

```

1  /*
2
3  PHD PROJECT: The role of depressive symptoms and cardiometabolic risk factors in the prediction
4  of dementia: a cross-country comparison in England, the United States and China
5
6  STUDY 3: Trajectories of depressive symptoms and their relationship with dementia
7
8  Method of analysis:
9  Group-based trajectory modelling (GBTM) approach
10 Latent Class Growth Analysis (LCGA)
11
12 TIMELINE
13
14 DEPRESS TRAJECTORIES: WV8 - WV11 (4 TIME POINTS)
15 DEMENTIA INCIDENCE AT YEAR 6 (WV11)
16 DEMENTIA INCIDENCE: W12 - WV14 (3 TIME POINTS)
17
18
19
20 */
21
22
23
24
25 * importing data (.dta)
26
27 use "S:\Research\pkstudies\Study1_traj_depression\HRS\1. hrs traj depress dementia
28 risk\hrs_data_w8_14_study1.dta"
29
30
31 * KEEP NECESSARY VARIABLES
32
33
34 keep HHID PN RAHHIDPN HHIDPN ///
35 H_sex H_age H_education_yrs H_education H_maritalstatus_3cat H_maritalstatus_4cat
36 H_wealthquintiles ///
37 H_ethnicity H_hispanic_ethnicity ///
38 H_smoking_2cat H_smoking_3cat H_physicalactivity H_alcohol_freq H_alcohol_status ///
39 H_heart_disease H_stroke H_cvd_comorbidity Hwv8_cognition ///
40 Hwv8_cesd_depressed Hwv8_cesd_effort Hwv8_cesd_sleep Hwv8_cesd_happy Hwv8_cesd_lonely
41 Hwv8_cesd_enlife Hwv8_cesd_sad Hwv8_cesd_going Hwv8_cesd_score cesd_0 Hwv8_depressive_symptoms ///
42 Hwv9_cesd_depressed Hwv9_cesd_effort Hwv9_cesd_sleep Hwv9_cesd_happy Hwv9_cesd_lonely
43 Hwv9_cesd_enlife Hwv9_cesd_sad Hwv9_cesd_going Hwv9_cesd_score cesd_1 Hwv9_depressive_symptoms ///
44 Hwv10_cesd_depressed Hwv10_cesd_effort Hwv10_cesd_sleep Hwv10_cesd_happy Hwv10_cesd_lonely
45 Hwv10_cesd_enlife Hwv10_cesd_sad Hwv10_cesd_going Hwv10_cesd_score cesd_2
46 Hwv10_depressive_symptoms ///
47 Hwv11_cesd_depressed Hwv11_cesd_effort Hwv11_cesd_sleep Hwv11_cesd_happy Hwv11_cesd_lonely
48 Hwv11_cesd_enlife Hwv11_cesd_sad Hwv11_cesd_going Hwv11_cesd_score cesd_3
49 Hwv11_depressive_symptoms ///
50 Hwv12_cesd_depressed Hwv12_cesd_effort Hwv12_cesd_sleep Hwv12_cesd_happy Hwv12_cesd_lonely
51 Hwv12_cesd_enlife Hwv12_cesd_sad Hwv12_cesd_going Hwv12_cesd_score Hwv12_cesd_sumscore
52 Hwv12_depressive_symptoms ///
53 Hwv13_cesd_depressed Hwv13_cesd_effort Hwv13_cesd_sleep Hwv13_cesd_happy Hwv13_cesd_lonely
54 Hwv13_cesd_enlife Hwv13_cesd_sad Hwv13_cesd_going Hwv13_cesd_score Hwv13_cesd_sumscore
55 Hwv13_depressive_symptoms ///
56 Hwv14_cesd_happy Hwv14_cesd_enlife Hwv14_cesd_depressed Hwv14_cesd_effort Hwv14_cesd_sleep
57 Hwv14_cesd_lonely Hwv14_cesd_sad Hwv14_cesd_going Hwv14_cesd_sumscore Hwv14_depressive_symptoms ///
58 Hwv8_memory_report Hwv9_memory_report Hwv10_anydementia_report Hwv11_anydementia_report
59 Hwv12_anydementia_report Hwv13_anydementia_report Hwv14_anydementia_report ///
60 Hwv8_interview_date Hwv9_interview_date Hwv10_interview_date Hwv11_interview_date
61 Hwv12_interview_date Hwv13_interview_date Hwv14_interview_date ///
62 Hwv12to14_dementia_sum Hwv12to14_dementia_event ///
63 Hwv12to14_newdementia_or_lastint Hwv12to14_time_dementia_months Hwv12to14_dementia_free_date
64 Hwv12to14_time_dementia_midpoint Hwv12to14_time_dementia_midpoint0 Hwv12to14_time_of_event_dementia
65 ///
66 t_0 t_1 t_2 t_3 nmisfollowup_cesd nmisfollowup_dementia_wv12to14

