```
/*

PHD PROJECT: The role of depressive symptoms and cardiometabolic risk factors in the prediction of dementia: a cross-country comparison in England, the United States and China

STUDY 3: Trajectories of depressive symptoms and their relationship with dementia

Method of analysis:
Group-based trajectory modelling (GBTM) approach
Latent Class Growth Analysis (LCGA)


TIMELINE

DEPRESS TRAJECTORIES: WV8 - WV11 (4 TIME POINTS)
DEMENTIA INCIDENCE AT YEAR 6 (WV11)
DEMENTIA INCIDENCE: W12 - WV14 (3 TIME POINTS)



*/




* importing data (.dta)

use "S:\Research\pkstudies\Study1_traj_depression\HRS\1. hrs traj depress dementia risk\hrs_data_w8_14_study1.dta"



* KEEP NECESSARY VARIABLES


keep HHID PN RAHHIDPN HHIDPN ///
H_sex H_age H_eduaction_yrs H_eduaction H_maritalstatus_3cat H_maritalstatus_4cat H_wealthquintiles ///
H_ethnicity H_hispanic_ethnicity ///
H_smoking_2cat H_smoking_3cat H_physicalactivity H_alcohol_freq H_alcohol_status ///
H_heart_disease H_stroke H_cvd_comorbidity Hwv8_cognition ///
Hwv8_cesd_depressed Hwv8_cesd_effort Hwv8_cesd_sleep Hwv8_cesd_happy Hwv8_cesd_lonely Hwv8_cesd_enlife Hwv8_cesd_sad Hwv8_cesd_going Hwv8_cesd_score cesd_0 Hwv8_depressive_symptoms ///
Hwv9_cesd_depressed Hwv9_cesd_effort Hwv9_cesd_sleep Hwv9_cesd_happy Hwv9_cesd_lonely Hwv9_cesd_enlife Hwv9_cesd_sad Hwv9_cesd_going Hwv9_cesd_score cesd_1 Hwv9_depressive_symptoms ///
Hwv10_cesd_depressed Hwv10_cesd_effort Hwv10_cesd_sleep Hwv10_cesd_happy Hwv10_cesd_lonely Hwv10_cesd_enlife Hwv10_cesd_sad Hwv10_cesd_going Hwv10_cesd_score cesd_2 Hwv10_depressive_symptoms ///
Hwv11_cesd_depressed Hwv11_cesd_effort Hwv11_cesd_sleep Hwv11_cesd_happy Hwv11_cesd_lonely Hwv11_cesd_enlife Hwv11_cesd_sad Hwv11_cesd_going Hwv11_cesd_score cesd_3 Hwv11_depressive_symptoms ///
Hwv12_cesd_depressed Hwv12_cesd_effort Hwv12_cesd_sleep Hwv12_cesd_happy Hwv12_cesd_lonely Hwv12_cesd_enlife Hwv12_cesd_sad Hwv12_cesd_going Hwv12_cesd_score Hwv12_cesd_sumscore Hwv12_depressive_symptoms ///
Hwv13_cesd_depressed Hwv13_cesd_effort Hwv13_cesd_sleep Hwv13_cesd_happy Hwv13_cesd_lonely Hwv13_cesd_enlife Hwv13_cesd_sad Hwv13_cesd_going Hwv13_cesd_score Hwv13_cesd_sumscore Hwv13_depressive_symptoms ///
Hwv14_cesd_happy Hwv14_cesd_enlife Hwv14_cesd_depressed Hwv14_cesd_effort Hwv14_cesd_sleep Hwv14_cesd_lonely Hwv14_cesd_sad Hwv14_cesd_going Hwv14_cesd_sumscore Hwv14_depressive_symptoms ///
Hwv8_memory_report Hwv9_memory_report Hwv10_anydementia_report Hwv11_anydementia_report Hwv12_anydementia_report Hwv13_anydementia_report Hwv14_anydementia_report ///
Hwv8_interview_date Hwv9_interview_date Hwv10_interview_date Hwv11_interview_date Hwv12_interview_date Hwv13_interview_date Hwv14_interview_date ///
Hwv12to14_dementia_sum Hwv12to14_dementia_event ///
Hwv12to14_newdementia_or_lastint Hwv12to14_time_dementia_months Hwv12to14_dementia_free_date Hwv12to14_time_dementia_midpoint Hwv12to14_time_dementia_midpoin0 Hwv12to14_time_of_event_dementia ///
t_0 t_1 t_2 t_3 nmisfollowup_cesd nmisfollowup_dementia_wv12to14
```

```
/* Latent class growth analysis (LCGA)  of depressive symptoms */


* installing traj command

net from http://www.andrew.cmu.edu/user/bjones/traj
net install traj, force
help traj



* Generate a set of time variables to pass to traj, from wave 2 to 9 -> 8 time points (t0-t7)
forval i = 0/3 {
  generate t_`i' = `i'
}

*recode time in months

recode t_1 (1=24)
recode t_2 (2=48)
recode t_3 (3=72)


*rename cesd score across the waves - discrete var min=0 max=8
* use Hwv8_cesd_sumscore continuous variables

rename Hwv8_cesd_sumscore cesd_0
rename Hwv9_cesd_sumscore cesd_1
rename Hwv10_cesd_sumscore cesd_2
rename Hwv11_cesd_sumscore cesd_3






*** Descriptive stats of depression and dementia

tabulate cesd_0
summarize cesd_0 , detail
histogram cesd_0, discrete frequency normal

misstable summarize cesd_0
misstable patterns cesd_0

tabulate Hwv8_depressive_symptoms
summarize Hwv8_depressive_symptoms

misstable summarize Hwv8_depressive_symptoms
misstable patterns Hwv8_depressive_symptoms

tabulate Hwv8_memory_report
summarize Hwv8_memory_report

misstable summarize Hwv8_memory_report
misstable patterns Hwv8_memory_report


tabulate Hwv9_memory_report
summarize Hwv9_memory_report

misstable summarize Hwv9_memory_report
misstable patterns Hwv9_memory_report
```

```
tabulate Hwv10_anydementia_report
summarize Hwv10_anydementia_report

misstable summarize Hwv10_anydementia_report
misstable patterns Hwv10_anydementia_report

tabulate Hwv11_anydementia_report
summarize Hwv11_anydementia_report

misstable summarize Hwv11_anydementia_report
misstable patterns Hwv11_anydementia_report


tabulate Hwv12_anydementia_report
summarize Hwv12_anydementia_report

misstable summarize Hwv12_anydementia_report
misstable patterns Hwv12_anydementia_report


tabulate Hwv13_anydementia_report
summarize Hwv13_anydementia_report

misstable summarize Hwv13_anydementia_report
misstable patterns Hwv13_anydementia_report

tabulate Hwv14_anydementia_report
summarize Hwv14_anydementia_report

misstable summarize Hwv14_anydementia_report
misstable patterns Hwv14_anydementia_report








*** CLEANING DATA

* 1. drop missing data depression and dementia at baseline
* drop 48 depression missing data
* no missing data for baseline dementia

drop if cesd_0== .
