Methods for Longitudinal Data
Categorical Response

Gilbert Ritschard

Institute for demographic and life course studies, University Geneva
http://mephisto.unige.ch

Doctoral Program, Lausanne, May 20, 2011
## Typology of methods for life course data

<table>
<thead>
<tr>
<th>Questions</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>duration/hazard</strong></td>
<td><strong>state/event sequencing</strong></td>
</tr>
<tr>
<td>descriptive</td>
<td>Survival curves:</td>
</tr>
<tr>
<td></td>
<td>Parametric (Weibull, Gompertz, ...)</td>
</tr>
<tr>
<td></td>
<td>and non parametric (Kaplan-Meier, Nelson-Aalen) estimators.</td>
</tr>
<tr>
<td>causality</td>
<td>Hazard regression models (Cox, ...)</td>
</tr>
<tr>
<td></td>
<td>Survival trees</td>
</tr>
<tr>
<td></td>
<td>Sequence clustering</td>
</tr>
<tr>
<td></td>
<td>Frequencies of given patterns</td>
</tr>
<tr>
<td></td>
<td>Discovering typical episodes</td>
</tr>
<tr>
<td></td>
<td>Markov models</td>
</tr>
<tr>
<td></td>
<td>Mobility trees</td>
</tr>
<tr>
<td></td>
<td>Association rules among episodes</td>
</tr>
</tbody>
</table>
Outline

1. Survival analysis
2. State sequence analysis: brief overview
3. Mobility and transition rates
4. Conclusion
Section outline

1 Survival analysis
   • Survival curves
   • Survival models and trees
Survival Approaches
Event history analysis

- **Survival or Event history analysis** (Mills, 2011)(Blossfeld and Rohwer, 2002)
  - Focuses on one event.
  - Concerned with duration until event occurs or with hazard of experiencing event.

- **Survival curves**: Distribution of duration until event occurs

  \[ S(t) = p(T \geq t) \] .

- **Hazard models**: Regression like models for \( S(t, x) \) or hazard

  \[ h(t) = p(T = t | T \geq t) \]

  \[ h(t, x) = g(t, \beta_0 + \beta_1 x_1 + \beta_2 x_2(t) + \cdots) \] .
Survival Approaches

Event history analysis

- **Survival or Event history analysis** (Mills, 2011) (Blossfeld and Rohwer, 2002)
  - Focuses on one event.
  - Concerned with duration until event occurs or with hazard of experiencing event.
- **Survival curves**: Distribution of duration until event occurs
  
  \[ S(t) = p(T \geq t) . \]

- **Hazard models**: Regression like models for \( S(t, x) \) or hazard
  
  \[ h(t) = p(T = t \mid T \geq t) \]

  \[ h(t, x) = g(t, \beta_0 + \beta_1 x_1 + \beta_2 x_2(t) + \cdots) . \]
Survival curves (Switzerland, SHP 2002 biographical survey)

- Leaving home
- Marriage
- 1st Childbirth
- Parents’ death
- Last child left
- Divorce
- Widowing
Section outline

1. Survival analysis
   - Survival curves
   - Survival models and trees
SHP biographical retrospective survey
http://www.swisspanel.ch

- SHP retrospective survey: 2001 (860) and 2002 (4700 cases).
- We consider only data collected in 2002.
- Data completed with variables from 2002 wave (language).

Characteristics of retained data for divorce
(individuals who get married at least once)

<table>
<thead>
<tr>
<th></th>
<th>men</th>
<th>women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1414</td>
<td>1656</td>
<td>3070</td>
</tr>
<tr>
<td>1st marriage dissolution</td>
<td>231</td>
<td>308</td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>16.3%</td>
<td>18.6%</td>
<td>17.6%</td>
</tr>
</tbody>
</table>
SHP retrospective survey: 2001 (860) and 2002 (4700 cases).

We consider only data collected in 2002.

Data completed with variables from 2002 wave (language).

### Characteristics of retained data for divorce
(individuals who get married at least once)

<table>
<thead>
<tr>
<th></th>
<th>men</th>
<th>women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>1414</td>
<td>1656</td>
<td>3070</td>
</tr>
<tr>
<td><strong>1st marriage dissolution</strong></td>
<td>231</td>
<td>308</td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>16.3%</td>
<td>18.6%</td>
<td>17.6%</td>
</tr>
</tbody>
</table>
Marriage duration until divorce
Survival curves

Duration of marriage, Women
- 1942 and before
- 1943-1952
- 1953 and after

Duration of marriage, Men
- 1942 and before
- 1943-1952
- 1953 and after
Marriage duration until divorce
Hazard model

