```
STOP



*****************************************************.
***  Longitudinal Data Analysis for Social Science Researchers
**
**
** ESRC Researcher Development Initiative training programme:
**
**
**   Training materials lab 3:
**     EVENT HISTORY DATA ANALYSIS AND DATA MANAGEMENT  .
**
**
**      www.longitudinal.stir.ac.uk
**      Paul Lambert / Vernon Gayle, 26 August 2007
*****************************************************.


**** STATA VERSION ************************************



*********************************************************.
*********************************************************.
** The file below covers examples of event history data managment and data analysis
**       primarily using data from the British Household Panel Survey 1991-2004:
**
**
***    Exercise 1:  Event history data construction example
***    Exercise 2:  Event history data: descriptive techniques
***    Exercise 3:  Looking at event history models
***    Exercise 4:  Event history analysis in discrete time
***    Exercise 5:  Event history analysis and competing risks
***    Exercise 6:  Evaluating alternative parametric event history models
***    Exercise 7:  Distinguishing Age, period and cohort effects in duration models
**
**
*********************************************************.



*********************************************************.
** GENERAL INSTRUCTIONS ON THESE FILES
**
** Work through this file in the interactive do-file editor, replicating
**  the STATA do-file commands. Further help on working with STATA is
**  available from the LDA web site.
**
**
*** This lab file assumes you have a number of files downloaded to your
**    machine. You will need the following:
*
*
*
*
** 1) Downloadable from the UK Data Archive:
*
* -All BHPS Waves 1-15 component files in Stata format (UK Data Archive Study number
*   5151 (June 2007 release) (extracted from the zip file 5151STATA8.ZIP)
*      +warning - these are a large volume of files, ~152 different files, ~ 600MB)
*
*
* - The six Stata format 'episode' files from the BHPS Derived life history files
*     (UKDA study number 3954, 5th Edition) (covering waves 1-14 only) (you want to access
*      the 3 files data files on the top directory of the '3954.zip' archive,
*      called newpan.dta, xlempe.dta, and xljobe.dta, plus the 3 files in the
*      'episode' subfolder of the '3954.zip' archive, called l*.dta)
*
*
*
*
```



```
*********************************************************.

** .


*********************************************************.
** NOTIFICATION OF FILE LOCATIONS / DIRECTORIES AND STATA SETUP
**
**
**
** i) File location declarations:
*** For the commands below to work, you should begin by running the following
**   macros, which tell Stata where to look for the relevant data files (mentioned
**   above) on your machine : .


global path1 "d:\lda\work\"
* (the location of your working directory -  where you will save
*    newly created data files and output) .


global path2 "d:\data\lda\"
* (the location of a folder where you have saved the
*   WEBCT sourced data files mentioned above ) .


global path3 "d:\data\bhps\w1to15\"
* (the location of a folder where you have saved the BHPS
*   panel data files mentioned above) .


global path3d "d:\data\bhps\lifehist\"
* (the location of a folder where you have saved the BHPS derived
*   life history data files [study number 3954] mentioned above) .



global path9 "d:\temp\"
* (a temporary folder where you can save intermediate files) .


**
************************************************


************************************************
**
** Stata session management:
*** The commands below are used to set some general preferences within Stata,
**    it is usually good advice to run these but it is not essential
clear
set more off
set memory 128M
capture log close
capture log using $path1\log_lab3.txt, replace text
**
**
*********************************************************.



*********************************************************************
*
*   Reminder: other support materials in working with Stata in the context
*    of longitudinal survey datasets are available from the LDA website,
*    http://www.longitudinal.stir.ac.uk/Stata_support.html
*
************************************************************
************************************************************


** ..finally, here is the lab exercise...


************************************************************
```

```
*************************************************************
*************************************************************




*************************************************************.
***  EXERCISE 1) EVENT HISTORY DATA CONSTRUCTION EXAMPLE  .
*************************************************************.


**********************************************************************.

** DATA CONSTRUCTION :.

** The original BHPS combined life history file :.
use $path3d\ljempe.dta , clear

sort pid date
list pid date enddate stemp nextemp duration in 20/50

** this is a 'multistate multiepisode' file, containing sequence of main job episodes
**  of lifetime  for all available BHPS respondents.

* (This derived file is based on BHPS data up until the 2000 wave of data collection).

** Simplify this: select out only the first FT job (or self-employed)  by taking the earliest spell
with
* the appropriate employment status :.

