Statistical Methods in Epidemiology
Lab 4-5.
Survival Analysis

Introduction
In order to analyse survival data it is necessary to specify (at a minimum) a variable representing survival time and a variable specifying whether or not the event of interest was observed (called the failure variable). Instead of specifying a variable representing survival time we can specify the entry and exit dates.

In Stata, these variables are specified once using the **stset** command and then used for all subsequent survival analysis (st) commands (until the next **stset** command). For example

```
. use melanoma
. stset surv_mm, failure(status==1)
```

The above code shows how we would **stset** the skin melanoma data in order to analyse cause specific survival with survival time in completed months (**surv_mm**) as the time variable. The variable **status** takes the values 0=alive, 1=dead due to cancer, and 2=dead due to other causes. We have specified that only **status=1** indicates an event (death due to melanoma) so Stata will consider observations with other values of status as being censored. If we wanted to analyse observed survival (where all deaths are considered to be events) we could use the following command

```
. stset surv_mm, failure(status==1,2)
```

Some of the Stata survival analysis (st) commands relevant to this course are given below. Further details can be found in the manuals or online help.

```
stset Declare data to be survival-time data
stsplit Split time-span records
stdes Describe survival-time data
stsum Summarize survival-time data
sts Generate, graph, list, and test the survivor and cumulative hazard functions
stir Report incidence-rate comparison
strate Tabulate failure rate
```
stptime Calculate person-time at risk and failure rates
stcox Estimate Cox proportional hazards model
stptest Test of Cox proportional hazards assumption
stphplot Graphical assessment of the Cox prop. hazards assumption
stcoxkm Graphical assessment of the Cox prop. hazards assumption
streg Estimate parametric survival models

Once the data have been stset we can use any of these commands without having to specify the survival time or failure time variables.
## Exercises

1. Using hand calculation (i.e. using a spreadsheet program or pen, paper, and a calculator) estimate the cause-specific survivor function for the sample of 35 patients diagnosed with colon carcinoma (see the table below) using both the Kaplan-Meier method (up to at least 30 months) and the life table method (at least the first 5 annual intervals).

