```
/*

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 4: Clustering of cardiometabolic risk factors and dementia incidence


Method of analysis:
Latent Class Analysis (LCA)

DATASET: CHARLS
baseline: wave 1 (2011) follow-up waves 2-4 (2013-2018)


TIMELINE

LATENT CLASSES OF CARDIOMETABOLIC MARKERS: WV1 (BASELINE)
DEMENTIA INCIDENCE: W2 - WV4 (3 TIME POINTS)
COVARIATES ADJUSTMENT FOR HR MODELS: WV1

*/




* KEEP NECESSARY VARIABLES

keep ID id_12char bloodweight ///
C_sex C_age C_eduaction C_educ_new C_maritalstatus_8cat C_maritalstatus_3cat C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
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_cognition ///
Cwv1_cesd_score Cwv1_depressive_symptoms ///
Cwv2_cesd_score Cwv2_depressive_symptoms ///
Cwv3_cesd_score Cwv3_depressive_symptoms ///
Cwv4_cesd_sumscore Cwv4_depressive_symptoms ///
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_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 ///
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 ///
Cwv2to4_dementia_sum Cwv2to4_dementia_event ///
C_time_dementia_midpoint C_time_dementia_midpoint_final C_time_of_event_dementia









/* Latent class analysis - LCA  of cardiomeatbolic risk factors for dementia

Useful links:
```

```
https://www.stata.com/meeting/uk18/slides/uk18_MacDonald.pdf

https://www.stata.com/meeting/mexico18/slides/5_Mexico18_Canette.pdf

https://www.bgsu.edu/content/dam/BGSU/college-of-arts-and-sciences/center-for-family-and-demographic-research/documents/Workshops/2020-latent-class-analysis.pdf

https://www.stata.com/features/overview/latent-class-analysis/

https://www.stata.com/manuals/semexample50g.pdf

https://www.stata.com/manuals/semexample51g.pdf

https://www.stata.com/manuals/semexample52g.pdf

https://www.ucl.ac.uk/population-health-sciences/sites/population_health_sciences/files/lca.pdf

https://www.stata.com/manuals/semgsemlclassoptions.pdf

https://www.stata.com/meeting/nordic-and-baltic17/slides/nordic-and-baltic17_Pitblado.pdf

https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00920/full

https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2774074

https://www.statalist.org/forums/forum/general-stata-discussion/general/1412686-calculating-entropy-for-lca-latent-class-analysis-in-stata-15

https://www.statalist.org/forums/forum/general-stata-discussion/general/1590174-how-to-calculate-entropy-for-lca-with-stata

https://www.statalist.org/forums/forum/general-stata-discussion/general/1390895-combine-marginsplot-problem-with-plot-options




* gsem command to fit a latent class model

* gsem (var1 var2 var3 <-), logit lclass(C 3)



OR TRY

gsem (var1 var2 var3 <-), logit lclass(C 3) ///
startvalues(randompr, draws(20) seed(15) difficult) ///
emopts(iterate(30) difficult)



Binary variables of cardiometabolic markers measured at wave 2

CRP: Cwv1_crp

HDL cholesterol: Cwv1_hdl_cholesterol

Obesity by waist cir: Cwv1_obesity_waist

systolic Blood pressure: Cwv1_systolic_bp

diastolic Blood pressure: Cwv1_diastolic_bp

Diabetes: Cwv1_diabetes_report

HbA1c: Cwv1_HbA1c
```

```
*/






*** Descriptive stats of cardiometabolic markers



tabulate Cwv1_crp
summarize Cwv1_crp

misstable summarize Cwv1_crp
misstable patterns Cwv1_crp

tabulate Cwv1_hdl_cholesterol
summarize Cwv1_hdl_cholesterol

misstable summarize Cwv1_hdl_cholesterol
misstable patterns Cwv1_hdl_cholesterol

tabulate Cwv1_obesity_waist
summarize Cwv1_obesity_waist

misstable summarize Cwv1_obesity_waist
misstable patterns Cwv1_obesity_waist

tabulate Cwv1_systolic_bp
summarize Cwv1_systolic_bp

misstable summarize Cwv1_systolic_bp
misstable patterns Cwv1_systolic_bp


tabulate Cwv1_diastolic_bp
summarize Cwv1_diastolic_bp

misstable summarize Cwv1_diastolic_bp
misstable patterns Cwv1_diastolic_bp


tabulate Cwv1_diabetes_report
summarize Cwv1_diabetes_report

misstable summarize Cwv1_diabetes_report
misstable patterns Cwv1_diabetes_report


tabulate Cwv1_HbA1c
summarize Cwv1_HbA1c

misstable summarize Cwv1_HbA1c
misstable patterns Cwv1_HbA1c


tabulate Cwv1_dementia_report
summarize Cwv1_dementia_report

misstable summarize Cwv1_dementia_report
misstable patterns Cwv1_dementia_report
```

```


*** CLEANING DATA


* 1. drop dementia cases at baseline

* drop dementia wave 2 missing data
drop if Cwv1_dementia_report==1
* (267 observations deleted)
drop if Cwv1_dementia_report== .
* (88 observations deleted)


* 2. drop missing values of cardiometabolic markers

drop if Cwv1_crp== .
* (180 observations deleted)

drop if Cwv1_hdl_cholesterol== .
* (2 observations deleted)

drop if Cwv1_obesity_waist== .
* (1688 observations deleted)

drop if Cwv1_systolic_bp== .
* (83 observations deleted)

drop if Cwv1_diastolic_bp== .
* (13 observations deleted)

drop if Cwv1_diabetes_report== .
* (91 observations deleted)

drop if Cwv1_HbA1c== .
* (71 observations deleted)



* 3. drop obs with no records on dementia at any wave from 2-4 follow-ups


search mdesc
search rmiss2
search mvpatterns

* see number of missing values vs non-missing in each variable
mdesc Cwv2_dementia_report Cwv3_dementia_report Cwv4_self_info_dementia



/* 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_wv2to4=rmiss2(Cwv2_dementia_report ///
Cwv3_dementia_report Cwv4_self_info_dementia)

tab nmisfollowup_dementia_wv2to4

* drop observations "nmisfollowup_dementia_wv2to4" > 2 (those with 3 missing data = no records at
any wave)
drop if nmisfollowup_dementia_wv2to4>2
*(342 observations deleted)


* FINAL SAMPLE -> 9022


```

```
/* Latent Class analysis - gsem

* 7 variables: Cwv1_crp Cwv1_hdl_cholesterol Cwv1_obesity_waist
 Cwv1_systolic_bp Cwv1_diastolic_bp Cwv1_diabetes_report Cwv1_HbA1c

*/


* change names to start with lowercase (STATA assumes variables starting with a capital letter are cont latent variables!)