* (48 observations deleted)


* 2. drop dementia cases between wv8 and wv11 (total: 547 cases)

drop if Hwv8_memory_report==1
* (222 observations deleted)

drop if Hwv9_memory_report==1
* (109 observations deleted)

drop if Hwv10_anydementia_report==1
* (104 observations deleted)



* 3. process to drop missing data depression in at least 2 follow-up waves

/*

```

```
check below how to see number of missing values in an observation (case) and patterns of missing values
https://stats.idre.ucla.edu/stata/faq/how-can-i-see-the-number-of-missing-values-and-patterns-of-missing-values-in-my-data-file/
install packages:
* install mdesc
* install tabmiss
* insatll dm31
* insall mvpatterna

*/

search mdesc
search rmiss2
search mvpatterns


* see number of missing values vs non-missing in each variable
mdesc cesd_0 cesd_1 cesd_2 cesd_3
*or
mdesc cesd_*


* number of missing values per observation
* the code below creates a variable called nmisfollowup that gives the number of missing values for each observation in the variables of interest
egen nmisfollowup_cesd=rmiss2(cesd_1 cesd_2 cesd_3)

tab nmisfollowup_cesd


* drop observations "nmisfollowup_cesd" > 1 (those with 2 or 3 follow ups of missing data)
drop if nmisfollowup>1
*(917 observations deleted)




* 4. drop obs with no records on dementia at any wave from 12-14 follow-ups

* see number of missing values vs non-missing in each variable
mdesc Hwv12_anydementia_report Hwv13_anydementia_report Hwv14_anydementia_report



/* number of missing values per observation
* the code below creates a variable called nmisfollowup that gives the number of missing values
for each observation in the variables of interest */
egen nmisfollowup_dementia_wv12to14=rmiss2(Hwv12_anydementia_report Hwv13_anydementia_report ///
Hwv14_anydementia_report)

tab nmisfollowup_dementia_wv12to14

* drop observations "nmisfollowup_dementia_wv6to9" > 2 (with 3 followups of missing data = no records at any wave)
drop if nmisfollowup_dementia_wv12to14>2
*(748 observations deleted)






*descriptive stats of depressive symptoms cesd


tabulate cesd_0
summarize cesd_0, detail
```

```
histogram cesd_0, discrete frequency normal

tabulate cesd_1
summarize cesd_1 , detail
histogram cesd_1, discrete frequency normal

tabulate cesd_2
summarize cesd_2, detail
histogram cesd_2, discrete frequency normal

tabulate cesd_3
summarize cesd_3, detail
histogram cesd_3, discrete frequency normal


ta cesd_0, miss
ta cesd_1, miss
ta cesd_2, miss
ta cesd_3, miss


tabstat cesd_0, by(H_sex)stats (mean v n)
tabstat cesd_1, by(H_sex)stats (mean v n)
tabstat cesd_2, by(H_sex)stats (mean v n)
tabstat cesd_3, by(H_sex)stats (mean v n)



tabstat cesd_*,s(n me sk) by(H_sex)




bysort H_sex: tab cesd_0
bysort H_sex: tab cesd_1
bysort H_sex: tab cesd_2
bysort H_sex: tab cesd_3


tabstat cesd_0 cesd_1 cesd_2 cesd_3, s(sk kur)
sktest cesd_0 cesd_1 cesd_2 cesd_3



* missingness pateterns
misstable patterns cesd_*
* “1” means that the variable is observed and a “0” represents missing


* box plots of the observations at each occasion
graph box cesd_0 cesd_1 cesd_2 cesd_3, ascategory intensity (0) medtype (line)










/*
LCGA analysis


useful sources
use http://www.andrew.cmu.edu/user/bjones/traj/data/cambrdge.dta,clear
https://ssrc.indiana.edu/doc/wimdocs/2013-03-29_nagin_trajectory_stata-plugin-info.pdf

```

```

traj [if], var(varlist) indep(varlist) model(modeltype)
        order(numlist) [additional options]


order(numlist)      0=intercept, 1=linear, 2=quadratic, 3=cubic -
                            polynomial type for each group trajectory


 ci                  parametric bootstrap confidence intervals of
                            individual distal outcome and probability of
                            group memberships.










Available Models -> command traj


Censored normal (CNORM) model distribution

traj, var(qcp*op) indep(age*) model(cnorm) min(0) max(999) order(1 3 2)

trajplot, xtitle(Age) ytitle(Opposition) xlabel(6(1)15)
      ylabel(0(1)6)

/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG3 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader

/* Including time-stable covariates (risk) associated with group membership */

traj, var(qcp*op) indep(age*) model(cnorm) min(0) max(10) order(1 3
          2) risk(scolmer scolper)



Zero Inflated poisson (ZIP) Model

It is an analysis of Poisson data with extra zeros


traj, model(zip) var(y*) indep(t*) order(2 1 3) iorder(1)

trajplot, xtitle(Age) ytitle(Opposition) ci



 Time-Stable Covariates for Group Membership

    traj, var(qcp*op) indep(age*) model(cnorm) min(0) max(10) order(1 3
          2) risk(scolmer scolper)

    trajplot, xtitle(Age) ytitle(Opposition)


Logistic (logit) model

    use http://www.andrew.cmu.edu/user/bjones/traj/data/cambrdge.dta,
          clear

    traj, var(p1-p23) indep(tt1-tt23) model(logit) order(0 3 3)
```

```
    trajplot, xtitle(Scaled Age) ytitle(Prevalence)

    /* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG3 if _n > 400, ab(12)




Model selection:

1. Type of model: The ‘traj’ can model normal, censored normal, zero-inflated Poisson and binary logit models.
Capacity for incorporating effect of time-stable and time-varying covariates,
subsequent outcomes and joint trajectory models.

