```
/*

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: WV1 - WV4 (4 TIME POINTS)
DEMENTIA INCIDENCE AT YEAR 6: WV4

*/




* importing data (.dta)

use "S:\Research\pkhec\Study1_traj_depression\CHARLS\charls_depress_traj_final_4class_model_with_all_var complete data"




* KEEP NECESSARY VARIABLES

keep ID id_12char bloodweight ///
C_sex C_age C_eduaction_10level C_eduaction_harmon_3cat ///
C_eduaction C_maritalstatus_8cat C_maritalstatus_3cat C_maritalstatus_4cat ///
Cwv1_smoking_2cat Cwv1_smoking_3cat Cwv1_physicalactivity Cwv1_alcohol_freq Cwv1_alcohol_status ///
C_cvd_comorbidity Cwv1_antidepressant Cwv1_psycholog_treat Cwv1_anytreat_psyche ///
Cwv1_memory_wordrecall Cwv1_concetration_serial7 Cwv1_orientation_time ///
Cwv1_executive_drawpicture Cwv1_cognition Cwv1_wealthquintiles Cwv1_netwealth_quintiles ///
Cwv1_cesd_depressed Cwv1_cesd_effort Cwv1_cesd_sleep Cwv1_cesd_lonely ///
Cwv1_cesd_bother Cwv1_cesd_going Cwv1_cesd_mind Cwv1_cesd_fear Cwv1_cesd_happy ///
Cwv1_cesd_hope Cwv1_cesd_sumscore Cwv1_cesd_score Cwv1_depressive_symptoms ///
Cwv2_cesd_depressed Cwv2_cesd_effort Cwv2_cesd_sleep Cwv2_cesd_lonely ///
Cwv2_cesd_bother Cwv2_cesd_going Cwv2_cesd_mind Cwv2_cesd_fear Cwv2_cesd_happy ///
Cwv2_cesd_hope Cwv2_cesd_sumscore Cwv2_cesd_score Cwv2_depressive_symptoms ///
Cwv3_cesd_depressed Cwv3_cesd_effort Cwv3_cesd_sleep Cwv3_cesd_lonely ///
Cwv3_cesd_bother Cwv3_cesd_going Cwv3_cesd_mind Cwv3_cesd_fear Cwv3_cesd_happy ///
Cwv3_cesd_hope Cwv3_cesd_sumscore Cwv3_cesd_score Cwv3_depressive_symptoms ///
Cwv4_cesd_bother Cwv4_cesd_mind Cwv4_cesd_depressed Cwv4_cesd_effort ///
Cwv4_cesd_fear Cwv4_cesd_sleep Cwv4_cesd_lonely Cwv4_cesd_going Cwv4_cesd_happy ///
Cwv4_cesd_hope Cwv4_cesd_sumscore Cwv4_depressive_symptoms ///
Cwv2to4_dementia_sum Cwv2to4_dementia_event ///
Cwv1_crp_level Cwv1_crp Cwv1_hdl_level Cwv1_male_hdl Cwv1_female_hdl ///
Cwv1_meds_dyslipid Cwv1_anymeds_dyslipid Cwv1_dyslipid_evr Cwv1_dyslipid_diagnosed ///
Cwv1_dyslipid_report_sum Cwv1_dyslipid_report Cwv1_hdl_sum Cwv1_hdl_cholesterol ///
Cwv1_waist Cwv1_malewaist_ao Cwv1_femalewaist_ao Cwv1_obesity_waist_sum ///
Cwv1_obesity_waist Cwv1_bmi_score Cwv1_obesity_bmi Cwv1_waist_bmi_sum ///
Cwv1_obesity Cwv1_tg_level Cwv1_tg Cwv1_triglyc_sum Cwv1_triglyc Cwv1_1systolic_bp ///
Cwv1_1diastolic_bp Cwv1_2systolic_bp Cwv1_2diastolic_bp Cwv1_3systolic_bp Cwv1_3diastolic_bp ///
Cwv1_systolic_mean Cwv1_diastolic_mean Cwv1_systolic_bp Cwv1_diastolic_bp Cwv1_meds_bp ///
Cwv1_anymeds_bp Cwv1_bp_evr Cwv1_bp_diagnosed Cwv1_bp_report_sum Cwv1_bp_report Cwv1_bp_sum ///
Cwv1_bp Cwv1_glucose_level Cwv1_glucose Cwv1_HbA1c_level Cwv1_HbA1c Cwv1_diabetes_evr ///
Cwv1_diabetes_diagnosed Cwv1_diabetes_report_sum Cwv1_diabetes_report Cwv1_meds_diabetes ///
Cwv1_anymeds_diabetes Cwv1_glucose_diabetes_sum Cwv1_glycemia ///
Cwv3_crp_level Cwv3_crp Cwv3_hdl_level Cwv3_male_hdl Cwv3_female_hdl ///