- Discrete time model (logistic regression on person-year data)
- \( \exp(B) \) gives the Odds Ratio, i.e. change in the odd \( \frac{h}{1-h} \) when covariate increases by 1 unit.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \exp(B) )</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthyr</td>
<td>1.0088</td>
<td>0.002</td>
</tr>
<tr>
<td>university</td>
<td>1.22</td>
<td>0.043</td>
</tr>
<tr>
<td>child</td>
<td>0.73</td>
<td>0.000</td>
</tr>
<tr>
<td>language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknwn</td>
<td>1.47</td>
<td>0.000</td>
</tr>
<tr>
<td>French</td>
<td>1.26</td>
<td>0.007</td>
</tr>
<tr>
<td>German</td>
<td>1</td>
<td>ref</td>
</tr>
<tr>
<td>Italian</td>
<td>0.89</td>
<td>0.537</td>
</tr>
<tr>
<td>Constant</td>
<td>0.000000000004</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Divorce, Switzerland, Relative risk

LIVES Doctoral Program: Categorical longitudinal data
Survival analysis
Survival models and trees
Hazard model with interaction

- Adding interaction effects detected with the tree approach
- Improves significantly the fit ($\Delta \chi^2 = 0.004$)

<table>
<thead>
<tr>
<th></th>
<th>$\exp(B)$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>born after 1940</td>
<td>1.78</td>
<td>0.000</td>
</tr>
<tr>
<td>university</td>
<td>1.22</td>
<td>0.049</td>
</tr>
<tr>
<td>child</td>
<td>0.94</td>
<td>0.619</td>
</tr>
<tr>
<td>language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unkwn</td>
<td>1.50</td>
<td>0.000</td>
</tr>
<tr>
<td>French</td>
<td>1.12</td>
<td>0.282</td>
</tr>
<tr>
<td>German</td>
<td>1</td>
<td>ref</td>
</tr>
<tr>
<td>Italian</td>
<td>0.92</td>
<td>0.677</td>
</tr>
<tr>
<td>b_before_40*French</td>
<td>1.46</td>
<td>0.028</td>
</tr>
<tr>
<td>b_after_40*child</td>
<td>0.68</td>
<td>0.010</td>
</tr>
<tr>
<td>Constant</td>
<td>0.008</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Outline

1. Survival analysis
2. State sequence analysis: brief overview
3. Mobility and transition rates
4. Conclusion
Illustrative \textit{mvad} data set

- McVicar and Anyadike-Danes (2002)’s study of transition from school to employment in North Ireland.
  - Survey of \textbf{712 Irish youngsters}.
  - Sequences describe their follow-up during the \textbf{6 years} after the end of compulsory school (16 years old) and are formed by \textbf{70 successive} monthly observed states between September 1993 and June 1999.

Sates are:  
\begin{itemize}
  \item EM \quad \text{Employment}
  \item FE \quad \text{Further education}
  \item HE \quad \text{Higher education}
  \item JL \quad \text{Joblessness}
  \item SC \quad \text{School}
  \item TR \quad \text{Training}
\end{itemize}
Sate sequences - mvad data set

- First sequences (first 20 months)

  Sequence

- compact representation (SPS format)

  Sequence
  [1]  (EM,4)-(TR,2)-(EM,64)
  [2]  (FE,36)-(HE,34)
  [3]  (TR,24)-(FE,34)-(EM,10)-(JL,2)
  [4]  (TR,47)-(EM,14)-(JL,9)
State sequences: Graphical display

I-plot, Individual sequences

F-plot, Most frequent patterns

R-plot, Representative sequences

D-plot, Successive transversal distributions

Ht-plot, Transversal entropies

Ms-plot, Sequence of modal states
Pairwise dissimilarities and cluster analysis

- Different metrics permit to compute pairwise dissimilarities between sequences
  - of which optimal matching (Abbott and Forrest, 1986) is perhaps the most popular in social sciences
- Once you have pairwise dissimilarities, you can do
  - cluster analysis of sequences
  - principal coordinate analysis
  - measure the discrepancy between sequences
  - Find representative sequences, either most central or with highest density neighborhood (Gabadinho et al., 2011b)
  - ANOVA-like analysis and Regression trees (Studer et al., 2011)
Cluster analysis: Outcome

- Rendering the cluster contents: transversal state distributions
LIVES Doctoral Program: Categorical longitudinal data
State sequence analysis: brief overview

Cluster analysis: Outcome (2)

- Mean time per state by cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mean time (weighted n)</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>226.47</td>
<td>EM</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>189.06</td>
<td>EM</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>196.82</td>
<td>EM</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>99.22</td>
<td>EM</td>
</tr>
</tbody>
</table>

- Employment
- Further education
- Higher education
- Joblessness
- School
- Training
Regression tree
Outline

1. Survival analysis
2. State sequence analysis: brief overview
3. Mobility and transition rates
4. Conclusion
Section outline

Mobility and transition rates
  - Markov process
  - Mobility tree
Markov process: Principle