rename Cwv1_crp crp_lca
rename Cwv1_hdl_cholesterol hdl_lca
rename Cwv1_obesity_waist obesity_lca
rename Cwv1_systolic_bp systolic_lca
rename Cwv1_diastolic_bp diastolic_lca
rename Cwv1_diabetes_report diabetes_lca
rename Cwv1_HbA1c hba1c_lca



* Corrrelation matrix of the CM variables

corr crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca hba1c_lca

pwcorr crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca hba1c_lca, sig


* to create quality table in word - asdoc -
* https://www.youtube.com/watch?v=XHBl6PHfOzs&ab_channel=StataProfessor

help asdoc

asdoc pwcorr crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca hba1c_lca, sig


asdoc pwcorr crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca hba1c_lca, nonum replace cor crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca hba1c_lca label replace star(.05) dec(2)



* LCA models


* one-class model

gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 1)

estimates store oneclass_cm

* two-class model

gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 2)


estimates store twoclass_cm


* three-class model

gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
```

```
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 3)

gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 3) ///
startvalues(randompr, draws(20) seed(15) difficult) ///
emopts(iterate(30) difficult)

estimates store threeclass_cm


* four-class model

gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 4)


gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 4) ///
startvalues(randompr, draws(20) seed(15) difficult) ///
emopts(iterate(30) difficult)


estimates store fourclass_cm



/*

** Evaluating Fit to choose the number of classes **

1. a priori theory
2. Information Statistics
        AIC, BIC, adjusted BIC
3. Chi-Square goodness of fit
4. Entropy

Others but not used here:
Lo-Mendell-Rubin (LMR)
        Not recommended (designed for normal Y)
Bootstrapped Likelihood Ratio Test

*/

* AIC and BIC to determine which of these models fits best

estimates stats oneclass_cm twoclass_cm threeclass_cm



* LCA postestimation
* Latent class marginal probabilities - lcprob -
* Latent class marginal means - lcmean -


estat lcprob

estat lcmean



* likelihood -ratio test (G2) to evaluate whether our model fits as well as the saturated model


estat lcgof

* if p value is sig means that we reject the null hypothesis and and the model doesn't fit well
* if p value is not sig means that we fail to reject the null hypothesis that our model fits as well as the saturated model.
```

```
* for a well-fitted model p value should be non-sig. A significant p-value indicates lack of model fit in absolute terms.

* 3-class model: p < 0.001


* Entropy

quietly predict classpost*, classposteriorpr
gen sum_p_lnp = 0
forvalues k = 1/2 {
        replace sum_p_lnp = sum_p_lnp + classpost`k'*ln(classpost`k')
}
summ sum_p_lnp, meanonly
scalar E = 1+`r(sum)'/(e(N)*ln(2))
drop classpost? sum_p_lnp
di E







/* We can use the predictions of the posterior probability of class membership to evaluate an
individual’s probability of being in each class.

*/

predict classpost1*, classposteriorpr
list in 1, abbrev(10)

/* We can determine the expected class for each individual based on whether the posterior probability
is greater than 0.5
*/

generate expclass1 = 1 + (classpost11>0.5)
tabulate expclass


generate expclass2 = 1 + (classpost12>0.5)
tabulate expclass2


generate expclass3 = 1 + (classpost13>0.5)
tabulate expclass3



/* We can determine expected classification for each individual in the dataset based on the predicted
posterior class probabilities.
*/

predict cpost*, classposteriorpr
egen max = rowmax(cpost*)


* generate classes var

generate predclass = 1 if cpost1==max

replace predclass = 2 if cpost2==max

replace predclass = 3 if cpost3==max

tabulate predclass
```

```
**********function to print out summary stats
program summary_table_procLCla
    preserve
    *look at the average posterior probability
    gen Mp = 0
    foreach i of varlist cpost* {
        replace Mp = `i' if `i' > Mp
    }
    sort predclass
    *and the odds of correct classification
    by predclass: gen countG = _N
    by predclass: egen groupAPP = mean(Mp)
    by predclass: 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 cpost* {
       scalar c = c + 1
       quietly summarize `i'
       replace TotProb = r(sum)/ _N if predclass == 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 predclass countG groupAPP occ occ_pp p TotProb if counter == 1
    restore
end

summary_table_procLCla




/*
Class variable manipulation
*/

* 3-class model: rename  predclass to C_lca_group3

rename  predclass C_lca_group3

* labelling variable of C_lca_group3

label var C_lca_group3 "Latent classes 3 groups of cardiometabolic markers"

* labelling values
lab def lca_group3 1 "Relatively healthy" 2 "Obesity and Hypertension" 3 "Complex cardiometabolic disorders"

* attach category labels to the variable through label value

lab val C_lca_group3 lca_group3

ta C_lca_group3
```

```
* margins and marginsplot

* use margins to calculate marginal predictions
* use marginsplot to graph marginal predictions


margins, predict(classpr class(1)) ///
         predict(classpr class(2)) ///
         predict(classpr class(3))

marginsplot, xtitle ("") ytitle ("") ///
             xlabel (1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title ("Predicted Latent Class Probabilities with 95% CI")


margins, predict(classpr class(1)) ///
         predict(classpr class(2)) ///
         predict(classpr class(3))
marginsplot, recast(bar) xtitle("") ytitle("") ///
             xlabel(1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title("Predicted Latent Class Probabilities with 95% CI")

margins, predict(outcome(hba1c_lca) class(1)) ///
         predict(outcome(hba1c_lca) class(2)) ///
         predict(outcome(hba1c_lca) class(3))
marginsplot, recast(bar) xtitle("") ytitle("") ///
             xlabel(1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title("Predicted Pr(HbA1c=1) with 95% CI")

* repeat with all CM var



* class 1

margins, predict(outcome(crp_lca) class(1)) ///
         predict(outcome(hdl_lca) class(1)) ///
         predict(outcome(obesity_lca) class(1)) ///
         predict(outcome(systolic_lca) class(1)) ///
         predict(outcome(diastolic_lca) class(1)) ///
         predict(outcome(diabetes_lca) class(1)) ///
         predict(outcome(hba1c_lca) class(1)) ///



marginsplot, recast(bar) title ("Class 1") xtitle("") ///
             xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
             5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
             ytitle("Predicted mean") ylabel(0(.20)1) name (class1)


* class 2

margins, predict(outcome(crp_lca) class(2)) ///
         predict(outcome(hdl_lca) class(2)) ///
         predict(outcome(obesity_lca) class(2)) ///
         predict(outcome(systolic_lca) class(2)) ///
         predict(outcome(diastolic_lca) class(2)) ///
         predict(outcome(diabetes_lca) class(2)) ///
         predict(outcome(hba1c_lca) class(2)) ///



marginsplot, recast(bar) title ("Class 2") xtitle("") ///
             xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
             5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
```

```
                ytitle("Predicted mean") ylabel(0(.20)1) name (class2)




* class 3

margins, predict(outcome(crp_lca) class(3)) ///
         predict(outcome(hdl_lca) class(3)) ///
         predict(outcome(obesity_lca) class(3)) ///
         predict(outcome(systolic_lca) class(3)) ///
         predict(outcome(diastolic_lca) class(3)) ///
         predict(outcome(diabetes_lca) class(3)) ///
         predict(outcome(hba1c_lca) class(3)) ///



marginsplot, recast(bar) title ("Class 3") xtitle("") ///
             xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
             5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
             ytitle("Predicted mean") ylabel(0(.20)1) name (class3)



graph combine class1 class2 class3, cols(3)






* ------------------------------- *




/* 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 ///
Cwv1_smoking_3cat Cwv1_physicalactivity Cwv1_alcohol_status ///
C_cvd_comorbidity Cwv1_memory_wordrecall Cwv1_depressive_symptoms



* examining missing data patterns

mi set wide

mi misstable summarize C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_physicalactivity Cwv1_alcohol_status ///
C_cvd_comorbidity ///
Cwv1_memory_wordrecall Cwv1_depressive_symptoms



mi misstable patterns C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_physicalactivity Cwv1_alcohol_status ///
C_cvd_comorbidity ///
Cwv1_memory_wordrecall Cwv1_depressive_symptoms


/*
 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:

    Cwv1_netwealth_quintiles
    Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity
    Cwv1_memory_wordrecall Cwv1_depressive_symptoms

```

```

Potential auxiliary var:
DV:  Cwv2to4_dementia_event
IV:  crp_lca hdl_lca obesity_lca Cwv1_tg systolic_lca diastolic_lca hba1c_lca diabetes_lca
other var: C_age C_sex C_eduaction C_maritalstatus_4cat

*/


* correlation

pwcorr Cwv1_netwealth_quintiles ///
    Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
    Cwv1_memory_wordrecall Cwv1_depressive_symptoms ///
    Cwv2to4_dementia_event crp_lca hdl_lca obesity_lca Cwv1_tg systolic_lca diastolic_lca
hba1c_lca diabetes_lca ///
    C_age C_sex C_eduaction C_maritalstatus_4cat, obs


/* The correlation showed that all the following var are good auxiliary:
Cwv2to4_dementia_event obesity_lca diabetes_lca C_age C_sex C_eduaction C_maritalstatus_4cat
* 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) ->
    C_cvd_comorbidity Cwv1_depressive_symptoms

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



- auxiliary var:

    DV -> Cwv2to4_dementia_event
    IV -> obesity_lca diabetes_lca
    other covariates -> C_age 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 Cwv1_netwealth_quintiles ///
    Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
    Cwv1_memory_wordrecall Cwv1_depressive_symptoms



mi impute chained (logit) C_cvd_comorbidity Cwv1_depressive_symptoms ///
(mlogit) Cwv1_netwealth_quintiles Cwv1_smoking_3cat Cwv1_alcohol_status ///
(regress) Cwv1_memory_wordrecall = Cwv2to4_dementia_event obesity_lca diabetes_lca ///
C_age 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 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 C_cvd_comorbidity_mean1, name(mice1,replace)legend(off) ytitle("Mean of cvd")
tsline Cwv1_depressive_symptoms_mean1, name(mice1,replace)legend(off) ytitle("Mean of depression")
tsline Cwv1_memory_wordrecall_mean1, name(mice1,replace)legend(off) ytitle("Mean of memory")


/*

All 20 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 (baseline and time 3 covariates) ----

General characteristics of participants

General characteristics of participipnats stratified for study inclusion

General characteristics of participants stratified for dementia occurence

Participant characteristics by CM 3-class groups

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 at baseline

* Socio-demographics
sum C_age
ta C_sex
ta C_educ_new
ta C_maritalstatus_4cat
ta Cwv1_netwealth_quintiles
* Cardiometabolic disorders
ta crp_lca
ta hdl_lca
ta obesity_lca
ta systolic_lca
ta diastolic_lca
ta diabetes_lca
ta hba1c_lca
* Cardiovascular health factors
ta Cwv1_smoking_3cat
ta Cwv1_alcohol_status
ta Cwv1_physicalactivity
ta C_cvd_comorbidity
* Depressive symptoms (cont and categ)
sum Cwv1_cesd_score
ta Cwv1_depressive_symptoms
* Memory score
sum Cwv1_memory_wordrecall



* General baseline characteristics of CHARLS participants by dementia status

* Socio-demographics
ttest C_age, by(Cwv2to4_dementia_event)
ta C_sex Cwv2to4_dementia_event, chi2 column row
ta C_educ_new Cwv2to4_dementia_event, chi2 column row
ta C_maritalstatus_4cat Cwv2to4_dementia_event, chi2 column row
ta Cwv1_netwealth_quintiles Cwv2to4_dementia_event, chi2 column row
* Cardiometabolic disorders
ta crp_lca Cwv2to4_dementia_event, chi2 column row
ta hdl_lca Cwv2to4_dementia_event, chi2 column row
ta obesity_lca Cwv2to4_dementia_event, chi2 column row
```

```
ta systolic_lca Cwv2to4_dementia_event, chi2 column row
ta diastolic_lca Cwv2to4_dementia_event, chi2 column row
ta diabetes_lca Cwv2to4_dementia_event, chi2 column row
ta hba1c_lca Cwv2to4_dementia_event, chi2 column row
* Cardiovascular health factors
ta Cwv1_smoking_3cat Cwv2to4_dementia_event, chi2 column row
ta Cwv1_alcohol_status Cwv2to4_dementia_event, chi2 column row
ta Cwv1_physicalactivity Cwv2to4_dementia_event, chi2 column row
ta C_cvd_comorbidity Cwv2to4_dementia_event, chi2 column row
* Depressive symptoms (cont and categ)
ttest Cwv1_cesd_score, by(Cwv2to4_dementia_event)
ta Cwv1_depressive_symptoms Cwv2to4_dementia_event, chi2 column row
* Memory score
ttest Cwv1_memory_wordrecall, by(Cwv2to4_dementia_event)




* Sample characteristics by CM 3-class groups
* crosstabs categ var (frequencies and chi2) !report column percentage!