```

```

51
52
53
54
55
56 /* Latent class growth analysis (LCGA) of depressive symptoms */
57
58
59 * installing traj command
60
61 net from http://www.andrew.cmu.edu/user/bjones/traj
62 net install traj, force
63 help traj
64
65
66
67 * Generate a set of time variables to pass to traj, from wave 2 to 9 -> 8 time points (t0-t7)
68 forval i = 0/3 {
69     generate t_`i' = `i'
70 }
71
72 *recode time in months
73
74 recode t_1 (1=24)
75 recode t_2 (2=48)
76 recode t_3 (3=72)
77
78
79 *rename cesd score across the waves - discrete var min=0 max=8
80 * use Hwv8_cesd_sumscore continuous variables
81
82 rename Hwv8_cesd_sumscore cesd_0
83 rename Hwv9_cesd_sumscore cesd_1
84 rename Hwv10_cesd_sumscore cesd_2
85 rename Hwv11_cesd_sumscore cesd_3
86
87
88
89
90
91
92 *** Descriptive stats of depression and dementia
93
94 tabulate cesd_0
95 summarize cesd_0 , detail
96 histogram cesd_0, discrete frequency normal
97
98 misstable summarize cesd_0
99 misstable patterns cesd_0
100
101 tabulate Hwv8_depressive_symptoms
102 summarize Hwv8_depressive_symptoms
103
104 misstable summarize Hwv8_depressive_symptoms
105 misstable patterns Hwv8_depressive_symptoms
106
107 tabulate Hwv8_memory_report
108 summarize Hwv8_memory_report
109
110 misstable summarize Hwv8_memory_report
111 misstable patterns Hwv8_memory_report
112
113
114 tabulate Hwv9_memory_report
115 summarize Hwv9_memory_report
116
117 misstable summarize Hwv9_memory_report
118 misstable patterns Hwv9_memory_report

```

```

119
120
121 tabulate Hwv10_anydementia_report
122 summarize Hwv10_anydementia_report
123
124 misstable summarize Hwv10_anydementia_report
125 misstable patterns Hwv10_anydementia_report
126
127 tabulate Hwv11_anydementia_report
128 summarize Hwv11_anydementia_report
129
130 misstable summarize Hwv11_anydementia_report
131 misstable patterns Hwv11_anydementia_report
132
133
134 tabulate Hwv12_anydementia_report
135 summarize Hwv12_anydementia_report
136
137 misstable summarize Hwv12_anydementia_report
138 misstable patterns Hwv12_anydementia_report
139
140
141 tabulate Hwv13_anydementia_report
142 summarize Hwv13_anydementia_report
143
144 misstable summarize Hwv13_anydementia_report
145 misstable patterns Hwv13_anydementia_report
146
147 tabulate Hwv14_anydementia_report
148 summarize Hwv14_anydementia_report
149
150 misstable summarize Hwv14_anydementia_report
151 misstable patterns Hwv14_anydementia_report
152
153
154
155
156
157
158
159
160 *** CLEANING DATA
161
162 * 1. drop missing data depression and dementia at baseline
163 * drop 48 depression missing data
164 * no missing data for baseline dementia
165
166 drop if cesd_0== .
167 * (48 observations deleted)
168
169
170 * 2. drop dementia cases between wv8 and wv11 (total: 547 cases)
171
172 drop if Hwv8_memory_report==1
173 * (222 observations deleted)
174
175 drop if Hwv9_memory_report==1
176 * (109 observations deleted)
177
178 drop if Hwv10_anydementia_report==1
179 * (104 observations deleted)
180
181
182
183 * 3. process to drop missing data depression in at least 2 follow-up waves
184
185 /*
186