gen job=(stemp==1 | stemp==2)
tab spellno
tab job
gsort +pid -job +spellno
gen fstjob=1
replace fstjob=0 if (pid==pid[_n-1] | job==0)

gen newp=1
replace newp=0 if (pid==pid[_n-1])

tab fstjob
* 15402 first jobs recorded from BHPS respondents.
tab newp
* This is out of 21815 different people in the sample.


* Make a variable 'status' which indicates if not censored.
gen status=(last==0)
label variable status "First job spell duration is not right-censored"
tab status
tab status fstjob
* (11062 of the first jobs are not right censored, ie they ended before the end of the
*    'observation window'; 4340 are right censored).

* Select out the first jobs, along with some info on the job, and save out :.
keep if (fstjob==1)
keep pid date enddate duration status stemp nextemp stsoc stgold sthgs
sav $path9\mtch1.dta, replace


*** Linkage of BHPS individual level information with the first job spells .
**  (file used below is from the general BHPS datasets).

use pid lksex lkdobm lkdoby using $path3\xwlsten.dta, clear
rename lksex sex
rename lkdobm dobm
rename lkdoby doby
summarize
sort pid
sav $path9\mtch2.dta, replace
use $path9\mtch1.dta, clear
sort pid
merge pid using $path9\mtch2.dta
```



```
tab _merge
keep if (_merge==3)
tab sex
tab dobm
tab doby

*** Look at this data: each entry is a spell for an individual (their 1st job spell).
list pid sex doby stemp status date duration enddate in 600/630

*** Calculate 'real' dates

gen syear=floor(date/12) + 1900
gen smonth=date - (floor(date/12))*12
list pid sex doby syear smonth stemp duration in 600/630


sav $path1\ljemp_ext1.dta, replace

* (This derived file will be used later on)



*************************************************************.



*************************************************************.
*************************************************************.




*************************************************************.
*** EXERCISE 2: EVENT HISTORY DATA - DESCRIPTIVE TECHNIQUES .
*************************************************************.



use $path1\ljemp_ext1.dta, clear
numlabel _all, add
summarize
* (This data file was produced in the WebCT demo exercise, 'session 4')


** Each case is an employment event (first job of BHPS respondent, one record per person)
summarize dur
histogram dur



*************************
** Naive description: differences in mean durations by explanatory categories

table stgold sex if stgold >= 1, c(mean duration sd duration n duration)
sort stgold sex
ci duration, by(stgold sex)


graph box duration, over(sex)
graph box duration if stgold >=1, over(stgold)

* Graphs: after a bit more specification on the display:
graph box duration if stgold >=1, over(stgold) asyvars  ///
   title("Duration of first job by Goldthorpe class", size(large) ) ///
   note("Duration in months of first job after leaving education") ///
   ylabel(,angle(45)) legend( order(1 2 3 4 5 6 7 8 9 10 11)   ///
   label(1 "Higher service") label(2 "Lower service") ///
   label(3 "Routine non-manual") label(4 "Personal services") ///
   label(5 "Small proprietors w/e") label(6 "Small proprietors w/o")  ///
   label(7 "Farmers")  label(8 "Foremen") ///
   label(9 "Skill manual")  label(10 "Unskilled manual") label(11 "Agricultural labour") )
graph box duration if stgold >=1, over(stgold, label(angle(60)  )   ///
   relabel(1 "Higher service" 2 "Lower service" 3 "Routine non-manual" 4 "Personal services" ///
   5 "Small proprietors w/e" 6 "Small proprietors w/o" 7 "Farmers" 8 "Foremen" ///
   9 "Skilled manual" 10 "Unskilled manual" 11 "Agricultural labour")  )  ///
```

```
   title("Duration of first job by Goldthorpe class", size(large) ) ///
   note("Duration in months of first job after leaving education") ///
   ylabel(,angle(45))  marker(1, msymbol(smcircle_hollow))

** Naive modelling : models predicting duration by gender, hgs,
gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
summarize duration hgs fem syear doby2
correlate duration hgs fem syear doby2
regress duration hgs fem syear doby2


*****************************************
** Problem: analyses of mean duration don't account for censoring (indicated by status)

tab status

** Stata solution: declare data as survival time

stset duration, failure(status)
stdes

* There are 15401 first jobs, but only 11061 of those events are observed to end;
*   the remainder are right censored .