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age at dx</th>
<th>Clinical stage</th>
<th>dx date mmyy</th>
<th>Survival time mm yy</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>male</td>
<td>72</td>
<td>Localized</td>
<td>2.89</td>
<td>2</td>
<td>Dead - other</td>
</tr>
<tr>
<td>2</td>
<td>female</td>
<td>82</td>
<td>Distant</td>
<td>12.91</td>
<td>2</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>3</td>
<td>male</td>
<td>73</td>
<td>Distant</td>
<td>11.93</td>
<td>3</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>4</td>
<td>male</td>
<td>63</td>
<td>Distant</td>
<td>6.88</td>
<td>5</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>5</td>
<td>male</td>
<td>67</td>
<td>Localized</td>
<td>5.89</td>
<td>7</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>6</td>
<td>male</td>
<td>74</td>
<td>Regional</td>
<td>7.92</td>
<td>8</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>7</td>
<td>female</td>
<td>56</td>
<td>Distant</td>
<td>1.86</td>
<td>9</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>8</td>
<td>female</td>
<td>52</td>
<td>Distant</td>
<td>5.86</td>
<td>11</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>9</td>
<td>male</td>
<td>64</td>
<td>Localized</td>
<td>11.94</td>
<td>13</td>
<td>Alive</td>
</tr>
<tr>
<td>10</td>
<td>female</td>
<td>70</td>
<td>Localized</td>
<td>10.94</td>
<td>14</td>
<td>Alive</td>
</tr>
<tr>
<td>11</td>
<td>female</td>
<td>83</td>
<td>Localized</td>
<td>7.90</td>
<td>19</td>
<td>Dead – other</td>
</tr>
<tr>
<td>12</td>
<td>male</td>
<td>64</td>
<td>Distant</td>
<td>8.89</td>
<td>22</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>13</td>
<td>female</td>
<td>79</td>
<td>Localized</td>
<td>11.93</td>
<td>25</td>
<td>Alive</td>
</tr>
<tr>
<td>14</td>
<td>female</td>
<td>70</td>
<td>Distant</td>
<td>6.88</td>
<td>27</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>15</td>
<td>male</td>
<td>70</td>
<td>Regional</td>
<td>9.93</td>
<td>27</td>
<td>Alive</td>
</tr>
<tr>
<td>16</td>
<td>female</td>
<td>68</td>
<td>Distant</td>
<td>9.91</td>
<td>28</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>17</td>
<td>male</td>
<td>58</td>
<td>Localized</td>
<td>11.90</td>
<td>32</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>18</td>
<td>male</td>
<td>54</td>
<td>Distant</td>
<td>4.90</td>
<td>32</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>19</td>
<td>female</td>
<td>86</td>
<td>Localized</td>
<td>4.93</td>
<td>32</td>
<td>Alive</td>
</tr>
<tr>
<td>20</td>
<td>male</td>
<td>31</td>
<td>Localized</td>
<td>1.90</td>
<td>33</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>21</td>
<td>female</td>
<td>75</td>
<td>Localized</td>
<td>1.93</td>
<td>35</td>
<td>Alive</td>
</tr>
<tr>
<td>22</td>
<td>female</td>
<td>85</td>
<td>Localized</td>
<td>11.92</td>
<td>37</td>
<td>Alive</td>
</tr>
<tr>
<td>23</td>
<td>female</td>
<td>68</td>
<td>Distant</td>
<td>7.86</td>
<td>43</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>24</td>
<td>male</td>
<td>54</td>
<td>Regional</td>
<td>6.85</td>
<td>46</td>
<td>Dead – cancer</td>
</tr>
<tr>
<td>25</td>
<td>male</td>
<td>80</td>
<td>Localized</td>
<td>6.91</td>
<td>54</td>
<td>Alive</td>
</tr>
<tr>
<td>26</td>
<td>female</td>
<td>52</td>
<td>Localized</td>
<td>7.89</td>
<td>77</td>
<td>Alive</td>
</tr>
<tr>
<td>27</td>
<td>male</td>
<td>52</td>
<td>Localized</td>
<td>6.89</td>
<td>78</td>
<td>Alive</td>
</tr>
<tr>
<td>28</td>
<td>male</td>
<td>65</td>
<td>Localized</td>
<td>1.89</td>
<td>83</td>
<td>Alive</td>
</tr>
<tr>
<td>29</td>
<td>male</td>
<td>60</td>
<td>Localized</td>
<td>11.88</td>
<td>85</td>
<td>Alive</td>
</tr>
<tr>
<td>30</td>
<td>female</td>
<td>71</td>
<td>Localized</td>
<td>11.87</td>
<td>97</td>
<td>Alive</td>
</tr>
<tr>
<td>31</td>
<td>male</td>
<td>58</td>
<td>Localized</td>
<td>8.87</td>
<td>100</td>
<td>Alive</td>
</tr>
<tr>
<td>32</td>
<td>female</td>
<td>80</td>
<td>Localized</td>
<td>5.87</td>
<td>102</td>
<td>Dead - cancer</td>
</tr>
<tr>
<td>33</td>
<td>male</td>
<td>66</td>
<td>Localized</td>
<td>1.86</td>
<td>103</td>
<td>Dead - other</td>
</tr>
<tr>
<td>34</td>
<td>male</td>
<td>67</td>
<td>Localized</td>
<td>3.87</td>
<td>105</td>
<td>Alive</td>
</tr>
<tr>
<td>35</td>
<td>female</td>
<td>56</td>
<td>Distant</td>
<td>12.86</td>
<td>108</td>
<td>Alive</td>
</tr>
</tbody>
</table>
2. Use Stata to confirm the results you obtained in question 1. After starting Stata, you will first have to specify the data set you wish to analyse, that is

```
use colon_sample
```