Cwv3_meds_dyslipid Cwv3_anymeds_dyslipid Cwv3_dyslipid_evr Cwv3_dyslipid_diagnosed ///
Cwv3_dyslipid_report_sum Cwv3_dyslipid_report Cwv3_hdl_sum Cwv3_hdl_cholesterol ///
```

```
Cwv3_waist Cwv3_malewaist_ao Cwv3_femalewaist_ao Cwv3_obesity_waist_sum ///
Cwv3_obesity_waist Cwv3_bmi_score Cwv3_obesity_bmi Cwv3_waist_bmi_sum Cwv3_obesity ///
Cwv3_tg_level Cwv3_tg Cwv3_triglyc_sum Cwv3_triglyc Cwv3_1systolic_bp Cwv3_1diastolic_bp ///
Cwv3_2systolic_bp Cwv3_2diastolic_bp Cwv3_3systolic_bp Cwv3_3diastolic_bp Cwv3_systolic_mean ///
Cwv3_diastolic_mean Cwv3_systolic_bp Cwv3_diastolic_bp Cwv3_meds_bp Cwv3_anymeds_bp ///
Cwv3_bp_evr Cwv3_bp_diagnosed Cwv3_bp_report_sum Cwv3_bp_report Cwv3_bp_sum Cwv3_bp Cwv3_glucose_level ///
Cwv3_glucose Cwv3_HbA1c_level Cwv3_HbA1c Cwv3_diabetes_evr Cwv3_diabetes_diagnosed ///
Cwv3_diabetes_report_sum Cwv3_diabetes_report Cwv3_meds_diabetes Cwv3_anymeds_diabetes ///
Cwv3_glucose_diabetes_sum Cwv3_glycemia ///
Cwv4_smoking_2cat Cwv4_smoking_3cat Cwv4_physicalactivity ///
Cwv4_alcohol_freq Cwv4_alcohol_status Cwv4_cvd_comorbidity ///
Cwv4_memory_wordrecall Cwv4_concetration_serial7 Cwv4_orientation_time ///
Cwv4_executive_drawpicture Cwv4_cognition ///
Cwv1_dementia_report Cwv2_dementia_report Cwv3_dementia_report Cwv4_self_info_dementia ///
Cwv1_interview_date Cwv2_interview_date Cwv3_interview_date Cwv4_interview_date ///
Cwv2to4_newdementia_or_lastinter Cwv2to4_dementia_free_date C_time_dementia_months ///
 C_time_dementia_midpoint C_time_dementia_midpoint_final C_time_of_event_dementia




/* 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 1 to 4(t0-t3)
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=84)


*rename cesd score across the waves - discrete var min=0 max=8

rename Cwv1_cesd_score cesd_0
rename Cwv2_cesd_score cesd_1
rename Cwv3_cesd_score cesd_2
rename Cwv4_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 Cwv1_depressive_symptoms
summarize Cwv1_depressive_symptoms
```

```

misstable summarize Cwv1_depressive_symptoms
misstable patterns Cwv1_depressive_symptoms

tabulate Cwv1_dementia_report
summarize Cwv1_dementia_report

misstable summarize Cwv1_dementia_report
misstable patterns Cwv1_dementia_report


tabulate Cwv2_dementia_report
summarize Cwv2_dementia_report

misstable summarize Cwv2_dementia_report
misstable patterns Cwv2_dementia_report


tabulate Cwv3_dementia_report
summarize Cwv3_dementia_report

misstable summarize Cwv3_dementia_report
misstable patterns Cwv3_dementia_report

tabulate Cwv4_self_info_dementia
summarize Cwv4_self_info_dementia

misstable summarize Cwv4_self_info_dementia
misstable patterns Cwv4_self_info_dementia








*** CLEANING DATA

* 1. drop missing data depression and dementia at baseline
* drop 663 depression missing data

drop if cesd_0== .
* (625 observations deleted)
drop if Cwv1_dementia_report== .
