Practical session 3:
Causal Survival Analysis using Stata

Network of Applied Statisticians in Health

Objectives

By the end of this practical, you will be able to:

1. Implement intention to treat analysis (ITT) and per-protocol (PP) analysis using inverse probability of weighting (IPW)
2. Estimate hazard ratios for the intention to treat and per-protocol analysis using pooled logistic regression
3. Estimate standardised risk curves (or survival curves) using pooled logistic regression and IPW for the intention to treat and the per-protocol analysis

We will use the "Practical TTE (causal survival analysis).dta" dataset, which contains information on 5K individuals on

a) ID
b) Actual time (in years)
c) Time (in years): "ceiling" of time, i.e. ceil(x): the unique integer n such that n - 1 < x ≤ n
d) A: exposure
e) L: continuous (time-dependent) confounder
f) Age of participants, at baseline
g) Sex
h) Y: Observed outcome (e.g CVD: yes/no) between at time (t-1 and t]
i) Y_t1: Observed outcome (e.g CVD: yes/no) at t+1, i.e. between at time (t and t+1]

```
  obs:        26,615
 vars:             9                          27 Nov 2022 14:24
 size:       958,140
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              float   %9.0g                 ID
actual_time     float   %9.0g                 actual time (in years)
time            float   %9.0g                 time (in years)
A               float   %9.0g                 Observed exposure at time t
male            float   %9.0g                 male
age             float   %9.0g                 age (in years)
L               float   %9.0g                 Observed t-v confounder at time t
Y               float   %9.0g                 Observed outcome Y between at time (t-1 and t]
Y_t1            float   %9.0g                 Observed outcome Y between at time (t and t+1]
-------------------------------------------------------------------------------
Sorted by: id  time
```

The dataset is in long format

Part A: Estimate hazard ratios for ITT and PP analysis

Use the script "Final dofile (Hazard Ratios for ITT and PP analysis).do"

1. **Load and describe the data, familiarize yourself with the dataset**

   Run in Stata

   browse
   desc

2. **Estimate the hazard ratio of initiating treatment A on Y using Cox regression by adjusting for baseline confounders in the outcome model [observational analogue of ITT analysis]**

   It is easier to convert the dataset in an one observation per subject format

3. **Estimate the hazard ratio of initiating treatment A on Y using pooled logistic regression by adjusting for baseline confounders in the outcome model for ITT analysis**

   Use the dataset and note that variable Y_t1 holds the value of Y at (t+1)

   Create a variable for Y_{k+1}, i.e. whether outcome would occur in the next time period (in this example, year) outcome

   Run $logit(\Pr(Y_{k+1} = 1 | A_0, L_0, \overline{Y_k} = 0)) = a_{0,t} + a_1 A_0 + a_2^T L_0$

   Y_{k+1}:outcome at next time period
   A_0: exposure at time 0
   L_0: confounders at time 0
   $c_{0,t}$: time-varying intercept, i.e. $a_{0,t} = b_0 + b_1 * t + b_2 * t^2 + b_3 * t^3$

4. **Estimate the hazard ratio of initiating treatment A on Y using pooled logistic regression by adjusting for baseline confounders through (stabilised and unstabilised) IPW [intention-to-treat (ITT) analysis]**

   For unstabilised weights

   Run $logit(\Pr(A_0 = 1 | L_0)) = c_0 + c_1^T L_0$

   Predict $\Pr(A_0 = 1 | L_0)$,
   estimate IPW=1/$\Pr(A_0 = 1 | L_0)$ if $A_0 = 1$
   or IPW=1/(1-$\Pr(A_0 = 1 | L_0)$) if $A_0 = 0$

   Run weighted regression (weighted by IPW)

   $$logit(\Pr(Y_{k+1} = 1 | A_0, \overline{Y_k} = 0) = d_{0,t} + d_1 A_0$$

   See Stata script for stabilized weights

5. Do we expect any differences in the results between questions 1,2 and 3

6. Estimate the hazard ratio of adhering on treatment A on Y using pooled logistic regression by adjusting for baseline and time-dependent confounders through (unstabilised) inverse probability of adherence weights [per-protocol analysis]

For this analysis you need to

- Censor individuals who do not adhere
- Use pooled logistic regression models to estimate the probability of adherence weights (IPW_A). These weights will create a pseudo-population where everyone either continuously adheres to treatment or no treatment
- Use pooled logistic regression to model the outcome, weighted by IPW_A