2. Number of groups/classes: determination of the optimal number of groups to compose the mixture

3. Shape of the trajectory: determination of the appropriate order of the
polynomial used to model each group's trajectory (linear, quadratic, cubic).



Model Fit Criteria to select the model with optimal class enumeration:

• Bayesian Information Criteria (BIC), where lower BIC or least negative BIC
(higher value closer to zero) represents a better fitting model.

• Bayes Factor greater than 10 indicates very strong evidence
to use the “more complex” model.

• Meaningful proportion of participants within each class
(smallest group percentage to be higher or equal to 5%).

• Average posterior probability (APP) to belong to each class higher than 0.70.

• Entropy to determine the accuracy of classification of individuals into the different latent classes
If entropy is near 1.0, then classification of individuals is assumed to be adequate.
If entropy is near 0, then classification is assumed to be poor.







**********function to print out summary stats
program summary_table_procTraj
    preserve
    *look at the average posterior probability
    gen Mp = 0
    foreach i of varlist _traj_ProbG* {
        replace Mp = `i' if `i' > Mp
    }
    sort _traj_Group
    *and the odds of correct classification
    by _traj_Group: gen countG = _N
    by _traj_Group: egen groupAPP = mean(Mp)
    by _traj_Group: gen counter = _n
    gen n = groupAPP/(1 - groupAPP)
    gen p = countG/ _N
    gen d = p/(1-p)
    gen occ = n/d
    *Estimated proportion for each group
    scalar c = 0
    gen TotProb = 0
    foreach i of varlist _traj_ProbG* {
```

```
       scalar c = c + 1
       quietly summarize `i'
       replace TotProb = r(sum)/ _N if _traj_Group == c
    }
    gen d_pp = TotProb/(1 - TotProb)
    gen occ_pp = n/d_pp
    *This displays the group number [_traj_~p],
    *the count per group (based on the max post prob), [countG]
    *the average posterior probability for each group, [groupAPP]
    *the odds of correct classification (based on the max post prob group assignment), [occ]
    *the odds of correct classification (based on the weighted post. prob), [occ_pp]
    *and the observed probability of groups versus the probability [p]
    *based on the posterior probabilities [TotProb]
    list _traj_Group countG groupAPP occ occ_pp p TotProb if counter == 1
    restore
end

summary_table_procTraj



******* to generate a plot of the individual trajectories

preserve
reshape long count_ t_, i(id)

gen count_jit = count_ + ( 0.2*runiform()-0.1 )
graph twoway scatter count_jit t_, c(L) by(_traj_Group) msize(tiny) mcolor(gray) lwidth(vthin) lcolor(gray)


******* to calculate the Bayes factor

log Bayes factor (2loge(B10) ≈ 2(ΔBIC)
This estimate approximately equals 2(BICcomplex model−BICnull model)

*/






/*

Depressive symptoms (CES-D 8 item)
The trajectory groups of the CES-D scores (as a discrete variable) are tested
alone with time as the only independent variable, with no covariates added that could influence class membership.
The Zero Inflated poisson model (‘cnorm’) is applied, given that the CES-D 8-item was a count of symptoms and the majority of individuals scored 0 at each time point.


Initially, for each model, the linear, quadratic, and cubic functions of each trajectory can be tested,
depending on the number of time points.
To ensure parsimony, consistent with the recommendations of Helgeson, Snyder, and Seltman (2004),
non-significant cubic and quadratic terms are removed from trajectories in a given model,
but linear parameters are retained irrespective of significance.

I tested the best fitting model with two, three, four five and then six trajectories following the same process.
The models were compared (in a table of comparison) using BIC statistics,
Bayes factor, entropy, percentage of each class and average posterior probabilities.


PROCESS TO SELECT THE BEST-FITTING MODEL

Shape and Classes
```

```

1. run one traj with quadratic (order 2)
 - If quadratic is not significant run with linear parameter (order 1)

2. model with 2 traj with quadratic (order 2 2)
 - If neither traj is significant rerun with linear (order 1 1)
 - If one not significant adapt accordingly (e.g. order 1 2 OR order 2 1)

3. Compare models (complex-simple) with statistic criteria

4. Repeat the process with an increasing number of traj

*/







*** ZIP MODEL


* 1 class
traj, var(cesd_*) indep(t_*) model(zip) order(2) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci

/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG1 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader




* 2 classes
traj, var(cesd_*) indep(t_*) model(zip) order(2 2) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci

/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG2 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader





* 3 classes
traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci

/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG3 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader






* 4 classes
traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2 2) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
```

```
/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG4 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader



* 5 classes
traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2 2 2) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci



/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader


* 6 classes
traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2 2 2 2) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci



/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader




** the 5-class model fits better


* 5 classes
traj, var(cesd_*) indep(t_*) model(zip) order(3 3 3 3 3) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci



/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader



*** OPTIMAL ZIP MODEL

* 5 groups - zip model - cubic and linear polynomial (1 1 3 3 1)
traj, var(cesd_*) indep(t_*) model(zip) order(1 1 3 3 1) iorder(1)
trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci


/* Shows the assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)

/* trajT = x-axis, Avg# = data averages, Est# = model estimates */
    matrix list e(plot1), format(%9.2f) noheader

```

```
** run after each traj model to estimate the average posterior probability (APP) for each group

program summary_table_procTraj
    preserve
    *look at the average posterior probability
    gen Mp = 0
    foreach i of varlist _traj_ProbG* {
        replace Mp = `i' if `i' > Mp
    }
    sort _traj_Group
    *and the odds of correct classification
    by _traj_Group: gen cesdG = _N
    by _traj_Group: egen groupAPP = mean(Mp)
    by _traj_Group: gen counter = _n
    gen n = groupAPP/(1 - groupAPP)
    gen p = cesdG/ _N
    gen d = p/(1-p)
    gen occ = n/d
    *Estimated proportion for each group
    scalar c = 0
    gen TotProb = 0
    foreach i of varlist _traj_ProbG* {