* Stata's st commands adjust for the censoring effect, eg
ci duration if sex==1 & stgold==1
stci if sex==1 & stgold==1, rmean
* => Taking account of censoring can makes a considerable difference.



***********************************************
** Some examples of stata descriptive analyses, taking account of censoring:


** Life tables :

ltable dur status, by(sex) interval(50)
* A life table 'death' is the non-censored end of an event - ie, end of spell

ltable dur status if sex==1 & stgold >= 1, by(stgold)  interval(120)
ltable dur status if sex==2 & stgold >= 1, by(stgold)  interval(120)

** Censoring-adjusted means / confidence intervals

stci if stgold >= 1, by(sex stgold) rmean


** Censoring adjusted survival graphs :

sts graph , by(sex) censored(single)
sts graph if stgold >= 1, by(stgold) censored(single)

sts graph if stgold >= 1, by(stgold)  ///
   title("Kaplan-Meir Survival Times ", size(large) ) ///
      subtitle("Time in first job, by Goldthorpe class") ///
    note("Duration in months of first job after leaving education") ///
              legend( order(1 2 3 4 5 6 7 8 9 10 11)   ///
           label(1 "Higher service")    label(2 "Lower service")   ///
           label(3 "Routine non-manual")   label(4 "Personal services")   ///
             label(5 "Small proprietors w/e") label(6 "Small proprietors w/o")  ///
           label(7 "Farmers")  label(8 "Foremen") ///
    label(9 "Skilled manual")  label(10 "Unskilled manual") label(11 "Agricultural labour") ) ///
    clpattern(solid dash dot dash_dot longdash_dot shortdash solid dash dot dash_dot longdash )  ///
    clwidth(thin thin thin thin thin thin medthick medthick medthick medthick medthick  )



************************************************************.
************************************************************.
```



```
************************************************************.
*** EXERCISE 3: LOOKING AT EVENT HISTORY MODELS .
************************************************************.



***************************************
********** Regression models on event history data

use $path1\ljemp_ext1.dta, clear
numlabel _all, add
gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
gen agestart=syear - doby2
scatter agestart syear
keep if agestart >= 16 & agestart <= 30
histogram agestart
summarize
stset duration, failure(status)
stdes



** Cox's regression is the most widely used Event History model:

summarize hgs fem  agestart  syear
* HGS : Hope-Goldthorpe score of job (measure of job advantage)
* Agestart: Age in years at start of job
* Syear: Year of start of job

regress hgs fem agestart syear
stcox hgs fem agestart syear
* Advantaged job, and older age at start, associated with longer first job duration
* Being female, and more recent starting year, associated with shorter first job duration


* The Cox model says nothing about the underlying shape of the Hazard function
* conditional on covariates; however it can be estimated by Stata
* ( but time covariates need to be placed within standardised times)
summarize agestart syear
gen agestz=agestart - 25
gen syearz=syear - 1960
stcox hgs fem agestz syearz , basesurv(cox1_s)
stcurve, survival
**(This is very close to the Kaplan Meir curve:
sts graph


** Interaction terms on Cox's regression?
gen femags=fem*agestart
gen femhgs=fem*hgs
gen agehgs=hgs*agestart
stcox hgs fem agestart syear femags femhgs agehgs
* a more advantaged job reduces the longer duration premium of being older at start
* gender doesn't interact with effects of job advantage or age at start
* - BUT: beware collinearity



** Another Cox regression example - with Goldthorpe classes
tab stgold
keep if stgold >= 1
xi: regress dur i.stgold fem agestart syear if stgold >=1
xi: stcox i.stgold fem agestart syear if stgold >=1



***************************************
***********  Semi- and Parametric regressions: Using alternative distributional assumptions
```

```
use $path1\ljemp_ext1.dta, clear
numlabel _all, add
gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
gen agestart=syear - doby2
keep if agestart >= 16 & agestart <= 30
stset duration, failure(status)
stdes
gen agestz=agestart - 25
gen syearz=syear - 1960
summarize dur hgs fem  agestart  syear  agestz syearz


*** Non-parametric survival curves
sts graph
sts graph, by(fem)
* Models make assumptions about the shape of the hazard rate -
* the non-parametric hazard is uneven:
sts graph, hazard
sts graph, hazard by(fem)



***** Cox's semi-parametric model: No assumption on hazard rate other than that proportional,
*  (but can generate an estimated hazard)
stcox hgs fem agestart syear
stcox hgs fem agestz syearz,  basesurv(cox1_s) basehc(cox1_h)
stcurve, survival
stcurve, hazard


***** Paramteric models: assume the hazard rate fits a parametric curve :

streg hgs fem agestart syear , distribution(exponential)
stcurve, survival
stcurve, hazard

streg hgs fem agestart syear , distribution(weibull)
stcurve, survival
stcurve, hazard

streg hgs fem agestart syear , distribution(gompertz)
stcurve, survival
stcurve, hazard

streg hgs fem agestart syear , distribution(lognormal)
stcurve, survival
stcurve, hazard


streg hgs fem agestart syear , distribution(loglogistic)
stcurve, survival
stcurve, hazard

streg hgs fem agestart syear , distribution(gamma)
stcurve, survival
stcurve, hazard

* (These hazards show the hazard rate by time at the mean value for all covariates)



***** [Another option: Piecewise constant hazards model :  ]
*  [requires installation of associated ado file to run 'stpiece' command]
*  [see http://econpapers.repec.org/software/bocbocode/s396801.htm ]