* (38 observations deleted)

* 2. drop dementia cases between wv1 and wv3 (total: 407 cases)

drop if Cwv1_dementia_report==1
* (234 observations deleted)

drop if Cwv2_dementia_report==1
* (73 observations deleted)

drop if Cwv3_dementia_report==1
* (100 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-mi
```

```
ssing-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 missing data)
drop if nmisfollowup>1
*(1549 observations deleted)



* 4. drop obs with no records on dementia at wave 4

drop if Cwv4_self_info_dementia== .
*(3823 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(C_sex)stats (mean v n)
tabstat cesd_1, by(C_sex)stats (mean v n)
tabstat cesd_2, by(C_sex)stats (mean v n)
```

```
tabstat cesd_3, by(C_sex)stats (mean v n)

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


bysort C_sex: tab cesd_0
bysort C_sex: tab cesd_1
bysort C_sex: tab cesd_2
bysort C_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
longitudinal nalysis of trajectories
GBTM model

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 10 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 censored normal distribution (‘cnorm’) is applied since the depressive symptom scores were negatively skewed.


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 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

*/


*** CNORM MODEL


* 1 class - cnorm model - quadratic polynomial (2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2)
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


/*
Trajectory shape
2 - p-value sig 0.0000
*/




* 2 classes - cnorm model - quadratic polynomial (2 2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2)
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 - cnorm model - quadratic polynomial (2 2 2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2 2)
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 - cnorm model - quadratic polynomial (2 2 2 2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2 2 2)
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 - cnorm model - quadratic polynomial (2 2 2 2 2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2 2 2 2)
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 - cnorm model - quadratic  anc cubic polynomial (2 2 2 2 2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2 3 3 3)
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 - cnorm model - quadratic  anc cubic polynomial (2 2 2 2 2 2)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2 2 2 2 2)
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-model depressive traj is selected to be tested in different shapes.




* 5 classes - cnorm model - cubic polynomial (3 3 3 3 3)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(3 3 3 3 3)
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 MODEL


* 5 classes - cnorm model - cubic polynomial (2 2 3 3 3)
traj, var(cesd_*) indep(t_*) model(cnorm) min(0) max(30) order(2 2 3 3 3)
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 (2 2 3 3 3)
*/




/*
Data and variable manipulation
*/

* 5-class model: rename  _traj_Group to C_traj_group5

rename  _traj_Group C_traj_group5
recode C_traj_group5 (3=4) (4=3)
ta C_traj_group5
rename _traj_ProbG1 C_depres_traj_1
rename _traj_ProbG2 C_depres_traj_2
rename _traj_ProbG3 C_depres_traj_3
rename _traj_ProbG4 C_depres_traj_4
rename _traj_ProbG5 C_depres_traj_5



* labelling variable of C_traj_group4

label var C_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 C_traj_group5 traj_depres

ta C_traj_group5
```

```
* Frequencies of covariates

tabulate C_age
summarize C_age


tabulate C_sex
summarize C_sex


tabulate C_eduaction
summarize C_eduaction


tabulate C_maritalstatus_4cat
summarize C_maritalstatus_4cat


tabulate Cwv1_netwealth_quintiles
summarize Cwv1_netwealth_quintiles


tabulate Cwv1_smoking_3cat
summarize Cwv1_smoking_3cat


tabulate Cwv1_physicalactivity
summarize Cwv1_physicalactivity



tabulate Cwv1_alcohol_status
summarize Cwv1_alcohol_status


tabulate Cwv1_cvd_comorbidity
summarize Cwv1_cvd_comorbidity


tabulate Cwv1_memory_wordrecall
summarize Cwv1_memory_wordrecall










/* MULTIPLE IMPUTATION (MI)

To handle with missing values of baseline and time 3 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 C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv4_smoking_3cat Cwv4_physicalactivity Cwv4_alcohol_status ///