In order to use the Stata `ltable` command (life table estimates of the survivor function) we must construct a new variable indicating whether the observation period ended with an event (the new variable is assigned code 1) or censoring (the new variable is assigned code 0). We will call this new variable `csr_fail` (cause-specific failure). The `ltable` command is not a standard Stata survival analysis (st) command and does not require that the data be `stset`.

```
. generate csr_fail=0
. replace csr_fail=1 if status==1
```

The following command will give the actuarial estimates

```
. ltable surv_yy csr_fail
```

Alternatively, we could use

```
. ltable surv_mm csr_fail, interval(12)
```

Using the option `graph` we can also plot the survivor function

```
. ltable surv_mm csr_fail, interval(12) graph
```

Before most Stata survival analysis commands can be used (`ltable` is an exception) we must first `stset` the data:

```
. stset surv_mm, failure(status==1)
```

A listing of the Kaplan-Meier estimates is then obtained as follows

```
. sts list
```

To graph the Kaplan-Meier estimates

```
. sts graph
```
Note that we only have to \texttt{stset} the data once. You can also tell Stata to show the number at risk or the number of censored observations on the Kaplan-Meier plot

\begin{verbatim}
. sts graph, atrisk
. sts graph, lost
\end{verbatim}

Titles and axis labels can also be specified.

\begin{verbatim}
. sts graph, atrisk title(Kaplan-Meier estimates of cause-specific survival) xtitle(Time since diagnosis in months)
\end{verbatim}

3. For the patients diagnosed with localised skin melanoma, use Stata to estimate the cause specific survivor function, using the Kaplan-Meier method with survival time in months, separately for each of the two calendar periods 1975–1984 and 1985–1994. The following commands can be used

\begin{verbatim}
. use melanoma
. keep if stage == 1
. stset surv_mm, failure(status==1)
. sts graph, by(year8594)
\end{verbatim}

(a) Without making reference to any formal statistical tests, does it appear that patient survival is superior during the most recent period?
(b) The following commands can be used to plot the hazard function (instantaneous mortality rate):

\begin{verbatim}
. sts graph, hazard by(year8594)
\end{verbatim}

At what point in the follow-up is mortality highest? Does this pattern seem reasonable? Is this pattern apparent when looking at the plot of the survivor function?

4. In question 3 we studied plots of the survivor function for patients diagnosed with localised skin melanoma by calendar period of diagnosis. Use the log rank test to determine whether there is a statistically significant difference in patient survival between the two periods. The following command can be used:

\begin{verbatim}
. sts test year8594
\end{verbatim}
What do you conclude?

An alternative test is the generalised Wilcoxon, which can be obtained as follows

```
.sks test year8594, wilcoxon
```

5. Let’s now read the melanoma data again, but study all stages.

```
use melanoma, clear
stset surv_mm, failure(status==1)
```

(a) Plot estimates of the survivor function and hazard function by stage. Does it appear that stage is associated with survival?

(b) Estimate the mortality rates for each stage using, for example, the strate command. What are the units of the estimated rates?

(c) If you haven’t already done so, estimate the mortality rates for each stage per 1000 person-years of follow-up.

(d) Study whether survival is different for males and females (by plotting the survivor function, tabulating mortality rates and conducting a log rank test).

(e) Test whether the survival probabilities differ significantly by gender whilst allowing for the effect of stage. What happens after adjusting for the variable stage?

6. **Localised melanoma: modelling cause-specific mortality using Cox regression.** We will now model cause-specific mortality using Cox regression. To fit a Cox proportional hazards model (for cause-specific survival) with calendar period as the only explanatory variable, the following commands can be used

```
use melanoma
keep if stage == 1
stset surv_mm, failure(status==1)
stcox year8594
```

(a) Interpret the estimated hazard ratio, including a comment on statistical significance.