* oneway ANOVA cont var (mean, sd)


* Socio-demographics
oneway C_age C_lca_group3, tabulate
ta C_sex C_lca_group3, chi2 column row
ta C_educ_new C_lca_group3, chi2 column row
ta C_maritalstatus_4cat C_lca_group3, chi2 column row
ta Cwv1_netwealth_quintiles C_lca_group3, chi2 column row
* Cardiovascular health factors
ta Cwv1_smoking_3cat C_lca_group3, chi2 column row
ta Cwv1_alcohol_status C_lca_group3, chi2 column row
ta Cwv1_physicalactivity C_lca_group3, chi2 column row
ta C_cvd_comorbidity C_lca_group3, chi2 column row
* Depressive symptoms (cont and categ)
oneway Cwv1_cesd_score C_lca_group3, tabulate
ta Cwv1_depressive_symptoms C_lca_group3, chi2 column row
* Memory score
oneway Cwv1_memory_wordrecall C_lca_group3, tabulate








/*
---- SURVIVAL ANALYSIS AT COMPLETE DATA ----

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 C_time_of_event_dementia Cwv2to4_dementia_event C_lca_group3 ///
C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_physicalactivity Cwv1_alcohol_status C_cvd_comorbidity ///
Cwv1_depressive_symptoms Cwv1_memory_wordrecall,compact




* Declare Data to be Survival Data
* Time to event: C_time_of_event_dementia (months)
* Censoring: Cwv2to4_dementia_event (1=dementia, 0=censored)
* Command is stset TIMETOEVENT, failure(CENSORVARIABLE)


stset C_time_of_event_dementia, failure (Cwv2to4_dementia_event==1) id(id_12char)



*describe survival data using commnad stsum

stsum

stsum, by(C_lca_group3)




* Kaplan Meier Curve estimation

sts list

sts list, by(C_lca_group3)



* 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=9022, # events=470") ///
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=9022, # events=470") 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 C_lca_group3

sts list, by(C_lca_group3)

* graph with frills

sts graph, by(C_lca_group3) xlabel(0(20)100) ylabel(0.80(.05)1) xtitle("Time in Months") ///
ytitle("Survival Prob") title("Kaplan Meier Curve") subtitle("n=9022, # events=470") ///
caption("graph04.png", size(vsmall))
```

```
* calculate person-time and incidence rates using command ststime

stptime,title(Person-years)

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

stptime, title(Person-years) per(10000)


* calculate person-time  by category of C_lca_group3

stptime, by(C_lca_group3)

stptime, by(C_lca_group3) per(1000)



* mean and median of follow-up
sum C_time_of_event_dementia
sum C_time_of_event_dementia, detail



/* Log Rank Test of equality of survival distributions
 (NULL: equality of survival distributions among C_lca_group3 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 C_lca_group3, logrank

sts test C_age, logrank

sts test C_sex, logrank

sts test C_eduaction, logrank

sts test C_maritalstatus_4cat, logrank

sts test Cwv1_netwealth_quintiles, logrank

sts test Cwv1_smoking_3cat, logrank

sts test Cwv1_physicalactivity, logrank

sts test Cwv1_alcohol_status, logrank

sts test C_cvd_comorbidity, logrank

sts test Cwv1_depressive_symptoms, logrank

sts test Cwv1_memory_wordrecall, logrank
```

```
/* Cox PH regression model

using the command stcox

--- Building the model ---

Model 1: unadjusted - single predictor of CM classes
Model 2: model 1 + sociodemographics: age sex education marital status and wealth
Model 3: model 2 + cvd health: smoking, alcohol consumption, cvd comorbidity
Model 4: model 3 + mental health: depressive symptoms
*/


* Unadjusted model - model 1 - single predictor

stcox C_lca_group3

* define design var by using i.(3 classes)

stcox i.C_lca_group3


* Adjusted models - multivariable Cox model
* controlling for covariates

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

stcox i.C_lca_group3 C_age C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles

* model 3: model 2 + adjust for cvd health

stcox i.C_lca_group3 C_age C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity


* model 4:  model 3 + adjust for depression

stcox i.C_lca_group3 C_age C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms





* Coefficients instead of hazard ratios by specifing the option nohr

stcox i.C_lca_group3, nohr


stcox i.C_lca_group3 C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms, nohr




* Multivariable model development
* Likelihood-ratio tests



*install eststo
findit eststo

```

```
* ---- rx controlling for age and sex -----*
quietly: stcox C_age i.C_sex
eststo modelagesex

quietly: stcox C_age i.C_sex i.C_lca_group3
eststo modelagesex_3group

lrtest modelagesex modelagesex_3group



* ---- rx controlling for sociodemographics -----*
quietly: stcox C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles
eststo modelsociodemo

quietly: stcox C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles i.C_lca_group3
eststo modelsociodemo_3group

lrtest modelsociodemo modelsociodemo_3group


* ---- rx controlling for cardiovascular health -----*
quietly: stcox i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity
eststo modelcardiovascular

quietly: stcox i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity i.C_lca_group3
eststo modelcardiovascular_3group

lrtest modelcardiovascular modelcardiovascular_3group


* ---- rx controlling for mental health-----*
quietly: stcox i.Cwv1_depressive_symptoms Cwv1_memory_wordrecall
eststo modelmentalcogn

quietly: stcox i.Cwv1_depressive_symptoms i.C_lca_group3
eststo modelmentalcogn_3group

lrtest modelmentalcogn modelmentalcogn_3group




* side-by-side comparison of models


quietly: stcox i.C_lca_group3
eststo model1

quietly: stcox  C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles i.C_lca_group3
eststo model2

quietly: stcox C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity i.C_lca_group3
eststo model3

quietly: stcox  C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms i.C_lca_group3
eststo model4



* Display Betas and Summary Statistics
estout model1 model2 model3 model4, 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 model4, eform stats(n chi2 bic, star(chi2)) prehead("Hazard Ratios")




* Postestimation tools for stcox

* Test of proportional hazards