Cwv4_cvd_comorbidity Cwv3_glycemia Cwv3_bp Cwv3_obesity Cwv3_hdl_cholesterol Cwv3_crp ///
Cwv4_memory_wordrecall


* examining missing data patterns

mi set wide

mi misstable summarize C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv4_smoking_3cat Cwv4_physicalactivity Cwv4_alcohol_freq Cwv4_alcohol_status ///
Cwv4_cvd_comorbidity Cwv3_glycemia Cwv3_bp Cwv3_obesity Cwv3_hdl_cholesterol Cwv3_crp ///
Cwv4_memory_wordrecall  Cwv1_antidepressant Cwv1_psycholog_treat Cwv1_anytreat_psyche
```

```
mi misstable patterns C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv4_smoking_3cat Cwv4_physicalactivity Cwv4_alcohol_freq Cwv4_alcohol_status ///
Cwv4_cvd_comorbidity Cwv3_glycemia Cwv3_bp Cwv3_obesity Cwv3_hdl_cholesterol Cwv3_crp ///
Cwv4_memory_wordrecall  Cwv1_antidepressant Cwv1_psycholog_treat Cwv1_anytreat_psyche



/*
 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:

    C_age Cwv1_netwealth_quintiles
    Cwv4_smoking_3cat Cwv4_alcohol_status
    Cwv4_cvd_comorbidity Cwv3_glycemia Cwv3_bp Cwv3_obesity Cwv3_hdl_cholesterol Cwv3_crp
    Cwv4_memory_wordrecall



Potential auxiliary var:
DV:  Cwv4_self_info_dementia
IV:  C_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3
other var: C_sex C_eduaction C_maritalstatus_4cat

*/


* correlation

pwcorr C_age Cwv1_netwealth_quintiles ///
    Cwv4_smoking_3cat Cwv4_alcohol_status ///
    Cwv4_cvd_comorbidity Cwv3_glycemia Cwv3_bp Cwv3_obesity Cwv3_hdl_cholesterol Cwv3_crp ///
    Cwv4_memory_wordrecall ///
    Cwv4_self_info_dementia C_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3 ///
    C_sex C_eduaction C_maritalstatus_4cat, 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) ->
    C_age Cwv1_memory_wordrecall

    logistic for the binary var (logit) ->
    Cwv1_cvd_comorbidity

    multinomial logistic for our nominal categorical var (mlogit) ->
    Cwv1_netwealth_quintiles
    Cwv1_smoking_3cat Cwv1_alcohol_status



- auxiliary var:

    DV -> Cwv4_self_info_dementia
    IV -> C_traj_group4
    other covariates -> C_sex C_eduaction C_maritalstatus_4cat
```

```
- imputation numbers (m) -> 20

    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 C_age Cwv1_netwealth_quintiles ///
    Cwv1_smoking_3cat Cwv1_alcohol_status ///
    Cwv1_cvd_comorbidity  ///
    Cwv1_memory_wordrecall





mi impute chained (logit) Cwv1_cvd_comorbidity ///
(mlogit) Cwv1_netwealth_quintiles Cwv1_smoking_3cat Cwv1_alcohol_status ///
(regress) C_age Cwv1_memory_wordrecall = Cwv4_self_info_dementia C_traj_group4 ///
C_sex C_eduaction C_maritalstatus_4cat, 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 C_age_mean1, name(mice1,replace)legend(off) ytitle("Mean of age")
tsline Cwv1_netwealth_quintiles_mean1, name(mice1,replace)legend(off) ytitle("Mean of wealth")
```

```
tsline Cwv1_smoking_3cat_mean1, name(mice1,replace)legend(off) ytitle("Mean of smoking")
tsline Cwv1_alcohol_status_mean1, name(mice1,replace)legend(off) ytitle("Mean of alcohol")
tsline Cwv1_cvd_comorbidity_mean1, name(mice1,replace)legend(off) ytitle("Mean of cvd")
tsline Cwv1_memory_wordrecall_mean1, name(mice1,replace)legend(off) ytitle("Mean of memory")


/*

All 10 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 Cwv1_memory_wordrecall_mean*, name(mice1,replace)legend(off) ytitle("Mean of memory")
tsline Cwv1_memory_wordrecall_sd*, name(mice2, replace) legend(off) ytitle("SD of memory")
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 CHARLS participants

* Demographics
sum C_age
ta C_sex
ta C_eduaction
ta C_maritalstatus_4cat
ta Cwv1_netwealth_quintiles
* Lifestyle factors
ta Cwv1_smoking_3cat
ta Cwv1_alcohol_status
ta Cwv1_physicalactivity
* Cardiometabolic health
ta Cwv1_cvd_comorbidity
```

```
ta Cwv1_diabetes_report
ta Cwv1_HbA1c
ta Cwv1_crp
ta Cwv1_hdl_cholesterol
ta Cwv1_waist
ta Cwv1_systolic_bp
ta Cwv1_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 at baseline
sum Cwv1_memory_wordrecall






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


* Demographics
oneway C_age C_traj_group5, tabulate
ta C_sex C_traj_group5, chi2 column row
ta C_eduaction C_traj_group5, chi2 column row
ta C_maritalstatus_4cat C_traj_group5, chi2 column row
ta Cwv1_netwealth_quintiles C_traj_group5, chi2 column row