You need to calculate unstabilised weights for censoring due to switching treatment

$$IPW\_A_k = \prod_{k=0}^{t} \frac{1}{f(A_k|\overline{A_{k-1}}, \overline{L_k}, \overline{Y_k} = 0)}$$

To estimate the denominator $f(A_k|\overline{A_{k-1}}, \overline{L_k}, \overline{Y_k} = 0)$ we run two models

The 1st model was fit to person-times who were untreated in the previous time point (that is, $A_{k-1} = 0$):

$$logit(\Pr(A_k = 1|\, A_{k-1} = 0, \overline{L_k}, \overline{Y_k} = 0)) = e_0 + e_1^T L_0 + e_2^T L_k$$

Then IPW_A at each time-point for the untreated will be

$$IPW\_A_{k,\ untreated} = \frac{1}{1 - (\Pr(A_k = 1|\, A_{k-1} = 0, \overline{L_k}, \overline{Y_k} = 0)}$$

The 2nd model was fit to person-times who were treated in the previous time point (that is, $A_{k-1} = 1$):

$$logit(\Pr(A_k = 1|\, A_{k-1} = 1, \overline{L_k}, \overline{Y_k} = 0)) = f_0 + f_1^T L_0 + f_2^T L_k$$

Then IPW_A at each time-point for the treated will be

$$IPW\_A_{k,\ treated} = \frac{1}{(\Pr(A_k = 1|\, A_{k-1} = 0, \overline{L_k}, \overline{Y_k} = 0)}$$

Covariate history $\overline{L_k}$ is summarized by baseline L_0 and the most recent measurement of L_k

You then multiply IPW_A across all time points for each individuals

(you need to truncate weights to 25 if max(IPW_A)>25 and p99(IPW_A)<25)

Or run weighted pooled logistic regression, weighted by IPW_A

$logit(\Pr(Y_{k+1} = 1 | A_0, L_0, \overline{Y_k} = 0, \overline{C_k} = 0)) = h_{0,t} + h_1 A_0 + h_2^T L_0$ among the uncensored individuals

7. What is the difference in the interpretation of the results of questions 3 and 5?

8. Summarise your findings in the following table

Table: Hazard ratios of the ITT and PP analysis

	HR (95% CI)
Effect of initiating treatment A on Y using Cox regression and adjusting for baseline confounders in the outcome model for ITT analysis (Q2)	
Effect of initiating treatment A on Y using pooled logistic regression and adjusting for baseline confounders in the outcome model for ITT analysis (Q3)	
Effect of initiating treatment A on Y using pooled logistic regression and adjusting for baseline confounders through unstabilised IPW for the intention-to-treat (ITT) analysis (Q4.a)	
Effect of initiating treatment A on Y using pooled logistic regression and adjusting for baseline confounders through stabilised IPW for the intention-to-treat (ITT) analysis (Q4.b)	
Effect of adhering on treatment A on Y using pooled logistic regression for the per-protocol (PP) analysis using unstabilised inverse probability of adherence weights for adherence (Q6)	

Part B: Estimate standardized risk curves

Use the script "Final dofile (Standardised risk curves).do"

In this script, we will estimate standardized risk curves for the PP analysis. The procedure is similar for the ITT analysis (changes only the initial model in step 1)

1. Fit a discrete-time hazards model (eg, a pooled logistic model with relatively short periods) that estimates, at each time and for each person, the probability of developing the outcome. Use the variable "time of follow-up,", along with time^2 and time^3 or a flexible functional form (splines) and estimate time-varying HRs by adding product terms between exposure and "time of follow-up."

2. Recreate all time points and two exposure groups for each subject (we want to estimate their risk curves both under exposure and under no exposure, regardless of the subject's exposure status)

3. Estimate the probability of being at risk at each time point

4. Estimate the probability of being healthy at each time point (1-probability of being at risk)

6. Multiply the model's predicted values of being healthy through time *t* to estimate the probability of remaining healthy (or probability of survival if the outcome is death) at *t* for subjects with their same combination of covariate values.

6. Predict the probability of being at risk at time *t* for each subject both under exposure and under no exposure, regardless of the subject's exposure status (1- probability of remaining healthy)

7. Separately average the adjusted risk curves under exposure and under no exposure, over all subjects. This last step effectively standardizes the curves to the empirical distribution of the covariates in the study, and results in 2 marginal risk curves: one under exposure, another under no exposure.

8. Estimate 95% CI using bootstrap