```

```

187 check below how to see number of missing values in an observation (case) and patterns of missing
188 values
189 https://stats.idre.ucla.edu/stata/faq/how-can-i-see-the-number-of-missing-values-and-patterns-of-mi
190 ssing-values-in-my-data-file/
191 install packages:
192 * install mdesc
193 * install tabmiss
194 * insatll dm31
195 * insall mvpatterna
196
197 */
198
199 search mdesc
200 search rmiss2
201 search mvpatterns
202
203 * see number of missing values vs non-missing in each variable
204 mdesc cesd_0 cesd_1 cesd_2 cesd_3
205 *or
206 mdesc cesd_*
207
208 * number of missing values per observation
209 * the code below creates a variable called nmisfollowup that gives the number of missing values
210 for each observation in the variables of interest
211 egen nmisfollowup_cesd=rmiss2(cesd_1 cesd_2 cesd_3)
212
213 tab nmisfollowup_cesd
214
215 * drop observations "nmisfollowup_cesd" > 1 (those with 2 or 3 follow ups of missing data)
216 drop if nmisfollowup>1
217 *(917 observations deleted)
218
219
220
221
222 * 4. drop obs with no records on dementia at any wave from 12-14 follow-ups
223
224 * see number of missing values vs non-missing in each variable
225 mdesc Hwv12_anydementia_report Hwv13_anydementia_report Hwv14_anydementia_report
226
227
228
229 /* number of missing values per observation
230 * the code below creates a variable called nmisfollowup that gives the number of missing values
231 for each observation in the variables of interest */
232 egen nmisfollowup_dementia_wv12to14=rmiss2(Hwv12_anydementia_report Hwv13_anydementia_report ///
233 Hwv14_anydementia_report)
234
235 tab nmisfollowup_dementia_wv12to14
236
237 * drop observations "nmisfollowup_dementia_wv6to9" > 2 (with 3 followups of missing data = no
238 records at any wave)
239 drop if nmisfollowup_dementia_wv12to14>2
240 *(748 observations deleted)
241
242
243
244
245
246 *descriptive stats of depressive symptoms cesd
247
248
249 tabulate cesd_0
250 summarize cesd_0, detail

```

```

251 histogram cesd_0, discrete frequency normal
252
253 tabulate cesd_1
254 summarize cesd_1 , detail
255 histogram cesd_1, discrete frequency normal
256
257 tabulate cesd_2
258 summarize cesd_2, detail
259 histogram cesd_2, discrete frequency normal
260
261 tabulate cesd_3
262 summarize cesd_3, detail
263 histogram cesd_3, discrete frequency normal
264
265
266 ta cesd_0, miss
267 ta cesd_1, miss
268 ta cesd_2, miss
269 ta cesd_3, miss
270
271
272 tabstat cesd_0, by(H_sex)stats (mean v n)
273 tabstat cesd_1, by(H_sex)stats (mean v n)
274 tabstat cesd_2, by(H_sex)stats (mean v n)
275 tabstat cesd_3, by(H_sex)stats (mean v n)
276
277
278
279 tabstat cesd_*,s(n me sk) by(H_sex)
280
281
282
283
284 bysort H_sex: tab cesd_0
285 bysort H_sex: tab cesd_1
286 bysort H_sex: tab cesd_2
287 bysort H_sex: tab cesd_3
288
289
290 tabstat cesd_0 cesd_1 cesd_2 cesd_3, s(sk kur)
291 sktest cesd_0 cesd_1 cesd_2 cesd_3
292
293
294
295 * missingness pateterns
296 misstable patterns cesd_*
297 * "1" means that the variable is observed and a "0" represents missing
298
299
300 * box plots of the observations at each occasion
301 graph box cesd_0 cesd_1 cesd_2 cesd_3, ascategory intensity (0) medtype (line)
302
303
304
305
306
307
308
309
310
311 /*
312 LCGA analysis
313
314
315 useful sources
316 use http://www.andrew.cmu.edu/user/bjones/traj/data/cambrdge.dta,clear
317 https://ssrc.indiana.edu/doc/wimdocs/2013-03-29\_nagin\_trajectory\_stata-plugin-info.pdf
318