* Lifestyle factors
ta Cwv1_smoking_3cat C_traj_group5, chi2 column row
ta Cwv1_alcohol_status C_traj_group5, chi2 column row
ta Cwv1_physicalactivity C_traj_group5, chi2 column row
* Cardiometabolic health
ta Cwv1_cvd_comorbidity C_traj_group5, chi2 column row
ta Cwv1_diabetes_report C_traj_group5, chi2 column row
ta Cwv1_HbA1c C_traj_group5, chi2 column row
ta Cwv1_crp C_traj_group5, chi2 column row
ta Cwv1_hdl_cholesterol C_traj_group5, chi2 column row
ta Cwv1_waist C_traj_group5, chi2 column row
ta Cwv1_systolic_bp C_traj_group5, chi2 column row
ta Cwv1_diastolic_bp C_traj_group5, chi2 column row
* Memory score
oneway Cwv1_memory_wordrecall C_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 Cwv4_self_info_dementia C_traj_group5

*OR

logit Cwv4_self_info_dementia C_traj_group5

*OR

logit Cwv4_self_info_dementia C_traj_group5, or



* define design var by using i.(decreasing, increasing, high ref: low)

logistic Cwv4_self_info_dementia i.C_traj_group5

*OR


logit Cwv4_self_info_dementia i.C_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 Cwv4_self_info_dementia i.C_traj_group5 ///
C_age C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles

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

logistic Cwv4_self_info_dementia i.C_traj_group5 ///
C_age C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.Cwv1_cvd_comorbidity









/* -------- BINOMIAL LOGISTIC REGRESSION IN IMPUTED DATASET using time 3 covariates --------- *

Command is

mi estimate : logit DV IV, or

    OR

mi estimate: logistic DV IV

*/



* to redefine reference group to be the trajectory of "minimal symptoms" then use ib2.exposure_var -- the ib and the nnumber of var make the change in reference groups

```

```
* Unadjusted model - model 1 - single predictor

mi estimate, eform("Odds Ratio"): logistic Cwv4_self_info_dementia C_traj_group5

*OR

mi estimate, eform("Odds Ratio"): logit Cwv4_self_info_dementia C_traj_group5, or



* define design var by using i.(decreasing, increasing, high, ref: low)

mi estimate, eform("Odds Ratio"): logistic Cwv4_self_info_dementia i.C_traj_group5

*OR

mi estimate, eform("Odds Ratio"): logit Cwv4_self_info_dementia i.C_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 Cwv4_self_info_dementia i.C_traj_group5 ///
C_age C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles

* model 3: model 2 + adjust for lifestyle factors

mi estimate, eform("Odds Ratio"): logistic Cwv4_self_info_dementia i.C_traj_group5 ///
C_age C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.Cwv1_cvd_comorbidity







/*

*** SENSITIVITY ANALYSES ***


1) LCGA logit trajectories with dichotomous variable


2) Complete data

*/





/*
1) Logistic model LCGA

use Ewv2_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 Cwv1_depressive_symptoms score across the waves

rename Cwv1_depressive_symptoms depress_0
rename Cwv2_depressive_symptoms depress_1
rename Cwv3_depressive_symptoms depress_2
rename Cwv4_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)

/* 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)

/* 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) ytitle(Depressive symptom caseness) ci

/* Assigned group and probabilties of group membership */
    list _traj_Group - _traj_ProbG4 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 2 3 3)
traj, var(depress_*) indep(t_*) model(logit) order(3 2 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]
    list _traj_Group cesdG groupAPP occ occ_pp p TotProb if counter == 1
    restore
end

summary_table_procTraj



ta _traj_Group






* IMPUTED DATA: Logistic regression (Odds Ratio)

* Unadjusted model (model 1)

mi estimate, eform("Odds Ratio"): logistic Cwv4_self_info_dementia i._traj_Group

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

mi estimate, eform("Odds Ratio"): logistic Cwv4_self_info_dementia i._traj_Group ///
C_age C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles

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

mi estimate, eform("Odds Ratio"): logistic Cwv4_self_info_dementia i._traj_Group ///
C_age C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.Cwv1_cvd_comorbidity



* 2) complete data analysis (see above)



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