       scalar c = c + 1
       quietly summarize `i'
       replace TotProb = r(sum)/ _N if _traj_Group == c
    }
    gen d_pp = TotProb/(1 - TotProb)
    gen occ_pp = n/d_pp
    *This displays the group number [_traj_~p],
    *the cesd per group (based on the max post prob), [countG]
    *the average posterior probability for each group, [groupAPP]
    *the odds of correct classification (based on the max post prob group assignment), [occ]
    *the odds of correct classification (based on the weighted post. prob), [occ_pp]
    *and the observed probability of groups versus the probability [p]
    *based on the posterior probabilities [TotProb]
    list _traj_Group cesdG groupAPP occ occ_pp p TotProb if counter == 1
    restore
end

summary_table_procTraj







/*
---- MODEL SELECTION ----
Best-fitting model to try survival analysis is the
5 class - order (1 1 3 3 1)
*/




/*
Data and variable manipulation
*/

* 5-class model: rename  _traj_Group to H_traj_group5
```

```
rename  _traj_Group H_traj_group5
recode H_traj_group5 (4=5) (5=4)
ta H_traj_group5
rename _traj_ProbG1 H_depres_traj_1
rename _traj_ProbG2 H_depres_traj_2
rename _traj_ProbG3 H_depres_traj_3
rename _traj_ProbG4 H_depres_traj_4
rename _traj_ProbG5 H_depres_traj_5


* labelling variable of H_traj_group5

label var H_traj_group5 "Traj 5 groups of depressive symptoms"

* labelling values
lab def traj_depres 1 "minimal" 2 "mild" 3 "increasing" 4 "decreasing" 5 "high"

* attach category labels to the variable through label value

lab val H_traj_group5 traj_depres

ta H_traj_group5





* Frequencies of covariates

tabulate H_age
summarize H_age


tabulate H_sex
summarize H_sex


tabulate H_eduaction
summarize H_eduaction


tabulate H_maritalstatus_4cat
summarize H_maritalstatus_4cat


tabulate H_wealthquintiles
summarize H_wealthquintiles


tabulate Hwv8_smoking_3cat
summarize Hwv8_smoking_3cat


tabulate Hwv8_physicalactivity
summarize Hwv8_physicalactivity


tabulate Hwv8_alcohol_status
summarize Hwv8_alcohol_status


tabulate Hwv8_cvd_comorbidity
summarize Hwv8_cvd_comorbidity


tabulate Hwv8_memory
summarize Hwv8_memory

```

```
/* MULTIPLE IMPUTATION (MI)

To handle with missing values of covariates


useful sources for MI and MICE:

https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/
https://www.stata.com/manuals/mi.pdf - see page 139
https://www.stata.com/meeting/switzerland16/slides/medeiros-switzerland16.pdf
https://www.youtube.com/watch?v=i6SOlq0mjuc&ab_channel=StataCorpLLC
https://dss.princeton.edu/training/MIStata.pdf



Preparing to conduct MI
1. examine the number and proportion of missing values among the variables of interest
    use the mdesc command
2. examine missing data patterns
    use commands mi set and mi misstable patterns
3. identify potential auxiliary variables


Run MI using chained equations (MICE)
using the commands
1. how (in what style) to store the imputations
 mi set wide
2. which variables will be imputed
 mi register imputed
3. optionally, which variables will not be imputed
 mi register regular
4. what imputation method is implemented to impute each of var - MICE
 mi impute chained

*/





/*

1. examining missing values
    install packages:
    * install mdesc
    * install tabmiss
    * insatll dm31
    * insall mvpatterna

*/

search mdesc
search rmiss2
search mvpatterns






* examining number of missing values vs non-missing in each variable

```

```
mdesc H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl Hwv12_crp ///
Hwv12_memory




* examining missing data patterns

mi set wide

mi misstable summarize H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl Hwv12_crp ///
Hwv12_memory


mi misstable patterns H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl Hwv12_crp ///
Hwv12_memory




/*
 identifying potential auxiliary var
* Auxiliary variables are either correlated with a missing variable(s)
(the recommendation is r > 0.4) or are believed to be associated with missingness
- a priori knowledge of var that would make good auxiliary var
- identify potential candidates by examining associations between missing var and other var in the dataset
    running correlation using the command: pwcorr v1 v2 v3, obs
    the recommnedation for good correlation is r > 0.4


Missing var to be imputed:

    Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status
    Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl
    Hwv12_memory



Potential auxiliary var:
DV:  Hwv12to14_dementia_event
IV:  H_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3
other var:
    H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles Hwv8_depressive_symptoms

*/


* correlation

pwcorr Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
    Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl ///
    Hwv12_memory ///
    Hwv12to14_dementia_event H_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3 ///
    H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
    Hwv8_depressive_symptoms, obs


* The correlation showed that all the above potential var are good auxiliary
* A good auxiliary does not have to be correlated with every variable to be useful
* And it's not problematic if it has missing info of it's own
```

```
/*
MI by chained equations (MICE)
    see: https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

MICE is known as the fully conditional specification or sequential generalized regression
does not assume a joint MVN distribution
but instead uses a separate conditional distribution for each imputed variable.

The multivariate normal (MVN) model - mi imputed mvn -
assumes multivariate normality of all var

The multivariate imputation by chained equations (MICE) - mi imputed chained -
offers flexibility in how each var is modeled

mi impute chained allows to specify models for a
variety of variable types, including
continuous, binary, ordinal, nominal, truncated, and count variables


The MICE distributions available in Stata are:
binary, ordered and multinomial logistic regression for categorical variables,
linear regression and predictive mean matching (PMM)* for continuous variables,
and Poisson and negative binomial regression for count variables.



IMPUTATION PHASES

1. mi set wide
    style to store imputations

2. mi register imputed
    identifies which variables in the imputation model have missing information.

3. mi register regular (! optional)
    which variables will not be imputed

4. mi impute chained
    where the user specifies the imputation model to be used
    and the number of imputed datasets to be created.
    Example:
        mi impute chained (regress) bmi age (logit) female ///
        (mlogit) race = bpdiast i.region, add(20)

5. mi estimate
    is used as a prefix to the standard regress command.
    This executes the specified estimation model within each of the 20 imputed datasets
    to obtain 20 sets of coefficients and standard errors.
    Stata then combines these estimates to obtain one set of inferential statistics.