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.C_lca_group3 C_age i.C_sex i.C_eduaction i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms, nohr ///
tvc(C_lca_group3 C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
Cwv1_depressive_symptoms) texp(ln(C_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 C_lca_group3 C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
Cwv1_depressive_symptoms Cwv1_memory_wordrecall, schoenfeld(sch*) scaledsch(sca*)
stphtest, detail
stphtest, plot(C_lca_group3) msym(oh)
stphtest, plot(C_age) msym(oh)
stphtest, plot(C_sex) msym(oh)
stphtest, plot(C_eduaction) msym(oh)
stphtest, plot(C_maritalstatus_4cat) msym(oh)
stphtest, plot(Cwv1_netwealth_quintiles) msym(oh)
```

```
stphtest, plot(C_cvd_comorbidity) msym(oh)
stphtest, plot(Cwv1_smoking_3cat) msym(oh)
stphtest, plot(Cwv1_alcohol_status) msym(oh)
stphtest, plot(Cwv1_depressive_symptoms) msym(oh)





stphplot, by(C_lca_group3) plot1(msym(oh)) plot2(msym(th))
stphplot, by(C_age) plot1(msym(oh)) plot2(msym(th))
stphplot, by(C_sex) plot1(msym(oh)) plot2(msym(th))
stphplot, by(C_eduaction) plot1(msym(oh)) plot2(msym(th))
stphplot, by(C_maritalstatus_4cat) plot1(msym(oh)) plot2(msym(th))
stphplot, by(Cwv1_netwealth_quintiles) plot1(msym(oh)) plot2(msym(th))
stphplot, by(C_cvd_comorbidity) plot1(msym(oh)) plot2(msym(th))
stphplot, by(Cwv1_smoking_3cat) plot1(msym(oh)) plot2(msym(th))
stphplot, by(Cwv1_alcohol_status) plot1(msym(oh)) plot2(msym(th))
stphplot, by(Cwv1_depressive_symptoms) plot1(msym(oh)) plot2(msym(th))



* Assessment of PH Assumption: adjust for age and sex
stphplot, by(C_lca_group3) adjust(C_age C_sex) nolntime plot1opts(symbol(none) color(black)
lpattern(dash)) ///
plot2opts(symbol(none) color(green)) plot3opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is C_lca_group3") xtitle("months")



* Assessment of PH Assumption: adjust for model 2
stphplot, by(C_lca_group3) adjust(C_age C_sex C_eduaction C_maritalstatus_4cat
Cwv1_netwealth_quintiles) ///
nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
plot2opts(symbol(none) color(green)) plot3opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is C_lca_group3") xtitle("months")



* Assessment of PH Assumption: adjust for model 3
stphplot, by(C_lca_group3) adjust(C_age C_sex C_eduaction C_maritalstatus_4cat
Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity) ///
nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
plot2opts(symbol(none) color(green)) plot3opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is C_lca_group3") xtitle("months")



* Assessment of PH Assumption: adjust for model 4
stphplot, by(C_lca_group3) adjust(C_age C_sex C_eduaction C_maritalstatus_4cat
Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
Cwv1_depressive_symptoms) ///
nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
plot2opts(symbol(none) color(green)) plot3opts(symbol(none) color(red)) ///
title("Assessment of PH Assumption") subtitle(" Predictor is C_lca_group3") 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 C_lca_group3 C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles ///
Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
Cwv1_depressive_symptoms, mgale(mgale)


stcoxgof

* Good. Do not reject. We do not have statistically significant evidence of a poor fit (p-value =
.38).




* by using the Cox-Snell residuals

quietly stcox C_lca_group3 C_age C_sex C_eduaction C_maritalstatus_4cat Cwv1_netwealth_quintiles
///
Cwv1_smoking_3cat Cwv1_alcohol_status C_cvd_comorbidity ///
Cwv1_depressive_symptoms
predict cs, csnell

* or

quietly stcox E_traj_group4
predict cs, csnell


stset cs, failure(Cwv2to4_dementia_event)
sts generate km = s
generate H = -ln(km)
line H cs cs, sort ytitle("") clstyle(. refline)





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


* Declare Data to be Survival Data by using mi

mi stset C_time_of_event_dementia, failure (Cwv2to4_dementia_event==1) id(id_12char)


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



* Unadjusted model - model 1 - single predictor

* Model 1 (default coefficents)
mi estimate: stcox C_lca_group3

```

```
* Model 1: define design var by using i.
mi estimate: stcox i.C_lca_group3


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

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3


* Adjusted models - multivariable Cox model
* controlling for covariates

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

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
C_age i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles

* Model 3: model 2 + adjust for cvd health

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
C_age i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity


* Model 4: model 3 + adjust for depression

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
C_age i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms








/*


*** SENSITIVITY ANALYSES ***

1) multigroup latent class model by sex

2) interactions with age and gender
survival analysis stratified by age
two age groups: <70 and >=70

3) exclude participants with cvd

4) Complete data
 Cox regression analysis on complete data (without imputed covariates)
 (see above)
*/



/*
1) Multigroup latent class model by sex


TWO STEP PROCESS

1) LCA by group (to build the model and get lcprob and lcmean and to get the marginplots for males)

gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 3) ///
```

```
group(C_sex) ginvariant(coef)

estat lcprob
estat lcmean
estat lcgof

2) LCA sort sex (to get the marginplots for females)

sort C_sex

by C_sex: gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 3)


*/


* LCA by group
* three-class model


gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 3) ///
group(C_sex) ginvariant(coef) ///
startvalues(randompr, draws(20) seed(15) difficult) ///
emopts(iterate(30) difficult)

estimates store threeclass_cm



* LCA postestimation
* Latent class marginal probabilities - lcprob -
* Latent class marginal means - lcmean -


estat lcprob

estat lcmean

estat lcgof


/* We can use the predictions of the posterior probability of class membership to evaluate an
individual's probability of being in each class.