(Brémaud, 1999; Berchtold and Raftery, 2002)

- Assume we have a sequence of states (not necessarily panel data)
- How is state in position \( t \) related to previous states?
- What is the probability to switch to state \( B \) in \( t \) when we are in state \( A \) in \( t - 1 \)?
  - Probability to fall next year into joblessness when we have a partial time job.
  - Probability to stay unemployed next \( t \) when we are currently unemployed.
  - Probability to recover from illness next month.
Homogenous Markov process: Assumptions

- transition probability is the same whatever $t$ (homogeneity)
- a few lagged states summarize all the sequence before $t$
- 1st order: state in $t - 1$ summarizes all the sequence before $t$; i.e.; state in $t$ depends only on state in $t - 1$
- 2nd order: states in $t - 1$ and $t - 2$ summarize all the sequence before $t$; i.e.; state in $t$ depends only on states in $t - 1$ and $t - 2$
- ...

19/5/2011gr 24/37
Homogenous Markov process: Assumptions

- transition probability is the same whatever \( t \) (homogeneity)
- a few lagged states summarize all the sequence before \( t \)
- 1st order: state in \( t - 1 \) summarizes all the sequence before \( t \); i.e.; state in \( t \) depends only on state in \( t - 1 \)
- 2nd order: states in \( t - 1 \) and \( t - 2 \) summarize all the sequence before \( t \); i.e.; state in \( t \) depends only on states in \( t - 1 \) and \( t - 2 \)
- ...

19/5/2011gr 24/37
Blossfeld and Rohwer (2002) sample of 600 job episodes extracted from the German Life History Study

Job episodes partitioned into 3 job length categories
- short (1) = \(\leq 3\) years
- medium (2) = \((3; 10]\) years
- long (3) = \(> 10\) years

Data reorganized into 162 sequences of 2 to 9 job episodes (units with single episode not considered)

How does present episode length depend upon those of preceding jobs?
LIVES Doctoral Program: Categorical longitudinal data
Mobility and transition rates

Markov process

Markov matrices of order 0, 1 and 2

\[
\begin{array}{ccc}
  t-2 & t-1 & t \\
  \uparrow & \uparrow & \uparrow \\
  \downarrow & \downarrow & \downarrow \\
  t-2 & t-1 & t
\end{array}
\]

<table>
<thead>
<tr>
<th></th>
<th>job length at $t$</th>
<th>half conf. interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Indep</td>
<td>.50</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.20</td>
</tr>
<tr>
<td>$t-1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>.55</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>.60</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$t-2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>.37</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>.50</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

19/5/2011gr 26/37
Main findings

- **First order:**
  - Probability to start short job (1) after a short one (1) is much higher than starting a medium (2) or long job (3)
  - not the case after a medium or long job

- **Second order:**
  - No clear evidence about impact of lag 2 job
  - Main difference concerns long job (3) (but not significant)
  - Confirmed by MTD model, which gives weight 0 to second lag
Two state hidden Markov model

<table>
<thead>
<tr>
<th>Hidden state at $t$</th>
<th>1</th>
<th>2</th>
<th>half conf. interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t-1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.78</td>
<td>.22</td>
<td>.12</td>
</tr>
<tr>
<td>2</td>
<td>.53</td>
<td>.47</td>
<td>.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hidden state</th>
<th>Job length</th>
<th>half conf. interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>.75</td>
<td>.23</td>
</tr>
<tr>
<td>2</td>
<td>.05</td>
<td>.58</td>
</tr>
</tbody>
</table>

Initial probabilities: 0.56, 0.44, 0.11
Hidden Markov Model (HMM)

- Relaxing homogeneity assumption with HMM
- Fitting a HMM with 2 hidden states
  - distribution of initial state of hidden variable
  - transition matrix of hidden process
  - distribution of transitions to the job length categories associated to each hidden state
LIVES Doctoral Program: Categorical longitudinal data

Mobility and transition rates

Mobility tree

Section outline

3. Mobility and transition rates
   - Markov process
   - Mobility tree
LIVES Doctoral Program: Categorical longitudinal data
Mobility and transition rates
Mobility tree

**Mobility tree**
Social transition tree with birth place covariate (Ritschard and Oris, 2005)

Low, Clock, High

![Mobility tree diagram]
Outline

1. Survival analysis
2. State sequence analysis: brief overview
3. Mobility and transition rates
4. Conclusion
Conclusion

- Now, it is your turn!
- To chose a method, you first have to
  - Clarify what you are looking for
    - typical patterns, departures from standards, ...
    - specific transitions or holistic view
    - relationships with context (covariates)
    - ...
  - Identify the nature of your data
    - Categorical vs numerical
    - Direct or indirect measures of variable of interest
    - Long or short sequences
    - ...
Thank You!
References I


References II