    In the output from mi estimate you will see some metrics: Imputation Diagnostics
    information for RVI (Relative Increase in Variance),
    FMI (Fraction of Missing Information),
    DF (Degrees of Freedom) ,
    RE (Relative Efficiency),
    and the between imputation and the within imputation variance estimates
    to examine how the standard errors (SEs) are calculated.


----------------

SELECTING MY IMPUTATION MODEL

- MICE -> mi impute chained

- var to be imputed:
```

```

    linear regression for continuous var (regress) ->
    Hwv8_memory

    logistic for the binary var (logit) ->
    H_cvd_comorbidity


    multinomial logistic for our nominal categorical var (mlogit) ->
    H_smoking_3cat H_physicalactivity H_alcohol_status



- auxiliary var:

    DV -> Hwv12to14_dementia_event
    IV -> H_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3
    other covariates -> H_age H_sex H_eduaction H_maritalstatus_4cat
                        H_wealthquintiles Hwv8_depressive_symptoms




- imputation numbers (m) -> 20

    ELSA data were imputed 20 numbers

    White et al. (2010) recommendation: use the rule that m should equal the percentage of incomplete cases


- rseed (53421) for reproducability reasons


- (! OPTIONAL) advance impute options -> force

    proceed with imputation, even when missing imputed values (e.g. auxiliary have missing data) are encountered

- impute options -> savetrace (trace1)

    specifies Stata to save the means and standard deviations of imputed values from each iteration to a Stata dataset named “trace1
*/


mi set wide


mi register imputed H_smoking_3cat H_physicalactivity H_alcohol_status H_cvd_comorbidity Hwv8_memory



mi impute chained (logit) H_cvd_comorbidity  ///
(mlogit) H_smoking_3cat H_physicalactivity H_alcohol_status ///
(regress) Hwv8_memory = Hwv12to14_dementia_event H_traj_group4 Hwv8_depressive_symptoms ///
H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles, add(20) rseed(53421) savetrace(trace1)


* save imputed data


* plot imputations

*it will open a file named trace1
use trace1,clear
```

```
describe


reshape wide *mean *sd, i(iter) j(m)
tsset iter



/*
The trace plot below graphs the predicted means value produced during the first imputation chain.
As before, the expectations is that the values would vary randomly to incorporate variation into
the predicted values for read.
*/

tsline H_smoking_3cat_mean1, name(mice1,replace)legend(off) ytitle("Mean of smoking")
tsline H_physicalactivity_mean1, name(mice1,replace)legend(off) ytitle("Mean of physical activity")
tsline H_alcohol_status_mean1, name(mice1,replace)legend(off) ytitle("Mean of alcohol status")
tsline H_cvd_comorbidity_mean1, name(mice1,replace)legend(off) ytitle("Mean of cvd")
tsline Hwv8_memory_mean1, name(mice1,replace)legend(off) ytitle("Mean of memory")


/*

All imputation chains can also be graphed simultaneously to make sure that nothing unexpected
occurred in a single chain.
Every chain is obtained using a different set of initial values and this should be unique.
Each colored line represents a different imputation.
So all 10 imputation chains are overlaid on top of one another.

*/


tsline H_alcohol_status_mean*, name(mice1,replace)legend(off) ytitle("Mean of alcohol status")
tsline H_alcohol_status_sd*, name(mice2, replace) legend(off) ytitle("SD of alcohol status")
graph combine mice1 mice2, xcommon cols(1) title(Trace plots of summaries of imputed values)

* repeat for each imputed var







/*
---- DESCRIPTIVE STATISTICS ----

General characteristics of participants

General characteristics of participnats stratified for study inclusion

General characteristics of participants stratified for dementia occurence

Participant characteristics by depressive symptom trajectory group


1. CHI-SQUARE (chi2) for categorical var (crosstabulation)
    Frequency tables -> two-way tables
        using the command tabulate, chi2
        reporting observations, column percentage (N, %) and p-value of Pearson's r


2. one-way ANOVA for continuous var
    check box plot
    using the command oneway
    reporting mean, sd (summary tables) and p-value of F
*/
```

```
* General characteristics of HRS participants


* Demographics
sum H_age
ta H_sex
ta H_eduaction
ta H_maritalstatus_4cat
ta H_wealthquintiles
* Lifestyle and health factors
ta H_smoking_3cat
ta H_physicalactivity
ta H_alcohol_status
ta H_cvd_comorbidity
ta Hwv8_diabetes_reportevr
ta Hwv8_HbA1c
ta Hwv8_crp
ta Hwv12_hdl
ta Hwv8_obesity_waist
ta Hwv8_systolic_bp
ta Hwv8_diastolic_bp
* Depressive symptoms t1-t3 (cont and categ)
sum cesd_0
sum cesd_1
sum cesd_2
sum cesd_3
ta depress_0
ta depress_1
ta depress_2
ta depress_3
* Memory score
sum Hwv8_memory






* Sample characteristics by depressive symptom trajectories
* crosstabs categ var (frequencies and chi2) !report column percentage!
* oneway ANOVA cont var (mean, sd)

* Demographics
oneway H_age H_traj_group5, tabulate
ta H_sex H_traj_group5, chi2 column row
ta H_eduaction H_traj_group5, chi2 column row
ta H_maritalstatus_4cat H_traj_group5, chi2 column row
ta H_wealthquintiles H_traj_group5, chi2 column row
* Lifestyle and health factors
ta H_smoking_3cat H_traj_group5, chi2 column row
ta H_physicalactivity H_traj_group5, chi2 column row
ta H_alcohol_status H_traj_group5, chi2 column row
ta H_cvd_comorbidity H_traj_group5, chi2 column row
ta Hwv8_diabetes_reportevr H_traj_group5, chi2 column row
ta Hwv8_HbA1c H_traj_group5, chi2 column row
ta Hwv8_crp H_traj_group5, chi2 column row
ta Hwv12_hdl H_traj_group5, chi2 column row
ta Hwv8_obesity_waist H_traj_group5, chi2 column row
ta Hwv8_systolic_bp H_traj_group5, chi2 column row
ta Hwv8_diastolic_bp H_traj_group5, chi2 column row
* Memory score
oneway Hwv8_memory H_traj_group5, tabulate
```

```
/*
---- BINOMIAL LOGISTIC REGRESSION ON COMPLETE DATA ----

Command is:
logistic DV IVs
        OR
logit DV IVs, or


--- Building the model using baseline covariates ---

Model 1: unadjusted - single predictor of depressive symptom trajectories C_traj_group5
Model 2: model 1 + sociodemographics: age sex education marital status and wealth
Model 3: model 2 + health behaviours: smoking, alcohol consumption


*/



* Unadjusted model - model 1 - single predictor

logistic Hwv11_anydementia_report H_traj_group5

*OR

logit Hwv11_anydementia_report H_traj_group5, or



* define design var by using i.