******.


************************************************************.



************************************************************.
************************************************************.
```



```
************************************************************.
*** EXERCISE 4: EVENT HISTORY ANALYSIS IN DISCRETE TIME .
************************************************************.



use $path1\ljemp_ext1.dta, clear
numlabel _all, add
gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
gen agestart=syear - doby2
keep if agestart >= 16 & agestart <= 30
summarize dur hgs fem  agestart  syear


stset duration, failure(status) id(pid)
stdes

* Current format is 1 record per episode; episode lengths in months, up to 40 years
* To convert to discrete time, want one record per person per time unit

sav $path9\eh1.dta, replace

** i) discrete time every month

* (this requires a lot of memory - may need to increase memory)
clear
set mem 164m
use $path9\eh1.dta, clear
gen statusi=status
stdes
stcox hgs fem agestart syear
summarize pid dur

stsplit distim, every(1)

stdes
summarize pid dur distim _st _d statusi
list  pid dur distim _st _d statusi in 1/200

logistic _d distim hgs fem agestart syear
* Linear hazard dependence

gen distim2=distim^2
logistic _d distim distim2 hgs fem agestart syear
* Quadratic hazard dependence

logit _d distim hgs fem agestart syear if statusi==1
* Excluding censored cases is not necessary: _d inherently controls for censoring

** ii) discrete time every 6 months

use $path9\eh1.dta, clear
gen statusi=status
stdes
stcox hgs fem agestart syear
summarize pid dur

stsplit distim, every(6)

stdes
summarize pid dur distim _st _d statusi
list  pid dur distim _st _d statusi in 1/200

logistic _d distim hgs fem agestart syear
* Linear hazard dependence

gen distim2=distim^2
logistic _d distim distim2 hgs fem agestart syear
* Quadratic hazard dependence
```

```
************************************************************.


************************************************************.
************************************************************.




************************************************************.
*** Exercise 5: EVENT HISTORY ANALYSIS AND COMPETING RISKS .
************************************************************.

** Comment: Stata doesn't have purpose built competing risks models
*   at present (they may be on the horizon..), but it is legitimate
*   simply to treat a competing risk as an additional covariate, eg:

use $path1\ljemp_ext1.dta, clear
numlabel _all, add
tab stemp
tab stgold
tab nextemp
gen next3=nextemp
recode next3 1/3=1 4=2 5/11=3 *=-999
mvdecode next3, mv(-999)
label define next3l 1 "Employed" 2 "Unemployed" 3 "Not working"
label values next3 next3l
numlabel next3l, add
tab next3
tab stgold
keep if (stgold >= 1 & stgold <= 11 & next3 >= 1 & next3 <= 3)
summarize
* (analysis: influences on lenght of first job, conditional upon how it ends

stset duration, failure(status)
stdes

* Describe differences of outcomes :

stci, by(sex next3) rmean

stci if sex==1, by(stgold next3) rmean

sts graph if stgold >= 1, by(next3 sex)  ///
   title("Kaplan-Meir Survival Times ", size(large) ) ///
      subtitle("Time in first job, by destination state and gender") ///
    note("Duration in months of first job after leaving education") ///
              legend( order(1 2 3 4 5 6)   ///
           label(1 "Males, employed")    label(2 "Males, unemployed")   ///
           label(3 "Males, not in work")   label(4 "Females, employed")   ///
             label(5 "Females, unemployed") label(6 "Females, not in work") ) ///
    clpattern(solid dash dot dash_dot longdash_dot shortdash )  ///
    clwidth(thin thin medthick medium medthick medthick )


* Models according to differences in outcome:

gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
gen agestart=syear - doby2
keep if agestart >= 16 & agestart <= 30
summarize
tab next3

stcox hgs fem agestart syear if next3==1
est store employed
stcox hgs fem agestart syear if next3==2
est store unemployed
stcox hgs fem agestart syear if next3==3
est store notwork
```



```
est table employed unemployed notwork, star stats(N ll bic)


