Event History Analysis (EHA) / Hazard Rate Analysis
Vartanian: SW 541

You use EHA when you have longitudinal data and you want to understand the timing of events.

Example:

1. You want to know the amount of time between getting married and having a child. What are some of the influences on the timing of having a child?

2. You want to know how long people spend on welfare. Or another words, how long does it take specific groups of people to get off of welfare. Which groups tend to spend a long time on welfare and which groups spend a short period of time.

3. How long do people keep their jobs? Is there a lot of turnover in jobs. Here the event may be leaving a job -- examining the length of time people spend on jobs. What are the characteristics that cause people to leave jobs.

4. How long do people stay on the wagon once off of a particular drug? Here the event may be drinking again. Is there a difference in treatments and other characteristics of individuals that cause some people to start drinking again.

5. What is the likelihood of death. What are the characteristics of people who die? Or who die from specific types of causes? We can examine the characteristics of the people over time to determine the likelihood of death, given that you have survived for x number of years or you are x years old.

How is this different from OLS regression. Why can't we simply use OLS regression to determine these events?

1. We can use time varying variables in these models, whereas these types of variables are more difficult to use in OLS models.

2. Censoring. Some people do not experience the event by the end of the sampling period.

Because OLS regression has difficulty or produces biased results because of these types of problems, we use what is known as EHA or hazard rate analysis.

1. Time varying variables:

Let's say you're interested in spells on welfare and the factors that affect time spent on welfare, such as the number of children in your household. In an OLS model, if we were examining the number of children in a longitudinal sample that lasts 12 years, what we could do is include the

C:\WP60\LECT2.PHD\Hazards\Hazardintro.doc
number of children in each of the 12 years of the sample. But what about those people who have an event -- they exit welfare in year1? What are we to do with them? Are we to exclude them from the study because we don't want to know how many children they've had after they've exited welfare? Using EHA, we are able to examine people up to the time of their event and not include information pertaining to periods outside of the period we are interested in.

Example:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Kids</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person 1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Person 2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Person 3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Person 4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

The bold indicates the time periods the individual was receiving welfare.

For person 1, the person was receiving welfare for 4 periods. To determine the number of children they have that we will use in a regression model, we could either average the number of children during the time they were on welfare or create a number of different # of children variables according to the year they're on so that we get a more precise measure of the factors related to getting off of welfare. An average measure doesn't tell us much about how the number of children affect the likelihood of leaving welfare. It only tells us the overall effect of having some number of children on how long you stay on welfare. In event history analysis and hazard rate analysis, we'd like to see what is going on in periods where you continue to receive welfare and in periods when you leave the state or have the event. So determining averages for variables doesn't work well.

The second method would be to create a variable for each year that a person is on welfare. Thus, we could have variables for each of the years that the person is on welfare, such as kid1, kid2, kid3, kid4,...,kid12. This would give us a more precise understanding of the number of children in the household for each of the periods the person is on welfare. The problem with this method is that some individuals will have valid data for kid1 to kid12 (person 3), while others will only have valid data for kid5 to kid8 (person1), and others for kid1 to kid5 (person2), or kid8 to kid12 (person 4). When we use these within a regression model, those observations with missing data for any of the variables will be deleted from the regression. Because person 3 has valid data on kid1 to kid12, person1 will have missing data for kid1 to kid4 and kid9 to kid12, person 3 will have missing data for kid6 to kid12, and person 4 will have missing data for kid1 to kid6. In other words, the regression will not use the information from persons 1, 2 and 4 because some of the independent variables have missing data. Thus, this method is not possible.
2. Censoring.

What happens when after 12 years of the sample the person does not exit welfare? (See persons 3 and 4 above)?

In an OLS model, we might simply have to say that the spell length was 12 years (for person 3) or 5 years (for person 4) and leave it at that.

However, the spell length very well may not be 12 or 5 years. Unless this person actually left AFDC in the next year, we would be underestimating the true length of her time on welfare.

One solution to this problem is to throw out all of those observations where the spell is censored so that we only have completed spells. However, this will introduce bias into our estimates. Thus, again, OLS estimates will be biased.

In EHA, we can take into account both time-varying variables and censored observations. Generally, we can only take into account right censoring.

Hazard Rates:

The most common use of hazard rates is for life tables. These predict how long we will live.

The way we figure this out is by examining the number of people who are alive and the likelihood of dying that year.

What we then say is that, given that you have made it to age 50, your chances of death are 3%. Given that you've made it to year 51, your chances of death are 4%. And so on.

You then look at the midpoint of the interval to determine the likely number of years you will live, given that you've made it this far.

The hazard rate is telling us, given that you're still in the pool, that pool of life, what are the chances of death.

<table>
<thead>
<tr>
<th>Age</th>
<th># Dying</th>
<th>Number in Pool of life</th>
<th>Estimated Hazard</th>
<th>Survival Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>10</td>
<td>100</td>
<td>10/100=.10</td>
<td>.90</td>
</tr>
<tr>
<td>56</td>
<td>10</td>
<td>90</td>
<td>10/90 =.11</td>
<td>.801</td>
</tr>
<tr>
<td>57</td>
<td>10</td>
<td>80</td>
<td>10/80 =.125</td>
<td>.700875</td>
</tr>
<tr>
<td>58</td>
<td>9</td>
<td>70</td>
<td>9/70 =.129</td>
<td>.610462</td>
</tr>
<tr>
<td>59</td>
<td>10</td>
<td>61</td>
<td>10/61 =.164</td>
<td>.510</td>
</tr>
<tr>
<td>60</td>
<td>20</td>
<td>51</td>
<td>20/51 =.392</td>
<td>.310</td>
</tr>
<tr>
<td>61</td>
<td>20</td>
<td>31</td>
<td>20/31 =.645</td>
<td>.110</td>
</tr>
</tbody>
</table>
Survival Formula:

Time 1: \( 1 - \text{hazard for time period 1} = s_1 \quad 1-.10 = .90 \)

Time 2: \( s_1 - \text{hazard for time period 2} \times s_1 = s_2 \quad .90 - .90 \times .11 = .801 \)

Time 3: \( s_2 - \text{hazard for time period 3} \times s_2 = s_3 \quad .801 - .801 \times .125 = .700875 \)

Time 4: ... \( .700875 - .700875 \times .129 = .610462 \)

The way to set these things up on the computer is as follows:

What you're trying to do is determine the effects of particular independent variables on some event, say an earnings exit off of AFDC. Some of your variables will change over time and some will not. What we have are what is called time-varying variables and time-constant variables.

What we want to do is examine the hazard rate: the likelihood of exiting AFDC via increased earnings over time. That is, we want to determine the likelihood of exiting AFDC via increased earnings over time, given that you are still on AFDC. Thus, as years go by, fewer and fewer people will be on AFDC.

How to set this up:

We will get all of the years on AFDC for all of the observations. We will then divide up these observations so that each year on AFDC is a different observation. This is called the discrete hazard.

Thus, someone with 1 year on AFDC will have a single observation within the model. Someone with 5 years on AFDC will have 5 different observations.

What we are able to do then is assign a different value for each of the different variables for each of the different years. For example, we may want to know the effects of having different numbers of children on AFDC spell length. With these time-varying models, we can introduce a new number of children for every year on AFDC. If the number of children does not change, then this variable will be the same for all the different years on AFDC.

After we run this hazard rate model, we will be left with is some b coefficient for kids. This is called a proportional hazard rate model because the model does not tell you the effects of having 3 children in year 3 as opposed to year 2, but the effects of having 3 children. That is, we could run a model that was a non-proportional hazard rate model that told you the effects of having a particular number of children in different years of your AFDC spell. The non-proportional hazard rate model will not do this.

What we then do is put in a set of dummy variables for each of the different years on AFDC.

For example, if it's the first year of the spell, you would have
If year=1 then time1=1.
Else time1=0.

If it's the second year:
time1=0
time2=1.
time3=0
time4=0
time5=0.

By using these dummy time variables, you see what the effects of spending longer and longer periods on welfare are. We can then determine significant levels for the coefficient estimates for the variables. These time variable coefficients and significance levels will indicate is the possibility of duration dependence. That is, does spending more time in a particular state (welfare, unemployment, drug use) decrease the likelihood of leaving the state? There has been little evidence of duration dependence in the welfare literature but several theories have been tested that relate to duration dependence. Some believe that there may be duration dependence for welfare recipients for the following two reasons:

1. Welfare recipients are less trained for the workforce the longer the period of time they're on AFDC.
2. Welfare recipients have become less dependent on men -- they have another means of support with welfare and are therefore less likely to leave welfare by means of marriage.

But, it could also be that those who are most prone to get off get off right away, leaving behind those who are very unlikely to get off. That is, the variables in the model can capture some or much of the difference in the population of welfare recipients, but it may not capture all of the differences between different populations. Thus, what we may be picking up is unobserved heterogeneity within this group.

We may not be picking up some willingness or desire to work or a willingness or desire to get married. There may be some who have a great desire to get off of welfare by work or marriage and others who do not have much of desire to do this. In other words, it's not the length of time that they spend on AFDC, but a general low desirability to get off of welfare. Thus, what we are actually seeing is not duration dependence, but heterogeneity within the AFDC population.

To come up with non-proportional hazard models, you can interact your independent variables with the set of time variables. This will indicate whether the independent variables have a non-proportional effect on the likelihood of leaving the state.

**Continuous Hazard Models:**

In the continuous hazard rate model, you would not divide these observations up, but would have a single variable. You could have time-varying independent variables but this requires a heavy duty statistical package. SAS and SPSS do not have the capabilities to handle time-varying independent variables in a
continuous model.

**Right Censoring:**

Regression and non-regression hazard models can take into account any censoring that takes place within your models. Why is censoring important?

1. The sample period ends, but individuals are still in the sample.
2. People become non-respondents -- drop out of the sample.
3. People end their spell.

Taking into account censoring means that we don't know when in the period the person actually left the sample. Censoring allows us to assume that the person left at some random point in the year that the person left. Thus, what we assume is that in a yearly sample, the person left at the midpoint of the sample year. Thus, for people who exit welfare, we assume that they exited at some midpoint of the year, instead of assuming that they spent the entire year on welfare. Even if someone leaves the sample before they exit the state they're in, we can still use the information from the periods they were in the state (of receiving welfare or unemployed, etc.). What we know is that these individuals had particular conditions (had x number of kids, lived in an area where the unemployment rate was x%,...) and they did not exit the state.

**Left-Censoring:**

Individuals are considered to be left censored when we cannot determine when they started their spell in the state. This is generally the case when the sample begins and they are in the state (of receiving welfare) in the first year or month. Thus, we cannot tell when they began their AFDC spell. Generally, we simply exclude those people who begin their spell before the sample begins, or in the first year of the sample. Some surveys, such as the Survey of Income and Program Participation (SIPP) from the Bureau of the Census ask respondents who receive public assistance when they started receiving this aid. With this information, we can use these data and determine the effects of time on exit probabilities.