<td>5.58</td>
<td>5.43</td>
<td>6.38</td>
<td>4.93</td>
</tr>
<tr>
<td>0.250</td>
<td>100</td>
<td>90.89</td>
<td>74.44</td>
<td>89.04</td>
<td>89.80</td>
<td>84.04</td>
<td>94.75</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>86.05</td>
<td>65.78</td>
<td>86.20</td>
<td>86.37</td>
<td>82.57</td>
<td>92.28</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>77.57</td>
<td>56.41</td>
<td>82.59</td>
<td>82.54</td>
<td>79.11</td>
<td>89.64</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>43.78</td>
<td>31.99</td>
<td>65.14</td>
<td>64.74</td>
<td>60.86</td>
<td>75.99</td>
</tr>
<tr>
<td>0.365</td>
<td>100</td>
<td>86.90</td>
<td>70.03</td>
<td>85.35</td>
<td>85.87</td>
<td>79.50</td>
<td>93.10</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>74.66</td>
<td>55.24</td>
<td>79.20</td>
<td>79.30</td>
<td>73.37</td>
<td>88.20</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>56.71</td>
<td>40.35</td>
<td>69.68</td>
<td>69.26</td>
<td>64.21</td>
<td>81.67</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>13.30</td>
<td>13.79</td>
<td>40.95</td>
<td>40.37</td>
<td>36.97</td>
<td>54.16</td>
</tr>
<tr>
<td>0.500</td>
<td>100</td>
<td>80.13</td>
<td>61.30</td>
<td>79.06</td>
<td>79.79</td>
<td>71.88</td>
<td>90.89</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>58.80</td>
<td>41.57</td>
<td>68.06</td>
<td>67.87</td>
<td>60.14</td>
<td>81.77</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>32.03</td>
<td>24.40</td>
<td>54.94</td>
<td>54.21</td>
<td>47.34</td>
<td>69.80</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>0.99</td>
<td>2.95</td>
<td>16.81</td>
<td>16.15</td>
<td>14.76</td>
<td>27.00</td>
</tr>
</tbody>
</table>

**Table 1:** Error rates obtained from data with 40% of exposed subjects and 68% of censure. (a) Type I errors rate in percentages. (b) Type II errors rate in percentages. 10 000 samples simulated for each scenario.
Conclusion of simulation study

- Best performances obtained by the multivariate Cox model.
- Matched model: loss of power.
- Univariate weighted Cox model: more important type I error rate.
- Among the three versions of the adjusted log-rank tests:
  - The one proposed by Xu and al. does not respect the type I error rate
  - The two others show type I and type II error rates slightly higher than those of the multivariate Cox model
  - Slightly better type I error rates for the one proposed by Xie and Liu
Two limitations appear in our study:

- We have only considered the case of a binary exposure.
  * Adjusted survival curves can be generalized to more than two groups (multinomial logistic regression)
  * Adjusted log-rank test requires further developments
- Only the context in which the PH assumption holds true was simulated.
• In conclusion, we retain two good methods:
  - The multivariate Cox model
  - The adjusted survival curves with the log-rank test proposed by Xie et Liu

• Multivariate Cox model: the most efficient, requires verification of assumptions, summarizes the results in one HR.

• Adjusted survival curves: illustrate more precisely the differences in survivals between groups, lower performances of the associated adjusted log-rank test.


- Rosenbaum PR. and Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*