(b) Stata reports a Wald test of the null hypothesis that survival is independent of calendar period. The test statistic (and associated P-value) is reported in the table of parameter estimates (labelled z). Under the null hypothesis, the test statistic has a standard normal (Z) distribution, so the square of the test statistic will have a chi square
distribution with one degree of freedom. Stata also reports a likelihood ratio test statistic of the null hypothesis that none of the parameters in the model are associated with survival (labelled LR chi2(1)). In general, this test statistic will have a chi-square distribution with degrees of freedom equal to the number of parameters in the model. For the current model, with only one parameter, the test statistic has a chi square distribution with one degree of freedom. Compare these two test statistics with each other and with the log rank test statistic (which also has a $\chi^2$ distribution) calculated in question 4. Would you expect these test statistics to be similar? Consider the null and alternative hypotheses of each test and the assumptions involved with each test.

(c) Now include sex and age (in categories) in the model.

```
.xi: stcox sex year8594 i.agegrp
```

i. Interpret the estimated hazard ratio for the parameter labelled Iagegr_2, including a comment on statistical significance.

ii. Is the effect of calendar period strongly confounded by age and sex? That is, does the inclusion of sex and age in the model change the estimate for the effect of calendar period?

iii. Perform a Wald test of the overall effect of age and interpret the results.

```
.test _Iagegrp_1 _Iagegrp_2 _Iagegrp_3
```

(d) Perform a likelihood ratio test of the overall effect of age and interpret the results.

The following commands can be used

```
.xi: stcox sex year8594 i.agegrp
.est store A
.stcox sex year8594
.lrtest A
```

Compare your findings to those obtained using the Wald test. Are the findings similar? Would you expect them to be similar?
7. **Examining the proportional hazards hypothesis (localised melanoma).**

(a) For the melanoma data, plot the log cumulative hazard function for each calendar period. The following command can be used

```
   . use melanoma
   . keep if stage == 1
   . stset surv_mm, failure(status==1)
   . stphplot, by(year8594)
```

Do you think that a proportional hazards assumption is appropriate for these data?

(b) Does the appropriateness of the proportional hazards assumption have any implications for the log rank test?

(c) From the plot, estimate the hazard ratio for patients diagnosed 1985–94 to those diagnosed 1975–84.

(d) Compare the estimated hazard ratio with the one from the fitted Cox model with period as the only explanatory variable. Should the estimates be similar? Are they similar?

(e) Fit the model containing sex, period, and age and test the assumption of proportional hazards.

```
   . xi: stcox sex year8594 i.agegrp, schoen(sch*) scaledsch(schs*)
   . stphtest, detail
```

Is an assumption of proportional hazards appropriate?

(f) Use graphical methods to explore the assumption of proportional hazards by age. For example,

```
   . stphplot, by(agegrp)
   . sts graph, hazard by(agegrp)
   . stphtest, plot(_Iagegrp_3)
```

What do you conclude?

(g) Use time-varying covariates to estimate separate age effects for the first two years of follow-up (and separate estimates for the remainder of the follow-up) while controlling for sex and period. Do the estimates for the effect of age differ between the two periods of follow-up?
8. **Cox regression with observed (all-cause) mortality as the outcome.** Now fit a model to the localised melanoma data where the outcome is observed survival (i.e. all deaths are considered to be events).

```
. stset surv_mm, failure(status==1,2)
. keep if stage==1
. xi: stcox sex year8594 i.agegrp
```

(a) Interpret the estimated hazard ratio for the parameter labelled `i.agegrp_2`, including a comment on statistical significance.

(b) On comparing the estimates between the observed and cause-specific survival models it appears that only the parameters for age have changed substantially. Can you explain why the estimates for the effect of age would be expected to change more than the estimates of the effect of sex and period?

9. **Cox model for cause-specific mortality for melanoma (all stages).** Use Cox regression to model the cause-specific survival of patients with skin melanoma (including all stages).

(a) First fit the model with sex as the only explanatory variable. Does there appear to be a difference in survival between males and females?

(b) Is the effect of sex confounded by other factors (e.g. age, stage, subsite, period)? After controlling for potential confounders, does there appear to a difference in survival between males and females?