```

```

319
320   traj [if], var(varlist) indep(varlist) model(modeltype)
321         order(numlist) [additional options]
322
323
324   order(numlist)      0=intercept, 1=linear, 2=quadratic, 3=cubic -
325                       polynomial type for each group trajectory
326
327
328   ci                  parametric bootstrap confidence intervals of
329                       individual distal outcome and probability of
330                       group memberships.
331
332
333
334
335
336
337
338
339   Available Models -> command traj
340
341
342   Censored normal (CNORM) model distribution
343
344   traj, var(qcp*op) indep(age*) model(cnorm) min(0) max(999) order(1 3 2)
345
346   trajplot, xtitle(Age) ytitle(Opposition) xlabel(6(1)15)
347             ylabel(0(1)6)
348
349   /* Shows the assigned group and probabilities of group membership */
350   list _traj_Group - _traj_ProbG3 if _n < 3, ab(12)
351
352   /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
353   matrix list e(plot1), format(%9.2f) noheader
354
355   /* Including time-stable covariates (risk) associated with group membership */
356
357   traj, var(qcp*op) indep(age*) model(cnorm) min(0) max(10) order(1 3
358             2) risk(scolmer scolper)
359
360
361
362   Zero Inflated poisson (ZIP) Model
363
364   It is an analysis of Poisson data with extra zeros
365
366
367   traj, model(zip) var(y*) indep(t*) order(2 1 3) iorder(1)
368
369   trajplot, xtitle(Age) ytitle(Opposition) ci
370
371
372
373   Time-Stable Covariates for Group Membership
374
375   traj, var(qcp*op) indep(age*) model(cnorm) min(0) max(10) order(1 3
376             2) risk(scolmer scolper)
377
378   trajplot, xtitle(Age) ytitle(Opposition)
379
380
381   Logistic (logit) model
382
383   use http://www.andrew.cmu.edu/user/bjones/traj/data/cambridge.dta,
384   clear
385
386   traj, var(p1-p23) indep(tt1-tt23) model(logit) order(0 3 3)

```

```

387
388     trajplot, xtitle(Scaled Age) ytitle(Prevalence)
389

```

```

390     /* Assigned group and probabilities of group membership */
391     list _traj_Group - _traj_ProbG3 if _n > 400, ab(12)
392
393
394
395

```

396 Model selection:

- 397
- 398 1. Type of model: The ‘traj’ can model normal, censored normal, zero-inflated Poisson and binary  
logit models.
- 399 Capacity for incorporating effect of time-stable and time-varying covariates,  
400 subsequent outcomes and joint trajectory models.
- 401
- 402 2. Number of groups/classes: determination of the optimal number of groups to compose the mixture
- 403
- 404 3. Shape of the trajectory: determination of the appropriate order of the  
405 polynomial used to model each group's trajectory (linear, quadratic, cubic).
- 406
- 407
- 408

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

- 410
- 411 • Bayesian Information Criteria (BIC), where lower BIC or least negative BIC  
412 (higher value closer to zero) represents a better fitting model.
  - 413
  - 414 • Bayes Factor greater than 10 indicates very strong evidence  
415 to use the “more complex” model.
  - 416
  - 417 • Meaningful proportion of participants within each class  
418 (smallest group percentage to be higher or equal to 5%).
  - 419
  - 420 • Average posterior probability (APP) to belong to each class higher than 0.70.
  - 421
  - 422 • Entropy to determine the accuracy of classification of individuals into the different latent  
classes
- 423 If entropy is near 1.0, then classification of individuals is assumed to be adequate.
- 424 If entropy is near 0, then classification is assumed to be poor.
- 425
- 426
- 427
- 428
- 429
- 430
- 431

```

432 *****function to print out summary stats
433 program summary_table_procTraj
434     preserve
435     *look at the average posterior probability
436     gen Mp = 0
437     foreach i of varlist _traj_ProbG* {
438         replace Mp = `i' if `i' > Mp
439     }
440     sort _traj_Group
441     *and the odds of correct classification
442     by _traj_Group: gen countG = _N
443     by _traj_Group: egen groupAPP = mean(Mp)
444     by _traj_Group: gen counter = _n
445     gen n = groupAPP/(1 - groupAPP)
446     gen p = countG/_N
447     gen d = p/(1-p)
448     gen occ = n/d
449     *Estimated proportion for each group
450     scalar c = 0
451     gen TotProb = 0
452     foreach i of varlist _traj_ProbG* {