************************************************************.


************************************************************.
************************************************************.




*****************************************************************************
** Exercise 6) Evaluating alternative parametric event history models :
*****************************************************************************


use $path1\ljemp_ext1.dta, clear
numlabel _all, add
gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
gen agestart=syear - doby2
keep if agestart >= 16 & agestart <= 30
stset duration, failure(status)
stdes
gen agestz=agestart - 25
gen syearz=syear - 1960
summarize dur hgs fem  agestart  syear  agestz syearz


** Choosing between parametric event history models :
** Model comparison : use the AIC criteria

streg hgs fem agestart syear , distribution(exponential)
est store expon
scalar e_ll=e(ll)
scalar e_aic=-2*e(ll) + 2*(4 + 1 + 0)

streg hgs fem agestart syear , distribution(weibull)
est store weibull
scalar w_ll=e(ll)
scalar w_aic=-2*e(ll) + 2*(4 + 1 + 1)

streg hgs fem agestart syear , distribution(gompertz)
est store gompertz
scalar g_ll=e(ll)
scalar g_aic=-2*e(ll) + 2*(4 + 1 + 1)

streg hgs fem agestart syear , distribution(lognormal)
est store lognormal
scalar n_ll=e(ll)
scalar n_aic=-2*e(ll) + 2*(4 + 1 + 1)


streg hgs fem agestart syear , distribution(loglogistic)
est store loglogist
scalar l_ll=e(ll)
scalar l_aic=-2*e(ll) + 2*(4 + 1 + 1)

streg hgs fem agestart syear , distribution(gamma)
est store gamma
scalar ga_ll=e(ll)
scalar ga_aic=-2*e(ll) + 2*(4 + 1 + 2)

display e_ll
display w_ll
display g_ll
display n_ll
display l_ll
```

```
display ga_ll

display e_aic
display w_aic
display g_aic
display n_aic
display l_aic
display ga_aic

* Here the generalised gamma model is favoured in all circumstances

est table expon weibull lognormal  gamma, star stats(N ll)
* Note that the distribution specifications do impact the coefficient estimates.

**



*****************************************************************
*******************************************
**** Exercise 7) Distinguishing Age, period and cohort effects in duration models
*****************************************************************



use $path1\ljemp_ext1.dta, clear
numlabel _all, add
gen hgs=sthgs if sthgs >= 1
gen fem=(sex==2)
gen doby2=doby if (doby >= 1)
gen agestart=syear - doby2
scatter agestart syear
keep if agestart >= 16 & agestart <= 30
summarize
stset duration, failure(status)
stdes

summarize doby2 syear date

correlate syear doby2 agestart
graph matrix syear doby2 agestart


gen lnages=ln(agestart)
histogram lnages
gen ages21=agestart >= 21
gen sy90=syear >= 1990
gen sy70=syear <= 1970
histogram syear



gen agesy=agestart*syear
gen agedob=agestart*doby2
gen dobsy=doby2*syear
gen cohort1=0
replace cohort1=1 if doby2>1932 & doby2<1956
gen cohort2=0
replace cohort2=1 if doby2>1955
summarize hgs fem syear doby2 agestart agesy agedob cohort1 cohort2

* A series of models capturing different age, period, cohort effects
stcox hgs fem syear
stcox hgs fem  doby2
stcox hgs fem agestart

stcox hgs fem syear doby2
stcox hgs fem syear cohort1 cohort2
stcox hgs fem syear doby2 dobsy
stcox hgs fem syear agestart
stcox hgs fem syear agestart agesy
stcox hgs fem agestart doby2
stcox hgs fem agestart doby2 agedob

* Age, period and cohort (this model is only identified because of the categorical cohort data:.
```



```
stcox hgs fem syear agestart cohort1 cohort2


*************************


*****************************************************************
*****************************************************************


capture log close


**** EOF .



*********************************************************.
```