logistic Hwv11_anydementia_report i.H_traj_group5

*OR

logit Hwv11_anydementia_report i.H_traj_group5, or


* Adjusted models - multivariable logistic regression
* controlling for covariates

* model 2: model 1 + adjust for demographics: age sex education marital status and wealth

logistic Hwv11_anydementia_report i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles

* model 3: model 2 + adjust for lifestyle factors

logistic Hwv11_anydementia_report i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
H_smoking_3cat H_alcohol_status H_cvd_comorbidity







/*
---- SURVIVAL ANALYSIS ----

Tests of proportional-hazards assumption
Kaplan Meier survival curves
Person-time
Cox proportional regression - Hazard ratios - stcox
Postestimation tools for stcox
```

```
Test of Goodness of Fit

*** Cox regression in full data, complete data (listwise deletion of missing data) and imputed data
Cox PH regression in complete data
Cox PH regression model in imputed dataset - mi estimate


*/



* check dataset variables of interest only

codebook Hwv12to14_time_of_event_dementia Hwv12to14_dementia_event H_traj_group4 ///
H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_freq  ///
Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl ///
Hwv12_loneliness_quintiles Hwv12_memory,compact



* Declare Data to be Survival Data
* Time to event: Hwv12to14_time_of_event_dementia (months)
* Censoring: Hwv12to14_dementia_event (1=dementia, 0=censored)
* Command is stset TIMETOEVENT, failure(CENSORVARIABLE)


stset Hwv12to14_time_of_event_dementia, failure (Hwv12to14_dementia_event==1) id(RAHHIDPN)



*describe survival data using commnad stsum

stsum

stsum, by(H_traj_group)




* Kaplan Meier Curve estimation

sts list

sts list, by(H_traj_group5)



* Kaplan Meier Curve Plot

* no frills plot

sts graph

* with frills

sts graph, xtitle("Time in Months") ytitle("Survival Prob") ///
title("Kaplan Meier Curve") subtitle("n=4475, # events=233") ///
caption("graph02.png", size(vsmall))


* With Greenwood CI limits

sts graph, gwood legend(off) xtitle("Time in Months") ytitle("Survival Prob") ///
title("Kaplan Meier Curve") subtitle("n=4475, # events=233") caption("graph03.png", size(vsmall))
```

```

* Group Kaplan-Meier Curve Estimation
* Command is sts graph, by(GROUPVAR) OPTION OPTION OPTION Note: Must have sorted by GROUPVAR first

sort H_traj_group5

sts list, by(H_traj_group5)

* graph with frills

sts graph, by(H_traj_group5) xlabel(0(20)120) ylabel(0.80(.05)1) xtitle("Time in Months") ///
ytitle("Survival Prob") title("Kaplan Meier Curve") subtitle("n=4475, # events=233") ///
caption("graph04.png", size(vsmall))




* calculate person-time and incidence rates using command ststime

stptime,title(Person-years)

stptime, title(Person-years) per(1000)


* calculate person-time  by category of H_traj_group5

stptime, by(H_traj_group5)

stptime, by(H_traj_group5) per(1000)


* calculate the median of the follow-up

sum Hwv12to14_time_of_event_dementia, detail


/* Log Rank Test of equality of survival distributions
 (NULL: equality of survival distributions among H_traj_group5 groups)
 We will consider including the predictor if the test has a p-value of 0.2 - 0.25 or less.
 If the predictor has a p-value greater than 0.25 in a univariate analysis
 it is highly unlikely that it will contribute anything to a model which includes other predictors.
 Command is sts test GROUPVAR
*/


sts test H_traj_group5, logrank

sts test H_age, logrank

sts test H_sex, logrank

sts test H_eduaction, logrank

sts test H_maritalstatus_4cat, logrank

sts test H_wealthquintiles, logrank

sts test H_smoking_3cat, logrank

sts test H_cvd_comorbidity, logrank





/* Cox PH regression model

```

```
using the command stcox

--- Building the model ---

Model 1: unadjusted - single predictor of depressive symptom trajectories E_traj_group4
Model 2: model 1 + sociodemographics: age sex education marital status and wealth
Model 3: model 2 + lifestyle and health : smoking, alcohol consumption, cvd

*/



* Unadjusted model - model 1 - single predictor

stcox H_traj_group5

* define design var by using i.(low, moderate, high, ref: minimal)

stcox i.H_traj_group5


* Adjusted models - multivariable Cox model
* controlling for covariates

* model 2: model 1 + adjust for demographics: age sex education marital status and wealth

stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles


* model 3: model 2 + adjust for lifestyle and health behaviours

stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity






* Coefficients instead of hazard ratios by specifing the option nohr

stcox i.H_traj_group5, nohr


stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status ///
i.H_cvd_comorbidity, nohr






* Multivariable model development
* Likelihood-ratio tests



*install eststo
findit eststo


* ---- rx controlling for age and sex -----*
quietly: stcox H_age i.H_sex
eststo modelagesex

quietly: stcox H_age i.H_sex i.H_traj_group5
eststo modelagesex_4group
```

```
lrtest modelagesex modelagesex_4group



* ---- rx controlling for sociodemographics -----*
quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles
eststo modelsociodemo

quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles i.
H_traj_group5
eststo modelsociodemo_4group

lrtest modelsociodemo modelsociodemo_4group


* ---- rx controlling for lifestyle and health behaviours-----*
quietly: stcox i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity
eststo modellifestyle

quietly: stcox i.H_smoking_3cat i.H_alcohol_status i.H_traj_group5
eststo modellifestyle_4group

lrtest modellifestyle modellifestyle_4group





* side-by-side comparison of models


quietly: stcox i.H_traj_group5
eststo model1


quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles i.