*/

predict m_classpost1*, classposteriorpr
list in 1, abbrev(10)

/* We can determine the expected class for each individual based on whether the posterior probability
is greater than 0.5
*/

generate m_expclass1 = 1 + (m_classpost11>0.5)
tabulate m_expclass1


generate m_expclass2 = 1 + (m_classpost12>0.5)
tabulate m_expclass2


generate m_expclass3 = 1 + (m_classpost13>0.5)
tabulate m_expclass3
```

```
/* We can determine expected classification for each individual in the dataset based on the predicted
posterior class probabilities.
*/

predict m_cpost*, classposteriorpr
egen m_max = rowmax(m_cpost*)


* generate classes var

generate m_predclass = 1 if m_cpost1==m_max

replace m_predclass = 2 if m_cpost2==m_max

replace m_predclass = 3 if m_cpost3==m_max

tabulate m_predclass



* margins and marginsplot for MALES

* use margins to calculate marginal predictions
* use marginsplot to graph marginal predictions


*Install/update combomarginsplot ado.

*https://www.statalist.org/forums/forum/general-stata-discussion/general/1425209-is-it-possible-to-do-multilevel-latent-class-analysis-with-stata-15-ic

ssc install combomarginsplot, replace



margins, predict(classpr class(1)) ///
         predict(classpr class(2)) ///
         predict(classpr class(3)) subpop(if C_sex==0) saving(margin_male, replace)
marginsplot, xtitle ("") ytitle ("") ///
             xlabel (1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title ("Predicted Latent Class Probabilities with 95% CI") ///
             name(margin_male, replace)


margins, predict(classpr class(1)) ///
         predict(classpr class(2)) ///
         predict(classpr class(3)) subpop(if C_sex==0) saving(margin_male, replace)
marginsplot, recast(bar) xtitle("") ytitle("") ///
             xlabel(1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title("Predicted Latent Class Probabilities with 95% CI") ///
             name(margin_male, replace)



* class 1

margins, predict(outcome(crp_lca) class(1)) ///
         predict(outcome(hdl_lca) class(1)) ///
         predict(outcome(obesity_lca) class(1)) ///
         predict(outcome(systolic_lca) class(1)) ///
         predict(outcome(diastolic_lca) class(1)) ///
         predict(outcome(diabetes_lca) class(1)) ///
         predict(outcome(hba1c_lca) class(1)) subpop(if C_sex==0) ///
         saving(class1_male, replace) ///

marginsplot, recast(bar) title ("Class 1") xtitle("") ///
```

```
                xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
                5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
                ytitle("Predicted mean") ylabel(0(.20)1) name (class1_male, replace)


* class 2

margins, predict(outcome(crp_lca) class(2)) ///
         predict(outcome(hdl_lca) class(2)) ///
         predict(outcome(obesity_lca) class(2)) ///
         predict(outcome(systolic_lca) class(2)) ///
         predict(outcome(diastolic_lca) class(2)) ///
         predict(outcome(diabetes_lca) class(2)) ///
         predict(outcome(hba1c_lca) class(2)) subpop(if C_sex==0) ///
          saving(class2_male, replace) ///

marginsplot, recast(bar) title ("Class 2") xtitle("") ///
                xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
                5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
                ytitle("Predicted mean") ylabel(0(.20)1) name (class2_male, replace)




* class 3

margins, predict(outcome(crp_lca) class(3)) ///
         predict(outcome(hdl_lca) class(3)) ///
         predict(outcome(obesity_lca) class(3)) ///
         predict(outcome(systolic_lca) class(3)) ///
         predict(outcome(diastolic_lca) class(3)) ///
         predict(outcome(diabetes_lca) class(3)) ///
         predict(outcome(hba1c_lca) class(3)) subpop(if C_sex==0) ///
         saving(class3_male, replace) ///

marginsplot, recast(bar) title ("Class 3") xtitle("") ///
                xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
                5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
                ytitle("Predicted mean") ylabel(0(.20)1) name (class3_male, replace)



graph combine class1_male class2_male class3_male, cols(3)




* LCA sort by sex
* three-class model

sort C_sex

by C_sex: gsem (crp_lca hdl_lca obesity_lca systolic_lca diastolic_lca diabetes_lca ///
hba1c_lca <- _cons), family(bernoulli) link(logit) lclass(C 3)


estat lcprob

estat lcmean

estat lcgof



/* We can use the predictions of the posterior probability of class membership to evaluate an
individual’s probability of being in each class.

*/
```

```
predict f_classpost1*, classposteriorpr
list in 1, abbrev(10)

/* We can determine the expected class for each individual based on whether the posterior probability
is greater than 0.5
*/

generate f_expclass1 = 1 + (f_classpost11>0.5)
tabulate f_expclass1


generate f_expclass2 = 1 + (f_classpost12>0.5)
tabulate f_expclass2


generate f_expclass3 = 1 + (f_classpost13>0.5)
tabulate f_expclass3



/* We can determine expected classification for each individual in the dataset based on the predicted
posterior class probabilities.
