What is discrete longitudinal data analysis?

Ivonne Solis-Trapala

NCRM Lancaster-Warwick Node

ESRC Research Festival 2008
Discrete longitudinal data analysis

The title explained:

- **Discrete variables**: Variables having only integer values, for example:
  Number of heart attacks (0, 1, 2...),
  Failure (0) or success (1) in a psychological test item.

- **Longitudinal data**: A variable for each of a number of subjects
  is measured a number of different time points.
Longitudinal data is not:

- **Time series data**: Single long series of measurements,
- **Multivariate data**: Single outcome of two or more different kinds of measurements on each subject;

but:

a large number of short time series
Example of time series data
Cardiovascular events during the FIFA world cup in 2006

Figure 1. Daily Cardiovascular Events in the Study Population from May 1 to July 31 in 2003, 2005, and 2006.

The FIFA World Cup 2006 in Germany started on June 9, 2006, and ended on July 9, 2006. The 2006 World Cup matches with German participation are indicated by numbers 1 through 7: match 1, Germany versus Costa Rica; match 2, Germany versus Poland; match 3, Germany versus Ecuador; match 4, Germany versus Sweden; match 5, Germany versus Argentina; match 6, Germany versus Italy; and match 7, Germany versus Portugal (for third-place standing). Match 8 was the final match, Italy versus France.

Example of multivariate data
Fair admission process to a university

\[ A = \text{student is admitted (yes/no)} \]
\[ G = \text{student’s gender (female/male)} \]
\[ D = \text{department (Mathematics, Medicine, Engineering, Biology)} \]

Fair: female admission rates similar to male admission rates at each department

Unfair: otherwise
Example of longitudinal data

Improving communication skills of oncologists

“Of course I’m listening to your expression of spiritual suffering. Don’t you see me making eye contact, striking an open posture, leaning towards you and nodding empathetically?”
Example of longitudinal data (cont.)

A randomised controlled trial

160 clinicians took part in baseline filming. They were randomised to one of four groups:

- **Group A**: written feedback and course
- **Group B**: course only
- **Group C**: written feedback only
- **Group D**: control

**Time 1**

- **Group A**: 3 day course, 1 month after feedback
- **Group B**: 3 day course, feedback given during course
- **Group C**
- **Group D**

Filming and assessments 3 months after course

**Time 2**

Filming and assessments 15 months after course

Doctors in groups C and D invited to attend the course

**Time 3**

Example of longitudinal data (cont.)
The MIPS data

- MIPS = Medical Interaction Process
- DATA: COUNTS of primary outcomes, i.e. leading questions, expressions of empathy, focused questions

- Participants: 160 doctors
- 2 consultations filmed for each doctor at TIMES 1, 2, 3
Example of longitudinal data (cont.)

Longitudinal performance

[Graph showing performance over time for control and course groups.]
Statistical aspects

Modelling independent discrete data

Continuous response

- Linear regression model
  - e.g. ANOVA
  - i) Independent observations
  - ii) Mean of the response is a linear function of some explanatory variables
  - iii) The variance of the response is constant

Discrete response

- Generalised linear model
  - e.g. logistic regression, Poisson regression
  - ii) Mean of the response is associated with a linear function of some explanatory variables through a link function
  - iii) The variance of the response is a function of the mean
Why the model assumptions are important

Consider the following four data sets

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.10</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.10</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>12.50</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>

Fitting a linear regression model
Same for four data sets

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.0001</td>
<td>1.1247</td>
<td>2.67</td>
<td>0.0257</td>
</tr>
<tr>
<td>x</td>
<td>0.5001</td>
<td>0.1179</td>
<td>4.24</td>
<td>0.0022</td>
</tr>
</tbody>
</table>

Multiple R-Squared: 0.6665

F-statistic: 17.99 on 1 and 9 DF, p-value: 0.002170

Equation of regression line: \( y = 3 + 0.5x \)
Assess the linearity assumption!

\[ R^2 = 0.67 \]
Some statistical approaches based on GLM for analysing longitudinal discrete data

Target of inference: mean response (—) vs. mean response of an individual (—)

- Marginal models
- Random effects models
- Transition models
Example of marginal models

Generalised estimating equations (GEE)

- describe the relationship between *response variable* and *explanatory variables* with a *population average* regression model
- the approach provides consistent regression coefficient estimates even if the correlation structure is mis-specified
How is GEE implemented?

1. specify the mean regression
2. make a plausible guess of the covariance matrix
3. fit the model
4. use the residuals to adjust standard errors
Can you read this?

Icdn’uolt blveiee taht I cluod aulaclty uesdnatnrd waht I was rdanieg: the phaonmneal pweor of the hmuan mnid. Aoccdrnig to a rsceearch taem at Cmabrigde Uinervtisy, it deosn’t mttaer in waht oredr the ltteers in a wrod are, the olny iprmoatnt tihng is taht the frist and lsat ltteer be in the rghit pclae. The rset can be a taotl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Such a cdonition is arppoiatrely cllaed Typoglycemia.
Finally...  
An example of misleading inferences when standard errors of regression parameters estimates are not adjusted: MIPS data revisited

Robust conditional Poisson regression models comparing $T_2$ (3 month post-course) to $T_1$ (baseline) assessment

<table>
<thead>
<tr>
<th>Behaviour</th>
<th>$\hat{\beta}_c$</th>
<th>naive SE</th>
<th>robust SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leading questions</td>
<td>-0.30</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>Focused questions</td>
<td>0.23</td>
<td>0.077</td>
<td>0.13</td>
</tr>
<tr>
<td>Focused and open questions</td>
<td>0.16</td>
<td>0.067</td>
<td>0.10</td>
</tr>
<tr>
<td>Expressions of empathy</td>
<td>0.41</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>Summarising of information</td>
<td>0.054</td>
<td>0.11</td>
<td>0.24</td>
</tr>
<tr>
<td>Interruptions</td>
<td>-0.15</td>
<td>0.30</td>
<td>0.41</td>
</tr>
<tr>
<td>Checking understanding</td>
<td>-0.18</td>
<td>0.15</td>
<td>0.22</td>
</tr>
</tbody>
</table>