Estimation from Censored Medical Count Data

Onur Baser, PhD
Economist, Outcomes Research
Medstat
777 E. Excenter Boulevard, Suite 410
Annapolis, Maryland 21403
301-951-1897
http://www.medstat.com • www.medstat.com

Objectives
To address the following research questions:
• How can we solve the selection bias in estimating censored count models with a method applied using standard statistical software programs?
• How can we not see the selection bias attributable to censoring?
• How do we apply the method to the hospitalization of lung-cancer patients?

Introduction
• Censoring is a common problem with medical cost and count data. Medical cost estimation from censored data has been extensively studied.1
• Censored medical count data, however, has not been analyzed.
• Methods from traditional count models such as Poisson, negative binomial regressions, etc. cannot be directly applied.

Previous Approaches/Tidbits
• Approach: Sample mean of observed counts from all cases.
  • Boot: Downward bias since we are not accounting for the costs after censoring.
• Approach: Sample mean of only uncensored patients.
  • Boot: Based toward patients with a shorter survival time since patients who have longer survival times are highly likely to be censored.
• Approach: Standard survival analysis (such as Kaplan-Meier, Cox)
  • Boot: Not applicable since the count variables during and after the study time are not independent.

Study Design and Methods
• The inverse probability weighted (IPW) least squares method is used to assess the effects of covariates (e.g., patient and clinical characteristics) on hospitalization with censored data. IPW simply weights each observation by the inverse probability of appearing in the sample.
• Suppose we are interested in hospitalization over period [a, b].
  • Since there is no further hospitalization after death, the total hospitalization over (a, b] is the same in the cumulative count at \( \tau = b \), where \( \tau \) is survival time. Let \( x = cdf(\tau) \), where \( c \) is the time of censoring.
• Suppose \( c = 0 \) and \( c \) are independent given \( x \), where \( x \) is the set of explanatory variables.
• Assumption:
  1. \( x \) is always observed; \( x_{\text{obs}} \) is only observed when \( c = 0 \)
  2. \( x_{\text{obs}} \) can be ignored in selection equation, conditional on \( x \).
• Calculate \( \hat{p}(x) \) on the coefficients of \( x \).
• In practice, selection probability \( p(\tau, x) \) is unknown.
• Suppose censoring is not covariate dependent, and define \( p(\tau, x) = p(x) \).
• Then we suggest the following estimator to replace unknown selection probability:
  • Use the sample \( \tau = cdf(\tau) \) where \( \tau = cdf(\tau) \) and construct the product-limit estimator \( p_{\text{IPW}}(x) = p(x) \).
• Note that we use both censored and uncensored observations to estimate selection probability. STATA command "pop" for \( p(\tau, x) \) is not sufficient since we may commit Type II error.

Application to Hospitalization
• In the application to estimation of hospitalization, 183 patients with incident cases of lung cancer were recruited from 24 Michigan community hospitals and their affiliated oncology settings.
• We obtained Medicare claim files for two years following cancer diagnosis.
• A detailed dataset description can be found in Baser et al., 2003.

Results

Table 1. Summary Statistics from the Lung-Cancer Study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitalization</td>
<td>Total number of hospitalizations</td>
<td>1.142105</td>
<td>0.266667</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Age</td>
<td>Age</td>
<td>72.875</td>
<td>1.0172</td>
<td>55</td>
<td>85</td>
</tr>
<tr>
<td>Symptoms</td>
<td>A count of all symptoms</td>
<td>10.1200</td>
<td>0.9675</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>Three or more comorbid conditions</td>
<td>0.6593</td>
<td>0.4757</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>Patient's physical functioning</td>
<td>71.4583</td>
<td>0.9379</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Late stage</td>
<td>Late stage regional or distant</td>
<td>8.6471</td>
<td>0.4892</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
<td>72.875</td>
<td>1.0172</td>
<td>55</td>
<td>85</td>
</tr>
<tr>
<td>Symptoms</td>
<td>A count of all symptoms</td>
<td>10.1200</td>
<td>0.9675</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>Three or more comorbid conditions</td>
<td>0.6593</td>
<td>0.4757</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>Patient's physical functioning</td>
<td>71.4583</td>
<td>0.9379</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Late stage</td>
<td>Late stage regional or distant</td>
<td>8.6471</td>
<td>0.4892</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Estimates of Hospitalization Through Negative Binomial and IPW Negative Binomial

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
<td>72.875</td>
<td>1.0172</td>
<td>55</td>
<td>85</td>
</tr>
<tr>
<td>Symptoms</td>
<td>A count of all symptoms</td>
<td>10.1200</td>
<td>0.9675</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>Three or more comorbid conditions</td>
<td>0.6593</td>
<td>0.4757</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>Patient's physical functioning</td>
<td>71.4583</td>
<td>0.9379</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Late stage</td>
<td>Late stage regional or distant</td>
<td>8.6471</td>
<td>0.4892</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion
• Our study proposes a method that can be applied to censored count data to solve possible censoring bias.
• A regression-based Hausman test is easy to apply and shows if censoring bias exists.

References