*/

predict f_cpost*, classposteriorpr
egen f_max = rowmax(f_cpost*)


* generate classes var

generate f_predclass = 1 if f_cpost1==f_max

replace f_predclass = 2 if f_cpost2==f_max

replace f_predclass = 3 if f_cpost3==f_max

tabulate f_predclass




* margins and marginsplot for FEMALES

* use margins to calculate marginal predictions
* use marginsplot to graph marginal predictions



margins, predict(classpr class(1)) ///
         predict(classpr class(2)) ///
         predict(classpr class(3)) subpop(if C_sex==1) saving(margin_female, replace)
marginsplot, xtitle ("") ytitle ("") ///
             xlabel (1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title ("Predicted Latent Class Probabilities with 95% CI") ///
             name(margin_female, replace)


margins, predict(classpr class(1)) ///
         predict(classpr class(2)) ///
         predict(classpr class(3)) subpop(if C_sex==1) saving(margin_female, replace)
marginsplot, recast(bar) xtitle("") ytitle("") ///
             xlabel(1 "Class 1" 2 "Class 2" 3 "Class 3") ///
             title("Predicted Latent Class Probabilities with 95% CI") ///
             name(margin_female, replace)

```

```

* class 1

margins, predict(outcome(crp_lca) class(1)) ///
         predict(outcome(hdl_lca) class(1)) ///
         predict(outcome(obesity_lca) class(1)) ///
         predict(outcome(systolic_lca) class(1)) ///
         predict(outcome(diastolic_lca) class(1)) ///
         predict(outcome(diabetes_lca) class(1)) ///
         predict(outcome(hba1c_lca) class(1)) subpop(if C_sex==1) ///
         saving(class1_female, replace) ///

marginsplot, recast(bar) title ("Class 1") xtitle("") ///
             xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
             5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
             ytitle("Predicted mean") ylabel(0(.20)1) name (class1_female, replace)


* class 2

margins, predict(outcome(crp_lca) class(2)) ///
         predict(outcome(hdl_lca) class(2)) ///
         predict(outcome(obesity_lca) class(2)) ///
         predict(outcome(systolic_lca) class(2)) ///
         predict(outcome(diastolic_lca) class(2)) ///
         predict(outcome(diabetes_lca) class(2)) ///
         predict(outcome(hba1c_lca) class(2)) subpop(if C_sex==1) ///
          saving(class2_female, replace) ///

marginsplot, recast(bar) title ("Class 2") xtitle("") ///
             xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
             5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
             ytitle("Predicted mean") ylabel(0(.20)1) name (class2_female, replace)




* class 3

margins, predict(outcome(crp_lca) class(3)) ///
         predict(outcome(hdl_lca) class(3)) ///
         predict(outcome(obesity_lca) class(3)) ///
         predict(outcome(systolic_lca) class(3)) ///
         predict(outcome(diastolic_lca) class(3)) ///
         predict(outcome(diabetes_lca) class(3)) ///
         predict(outcome(hba1c_lca) class(3)) subpop(if C_sex==1)  ///
         saving(class3_female, replace) ///

marginsplot, recast(bar) title ("Class 3") xtitle("") ///
             xlabel(1 "crp" 2 "hdl" 3 "obesity" 4 "systolic BP" ///
             5 "diastolic BP" 6 "diabetes" 7 "hba1c", angle(45)) ///
             ytitle("Predicted mean") ylabel(0(.20)1) name (class3_female, replace)



graph combine class1_female class2_female class3_female, cols(3)






*combine margins male and female class probabilities

graph combine margin_male margin_female, cols(3)

*combine margins male and female 3 classes mean

```

```
graph combine class1_male class2_male class3_male class1_female class2_female class3_female, cols(3)

/* 2) Interaction with age and gender
Survival analysis stratified by age

generate age group variable
Age groups: 1) young old (< 70) 2) old old (>= 70)

Kaplan Meier curves
Cox regression models in imputed data

young old <70
if C_age_group==1

old old >70
if C_age_group==2


*/


gen C_age_group=1 if C_age < 70
replace C_age_group=2 if C_age >=70 & ///
!missing(C_age)

label var C_age_group "Age groups <70 young-old / 70 old-old"
lab def age_group 1 "young old <70" 2 "old old >70"
lab val C_age_group age_group

tab C_age_group


stset C_time_of_event_dementia, failure (Cwv2to4_dementia_event==1) id(id_12char)


* YOUNG OLD <70 Kaplan Meier

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

sort C_lca_group3

sts list if C_age_group==1, by(C_lca_group3)

* graph with frills in males

sts graph if C_age_group==1, by(C_lca_group3) xlabel(0(20)100) ylabel(0.80(.05)1) xtitle("Time in Months") ///
ytitle("Survival Prob") title("Kaplan Meier Curve <70 years") caption("graph04.png", size(vsmall))


* calculate person-time and incidence rates using command ststime

stptime if C_age_group==1,title(Person-years)

stptime if C_age_group==1, title(Person-years) per(1000)


* calculate person-time  by category of C_lca_group3
```

```
stptime if C_age_group==1, by(C_lca_group3)

stptime if C_age_group==1, by(C_lca_group3) per(1000)



* OLD OLD >70 Kaplan Meier

sts list if C_age_group==2, by(C_lca_group3)

* graph with frills in males

sts graph if C_age_group==2, by(C_lca_group3) xlabel(0(20)100) ylabel(0.80(.05)1) xtitle("Time in Months") ///
ytitle("Survival Prob") title("Kaplan Meier Curve >= 70 years") caption("graph04.png", size(vsmall))



* calculate person-time and incidence rates using command ststime

stptime if C_age_group==2,title(Person-years)

stptime if C_age_group==2, title(Person-years) per(1000)


* calculate person-time  by category of C_lca_group3

stptime if C_age_group==2, by(C_lca_group3)

stptime if C_age_group==2, by(C_lca_group3) per(1000)






* COX PH REGRESSION MODEL IN IMPUTED DATASET


* Declare Data to be Survival Data by using mi

mi stset C_time_of_event_dementia, failure (Cwv2to4_dementia_event==1) id(id_12char)





*** INTERACTION gender*cardiometabolic cluster ***

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 i.C_sex#i.C_lca_group3



mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
C_age i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms i.C_sex#i.C_lca_group3




*** INTERACTION age*cardiometabolic cluster ***

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 c.C_age#i.C_lca_group3
```

```
mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms c.C_age#i.C_lca_group3