H_traj_group5
eststo model2



quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity i.H_traj_group4
eststo model3




* Display Betas and Summary Statistics
estout model1 model2 model3, stats(n chi2 bic, star(chi2)) prehead("Betas")

/* Key Interpretattion
Chi2 = Value of LR test comparing the model fit (“full”) to intercept only (“reduced”)
bic = Schwarz’ Bayesian Information Criterion = It is a function of the log-likelihood.
Smaller values indicate a better fit.
*/

* Display Hazard Ratios and Model Fit Statistics. Option eform produces hazard ratios
estout model1 model2 model3, eform stats(n chi2 bic, star(chi2)) prehead("Hazard Ratios")




* Postestimation tools for stcox

/* Test of proportional hazards

```

```
If the tests in the table are not significance (p-values over 0.05)
 then we can not reject proportionality and we assume
 that we do not have a violation of the proportional assumption.
*/

estat phtest, detail


/* Proportionality Assumption - method 1
We will check proportionality by including time-dependent covariates in the model
by using the tvc and the texp options in the stcox command.
Time dependent covariates are interactions of the predictors and time.
In this analysis we choose to use the interactions with log(time)
because this is the most common function of time used in time-dependent covariates
but any function of time could be used.
If a time-dependent covariate is significant this indicates
a violation of the proportionality assumption for that specific predictor.
The conclusion is that all of the time-dependent variables are not significant
either collectively or individually thus supporting the assumption of proportional hazard.
*/



stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity, nohr ///
tvc(H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
H_smoking_3cat H_alcohol_status H_cvd_comorbidity) texp(ln(Hwv12to14_time_of_event_dementia))



/* Proportionality Assumption - method 2
by using the Schoenfeld and scaled Schoenfeld residuals
In the stphtest command we test the proportionality of the model as a whole
and by using the detail option we get a test of proportionality for each predictor.
By using the plot option we can also obtain a graph of the scaled Schoenfeld assumption.
If the tests in the table are not significance (p-values over 0.05)
then we can not reject proportionality and we assume
that we do not have a violation of the proportional assumption.
The stphplot command uses log-log plots to test proportionality
and if the lines  in these plots are parallel then we have further indication
that the predictors do not violate the proportionality assumption.
*/

quietly stcox H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
H_smoking_3cat H_alcohol_status H_cvd_comorbidity, schoenfeld(sch*) scaledsch(sca*)
stphtest, detail
stphtest, plot(H_traj_group5) msym(oh)
stphtest, plot(H_age) msym(oh)
stphtest, plot(H_sex) msym(oh)
stphtest, plot(H_eduaction) msym(oh)
stphtest, plot(H_maritalstatus_4cat) msym(oh)
stphtest, plot(H_wealthquintiles) msym(oh)
stphtest, plot(H_smoking_3cat) msym(oh)
stphtest, plot(H_alcohol_status) msym(oh)
stphtest, plot(H_cvd_comorbidity) msym(oh)





stphplot, by(H_traj_group5) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_age) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_sex) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_eduaction) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_maritalstatus_4cat) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_wealthquintiles) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_smoking_3cat) plot1(msym(oh)) plot2(msym(th))
stphplot, by(H_alcohol_freq) plot1(msym(oh)) plot2(msym(th))
```

```
stphplot, by(H_cvd_comorbidity) plot1(msym(oh)) plot2(msym(th))


* Assessment of PH Assumption: adjust for age and sex
stphplot, by(H_traj_group5) adjust(H_age H_sex) nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
plot2opts(symbol(none) color(navy)) plot3opts(symbol(none) color(green)) plot4opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is H_tarj_group5") xtitle("months")



* Assessment of PH Assumption: adjust for model 2
stphplot, by(H_traj_group5) adjust(H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles) ///
nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
plot2opts(symbol(none) color(navy)) plot3opts(symbol(none) color(green)) plot4opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is H_tarj_group5") xtitle("months")



* Assessment of PH Assumption: adjust for model 3
stphplot, by(H_traj_group4) adjust(H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
H_smoking_3cat H_alcohol_status H_cvd_comorbidity) ///
nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
plot2opts(symbol(none) color(navy)) plot3opts(symbol(none) color(green)) plot4opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is H_tarj_group5") xtitle("months")






/* Test of overall goodness of fit
Goodness of fit of the final model
2 methods:
 - by using the commnad stcoxgof (good fit = non sig p-value)
 - by using the Cox-Snell residuals
    to create the Nelson-Aalen cumulative hazard function
    If the hazard function follows the 45 degree line then we know that it approximately
    has an exponential distribution with a hazard rate of one and that the model fits the data well.
    If the model fits the data, the plot of the cumulative hazard versus cs
    should approximate a straight line with slope 1.
*/


* by using the commnad stcoxgof

* install stcoxgof
findit stcoxgof


stcox H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
H_smoking_3cat H_alcohol_status H_cvd_comorbidity, mgale(mgale)


stcoxgof




* by using the Cox-Snell residuals

quietly stcox H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
```

```
H_smoking_3cat H_alcohol_status H_cvd_comorbidity
predict cs, csnell

* or

quietly stcox H_traj_group5
predict cs, csnell


stset cs, failure(Hwv12to14_dementia_event)
sts generate km = s
generate H = -ln(km)
line H cs cs, sort ytitle("") clstyle(. refline)










/* -------- BINOMIAL LOGISTIC REGRESSION IN IMPUTED DATASET using baseline covariates --------- *

Command is

mi estimate : logit DV IV, or

    OR

mi estimate: logistic DV IV

*/



* Unadjusted model - model 1 - single predictor

mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report H_traj_group5

*OR

mi estimate, eform("Odds Ratio"): logit Hwv11_anydementia_report H_traj_group5, or



* define design var by using i.

mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report i.H_traj_group5

*OR

mi estimate, eform("Odds Ratio"): logit Hwv11_anydementia_report i.H_traj_group5, or


* Adjusted models - multivariable logistic regression
* controlling for covariates

* model 2: model 1 + adjust for demographics: age sex education marital status and wealth

mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles

* model 3: model 2 + adjust for lifestyle factors

mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
H_smoking_3cat H_alcohol_status H_cvd_comorbidity
```

```
* -------- COX PH REGRESSION MODEL IN IMPUTED DATASET --------- *


* Declare Data to be Survival Data by using mi

mi stset Hwv12to14_time_of_event_dementia, failure (Hwv12to14_dementia_event==1) id(RAHHIDPN)


* Run Cox regression analysis in imputed dataset by using "mi estimate:"
* Building the Model: Model 1 (unadjusted), Model 2, Model 3, Model 4
* Interactions


* Model 1 (default coefficents)
mi estimate: stcox H_traj_group5

* Model 1: define design var by using i.(low, moderate, high, ref: minimal)
mi estimate: stcox i.H_traj_group5


* Model 1 ask for hazard ratio by using the option eform("Haz.Ratio")

mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5


* Model 2: sociodemographics

mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles


* Model 3: lifestyle factors

mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity






/*

*** SENSITIVITY ANALYSES ***

1) single assessment of depressive symptoms and dementia risk at t0 and t3
continuous var of CES-D 8 items (0-8)
model 3 was further adjusted for cesd_0 and cesd_3

2) LCGA logit trajectories with dichotomous variable


3) Complete data

*/



* 1) Single assessment of cesd_0 and cesd_3 at model 3
```

```
* IMPUTED DATA


* single assessment model 3 adjust for cesd_0 and cesd_3


mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity ///
cesd_0


mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity ///
cesd_3








/*
2) Logistic model LCGA

use  Hwv8_depressive_symptoms dichotomous variables (0-1)

Logistic (logit) model

    use http://www.andrew.cmu.edu/user/bjones/traj/data/cambrdge.dta,
          clear

    traj, var(p1-p23) indep(tt1-tt23) model(logit) order(0 3 3)

    trajplot, xtitle(Scaled Age) ytitle(Prevalence)

    /* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG3 if _n > 400, ab(12)

*/






*rename Ewv2_depressive_symptoms score across the waves

rename Hwv8_depressive_symptoms depress_0
rename Hwv9_depressive_symptoms depress_1
rename Hwv10_depressive_symptoms depress_2
rename Hwv11_depressive_symptoms depress_3


net from http://www.andrew.cmu.edu/user/bjones/traj
net install traj, force
help traj




*** LOGIT MODEL
```

```
* 1 class - logit model - quadratic polynomial (2)
traj, var(depress_*) indep(t_*) model(logit) order(2)
trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG1 if _n < 3, ab(12)




* 2 class - logit model - quadratic polynomial (2 2)
traj, var(depress_*) indep(t_*) model(logit) order(2 2)
trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG2 if _n < 3, ab(12)





* 3 class - logit model - quadratic polynomial (2 2 2)
traj, var(depress_*) indep(t_*) model(logit) order(2 2 2)
trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG3 if _n < 3, ab(12)





* 4 class - logit model - quadratic polynomial (2 2 2 2)
traj, var(depress_*) indep(t_*) model(logit) order(2 2 2 2)
trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG4 if _n < 3, ab(12)






* 5 class - logit model - quadratic polynomial (2 2 2 2 2)
traj, var(depress_*) indep(t_*) model(logit) order(2 2 2 2 2)
trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)





* The 4-model depressive traj is selected to be tested in different shapes.

* 4 class - logit model - quadratic polynomial (3 3 3 3)
traj, var(depress_*) indep(t_*) model(logit) order(3 3 3 3)
trajplot, xtitle(Time in Months) ylabel(0(.20)1) ytitle(Depressive symptom caseness)

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG4 if _n < 3, ab(12)
```

```
program summary_table_procTraj
    preserve
    *look at the average posterior probability
    gen Mp = 0
    foreach i of varlist _traj_ProbG* {
        replace Mp = `i' if `i' > Mp
    }
    sort _traj_Group
    *and the odds of correct classification
    by _traj_Group: gen cesdG = _N
    by _traj_Group: egen groupAPP = mean(Mp)
    by _traj_Group: gen counter = _n
    gen n = groupAPP/(1 - groupAPP)
    gen p = cesdG/ _N
    gen d = p/(1-p)
    gen occ = n/d
    *Estimated proportion for each group
    scalar c = 0
    gen TotProb = 0
    foreach i of varlist _traj_ProbG* {
       scalar c = c + 1
       quietly summarize `i'
       replace TotProb = r(sum)/ _N if _traj_Group == c
    }
    gen d_pp = TotProb/(1 - TotProb)
    gen occ_pp = n/d_pp
    *This displays the group number [_traj_~p],
    *the cesd per group (based on the max post prob), [countG]
    *the average posterior probability for each group, [groupAPP]
    *the odds of correct classification (based on the max post prob group assignment), [occ]
    *the odds of correct classification (based on the weighted post. prob), [occ_pp]
    *and the observed probability of groups versus the probability [p]
    *based on the posterior probabilities [TotProb]
    lis_traj_Group cesdGroupAPPcc_pp p TotProbounter == 1
    restore
end

summary_table_procTraj



ta_traj_Group

rec_traj_Group 1=2 3=4










* IMPUTED DATA: Logistic regression (Odds Ratio)

* Unadjusted model (model 1)

mi estimate, or: logistic H_any_dementia_report

* model 2: model 1 + adjust for demographics: age sex education marital status and wealth

mi estimate, or: logistic H_any_dementia_report ///
H_age_sex_educ_marital_wealth_quintiles

* model 3: model 2 + adjust for lifestyle and health factors

mi estimate, or: logistic H_any_dementia_report ///
```

```
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity







* IMPUTED DATA: Cox regression (Hazard Ratio)


* Declare Data to be Survival Data by using mi

mi stset Hwv12to14_time_of_event_dementia, failure (Hwv12to14_dementia_event==1) id(RAHHIDPN)



* Unadjusted model (model 1)

mi estimate, eform("Haz. Ratio"): stcox i._traj_Group

* Model 2: sociodemographics

mi estimate, eform("Haz. Ratio"): stcox i._traj_Group ///
 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles


* Model 3: lifestyle and health factors

mi estimate, eform("Haz. Ratio"): stcox i._traj_Group ///
H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity




* 3) complete data analysis (see above)



* -------------------------- *
```