* YOUNG OLD <70 Cox regression models

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

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 if C_age_group==1


* Adjusted models - multivariable Cox model
* controlling for covariates

* Model 2: model 1 + adjust for sociodemographics: age education marital status and wealth

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles if C_age_group==1

* Model 3: model 2 + adjust for cvd health

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity if C_age_group==1


* Model 4: model 3 + adjust for depression

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms if C_age_group==1




* OLD OLD >70 Cox regression models

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

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 if C_age_group==2


* Adjusted models - multivariable Cox model
* controlling for covariates

* Model 2: model 1 + adjust for sociodemographics: sex education marital status and wealth

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles if C_age_group==2

* Model 3: model 2 + adjust for cvd health

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity if C_age_group==2


* Model 4: model 3 + adjust for depression

mi estimate, eform("Haz. Ratio"): stcox i.C_lca_group3 ///
i.C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
```

```
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status i.C_cvd_comorbidity ///
i.Cwv1_depressive_symptoms if C_age_group==2







/*

3) exclude participants with cvd

use the command if C_cvd_comorbidity==0

*/



* Declare Data to be Survival Data by using mi

stset C_time_of_event_dementia, failure (Cwv2to4_dementia_event==1) id(id_12char)



* define design var by using i.(3 classes)

stcox i.C_lca_group3 if C_cvd_comorbidity==0


* Adjusted models - multivariable Cox model
* controlling for covariates

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

stcox i.C_lca_group3 C_age C_sex i.C_educ_new i.C_maritalstatus_4cat ///
i.Cwv1_netwealth_quintiles if C_cvd_comorbidity==0

* model 3: model 2 + adjust for cvd health

stcox i.C_lca_group3 C_age C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status if C_cvd_comorbidity==0


* model 4:  model 3 + adjust for depression
stcox i.C_lca_group3 C_age C_sex i.C_educ_new i.C_maritalstatus_4cat i.Cwv1_netwealth_quintiles ///
i.Cwv1_smoking_3cat i.Cwv1_alcohol_status ///
i.Cwv1_depressive_symptoms if C_cvd_comorbidity==0





* 4) complete data (see above)


* ------------------------------ *






*** EXTRA SENSITIVITY ANALYSES FOR THE PAPER ***

/*

compare baseline characteristics between complete sample (before exclusion) and sample with
missing data (overall after exclusion)
```

```
*/



* General characteristics of CHARLS participants at baseline

* Socio-demographics
sum C_age
ta C_sex
ta C_educ_new
ta C_maritalstatus_4cat
ta Cwv1_netwealth_quintiles
* Cardiometabolic disorders
ta Cwv1_crp
ta Cwv1_hdl_cholesterol
ta Cwv1_obesity_waist
ta Cwv1_systolic_bp
ta Cwv1_diastolic_bp
ta Cwv1_diabetes_report
ta Cwv1_HbA1c
* Cardiovascular health factors
ta Cwv1_smoking_3cat
ta Cwv1_alcohol_status
ta Cwv1_physicalactivity
ta C_cvd_comorbidity
* Depressive symptoms
ta Cwv1_depressive_symptoms
* Memory score
sum Cwv1_memory_wordrecall




* compare health characteristics between those survived and dropped out



*** CLEANING DATA to keep those who dropped out



* 1. drop dementia cases at baseline

* drop dementia wave 2 missing data
drop if Cwv1_dementia_report==1
* (267 observations deleted)
drop if Cwv1_dementia_report== .
* (88 observations deleted)


* 2. drop missing values of cardiometabolic markers

drop if Cwv1_crp== .
* (180 observations deleted)

drop if Cwv1_hdl_cholesterol== .
* (2 observations deleted)

drop if Cwv1_obesity_waist== .
* (1688 observations deleted)

drop if Cwv1_systolic_bp== .
* (83 observations deleted)

drop if Cwv1_diastolic_bp== .
* (13 observations deleted)

```

```
drop if Cwv1_diabetes_report== .
* (91 observations deleted)

drop if Cwv1_HbA1c== .
* (71 observations deleted)



* 3. drop obs with no records on dementia at any wave from 2-4 follow-ups


search mdesc
search rmiss2
search mvpatterns

* see number of missing values vs non-missing in each variable
mdesc Cwv2_dementia_report Cwv3_dementia_report Cwv4_self_info_dementia



/* 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_wv2to4=rmiss2(Cwv2_dementia_report ///
Cwv3_dementia_report Cwv4_self_info_dementia)

tab nmisfollowup_dementia_wv2to4

* drop observations "nmisfollowup_dementia_wv2to4" < 3
drop if nmisfollowup_dementia_wv2to4<3


* General characteristics of CHARLS participants at baseline

* Socio-demographics
sum C_age
ta C_sex
ta C_educ_new
ta C_maritalstatus_4cat
ta Cwv1_netwealth_quintiles
* Cardiometabolic disorders
ta Cwv1_crp
ta Cwv1_hdl_cholesterol
ta Cwv1_obesity_waist
ta Cwv1_systolic_bp
ta Cwv1_diastolic_bp
ta Cwv1_diabetes_report
ta Cwv1_HbA1c
* Cardiovascular health factors
ta Cwv1_smoking_3cat
ta Cwv1_alcohol_status
ta Cwv1_physicalactivity
ta C_cvd_comorbidity
* Depressive symptoms
ta Cwv1_depressive_symptoms
* Memory score
sum Cwv1_memory_wordrecall



* compare health characteristics bewteen <70 and >=70



* General baseline characteristics of CHARLS participants by age group

* Socio-demographics
ttest C_age, by(C_age_group)
```

```
ta C_sex C_age_group, chi2 column row
ta C_educ_new C_age_group, chi2 column row
ta C_maritalstatus_4cat C_age_group, chi2 column row
ta Cwv1_netwealth_quintiles C_age_group, chi2 column row
* Cardiometabolic disorders
ta crp_lca C_age_group, chi2 column row
ta hdl_lca C_age_group, chi2 column row
ta obesity_lca C_age_group, chi2 column row
ta systolic_lca C_age_group, chi2 column row
ta diastolic_lca C_age_group, chi2 column row
ta diabetes_lca C_age_group, chi2 column row
ta hba1c_lca C_age_group, chi2 column row
* Cardiovascular health factors
ta Cwv1_smoking_3cat C_age_group, chi2 column row
ta Cwv1_alcohol_status C_age_group, chi2 column row
ta Cwv1_physicalactivity C_age_group, chi2 column row
ta C_cvd_comorbidity C_age_group, chi2 column row
* Depressive symptoms
ta Cwv1_depressive_symptoms C_age_group, chi2 column row
* Memory score
ttest Cwv1_memory_wordrecall, by(C_age_group)
ta C_lca_group3 C_age_group, chi2 column row




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