```

```

453         scalar c = c + 1
454         quietly summarize `i'
455         replace TotProb = r(sum)/ _N if _traj_Group == c
456     }
457     gen d_pp = TotProb/(1 - TotProb)
458     gen occ_pp = n/d_pp
459     *This displays the group number [_traj_~p],
460     *the count per group (based on the max post prob), [countG]
461     *the average posterior probability for each group, [groupAPP]
462     *the odds of correct classification (based on the max post prob group assignment), [occ]
463     *the odds of correct classification (based on the weighted post. prob), [occ_pp]
464     *and the observed probability of groups versus the probability [p]
465     *based on the posterior probabilities [TotProb]
466     list _traj_Group countG groupAPP occ occ_pp p TotProb if counter == 1
467     restore
468 end
469
470 summary_table_procTraj
471
472
473
474 ***** to generate a plot of the individual trajectories
475
476 preserve
477 reshape long count_ t_, i(id)
478
479 gen count_jit = count_ + ( 0.2*runiform()-0.1 )
480 graph twoway scatter count_jit t_, c(L) by(_traj_Group) msize(tiny) mcolor(gray) lwidth(vthin)
481     lcolor(gray)
482
483 ***** to calculate the Bayes factor
484
485 log Bayes factor (2loge(B10)  $\approx$  2( $\Delta$ BIC)
486 This estimate approximately equals 2(BICcomplex model-BICnull model)
487
488 */
489
490
491
492
493
494 /*
495
496 Depressive symptoms (CES-D 8 item)
497 The trajectory groups of the CES-D scores (as a discrete variable) are tested
498 alone with time as the only independent variable, with no covariates added that could influence
499 class membership.
500 The Zero Inflated poisson model ('cnorm') is applied, given that the CES-D 8-item was a count of
501 symptoms and the majority of individuals scored 0 at each time point.
502
503 Initially, for each model, the linear, quadratic, and cubic functions of each trajectory can be
504 tested,
505 depending on the number of time points.
506 To ensure parsimony, consistent with the recommendations of Helgeson, Snyder, and Seltman (2004),
507 non-significant cubic and quadratic terms are removed from trajectories in a given model,
508 but linear parameters are retained irrespective of significance.
509
510 I tested the best fitting model with two, three, four five and then six trajectories following
511 the same process.
512 The models were compared (in a table of comparison) using BIC statistics,
513 Bayes factor, entropy, percentage of each class and average posterior probabilities.
514
515 PROCESS TO SELECT THE BEST-FITTING MODEL
516
517 Shape and Classes

```



```

516
517 1. run one traj with quadratic (order 2)
518 - If quadratic is not significant run with linear parameter (order 1)
519
520 2. model with 2 traj with quadratic (order 2 2)
521 - If neither traj is significant rerun with linear (order 1 1)
522 - If one not significant adapt accordingly (e.g. order 1 2 OR order 2 1)
523
524 3. Compare models (complex-simple) with statistic criteria
525
526 4. Repeat the process with an increasing number of traj
527
528 */
529
530
531
532
533
534
535
536 *** ZIP MODEL
537
538
539 * 1 class
540 traj, var(cesd_*) indep(t_*) model(zip) order(2) iorder(1)
541 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
542
543 /* Shows the assigned group and probabilities of group membership */
544 list _traj_Group - _traj_ProbG1 if _n < 3, ab(12)
545
546 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
547 matrix list e(plot1), format(%9.2f) noheader
548
549
550
551
552 * 2 classes
553 traj, var(cesd_*) indep(t_*) model(zip) order(2 2) iorder(1)
554 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
555
556 /* Shows the assigned group and probabilities of group membership */
557 list _traj_Group - _traj_ProbG2 if _n < 3, ab(12)
558
559 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
560 matrix list e(plot1), format(%9.2f) noheader
561
562
563
564
565
566 * 3 classes
567 traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2) iorder(1)
568 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
569
570 /* Shows the assigned group and probabilities of group membership */
571 list _traj_Group - _traj_ProbG3 if _n < 3, ab(12)
572
573 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
574 matrix list e(plot1), format(%9.2f) noheader
575
576
577
578
579
580
581 * 4 classes
582 traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2 2) iorder(1)
583 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci

```

```

584
585
586 /* Shows the assigned group and probabilities of group membership */
587     list _traj_Group - _traj_ProbG4 if _n < 3, ab(12)
588
589 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
590     matrix list e(plot1), format(%9.2f) noheader
591
592
593
594 * 5 classes
595 traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2 2 2) iorder(1)
596 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
597
598
599
600 /* Shows the assigned group and probabilities of group membership */
601     list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)
602
603 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
604     matrix list e(plot1), format(%9.2f) noheader
605
606
607 * 6 classes
608 traj, var(cesd_*) indep(t_*) model(zip) order(2 2 2 2 2 2) iorder(1)
609 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
610
611
612
613 /* Shows the assigned group and probabilities of group membership */
614     list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)
615
616 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
617     matrix list e(plot1), format(%9.2f) noheader
618
619
620
621
622 ** the 5-class model fits better
623
624
625 * 5 classes
626 traj, var(cesd_*) indep(t_*) model(zip) order(3 3 3 3 3) iorder(1)
627 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
628
629
630
631 /* Shows the assigned group and probabilities of group membership */
632     list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)
633
634 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
635     matrix list e(plot1), format(%9.2f) noheader
636
637
638
639 *** OPTIMAL ZIP MODEL
640
641 * 5 groups - zip model - cubic and linear polynomial (1 1 3 3 1)
642 traj, var(cesd_*) indep(t_*) model(zip) order(1 1 3 3 1) iorder(1)
643 trajplot, xtitle(Time in Months) ytitle(Depressive symptoms CES-D) ci
644
645
646 /* Shows the assigned group and probabilities of group membership */
647     list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)
648
649 /* trajT = x-axis, Avg# = data averages, Est# = model estimates */
650     matrix list e(plot1), format(%9.2f) noheader
651

```

```

652
653
654
655
656
657
658
659 ** run after each traj model to estimate the average posterior probability (APP) for each group
660
661 program summary_table_procTraj
662     preserve
663     *look at the average posterior probability
664     gen Mp = 0
665     foreach i of varlist _traj_ProbG* {
666         replace Mp = `i' if `i' > Mp
667     }
668     sort _traj_Group
669     *and the odds of correct classification
670     by _traj_Group: gen cesdG = _N
671     by _traj_Group: egen groupAPP = mean(Mp)
672     by _traj_Group: gen counter = _n
673     gen n = groupAPP/(1 - groupAPP)
674     gen p = cesdG/_N
675     gen d = p/(1-p)
676     gen occ = n/d
677     *Estimated proportion for each group
678     scalar c = 0
679     gen TotProb = 0
680     foreach i of varlist _traj_ProbG* {
681         scalar c = c + 1
682         quietly summarize `i'
683         replace TotProb = r(sum)/ _N if _traj_Group == c
684     }
685     gen d_pp = TotProb/(1 - TotProb)
686     gen occ_pp = n/d_pp
687     *This displays the group number [_traj_~p],
688     *the cesd per group (based on the max post prob), [countG]
689     *the average posterior probability for each group, [groupAPP]
690     *the odds of correct classification (based on the max post prob group assignment), [occ]
691     *the odds of correct classification (based on the weighted post. prob), [occ_pp]
692     *and the observed probability of groups versus the probability [p]
693     *based on the posterior probabilities [TotProb]
694     list _traj_Group cesdG groupAPP occ occ_pp p TotProb if counter == 1
695     restore
696 end
697
698 summary_table_procTraj
699
700
701
702
703
704
705 /*
706 ---- MODEL SELECTION ----
707 Best-fitting model to try survival analysis is the
708 5 class - order (1 1 3 3 1)
709 */
710
711
712
713
714 /*
715 Data and variable manipulation
716 */
717
718 * 5-class model: rename _traj_Group to H_traj_group5
719

```

```

720 rename _traj_Group H_traj_group5
721 recode H_traj_group5 (4=5) (5=4)
722 ta H_traj_group5
723 rename _traj_ProbG1 H_depres_traj_1
724 rename _traj_ProbG2 H_depres_traj_2
725 rename _traj_ProbG3 H_depres_traj_3
726 rename _traj_ProbG4 H_depres_traj_4
727 rename _traj_ProbG5 H_depres_traj_5
728
729
730 * labelling variable of H_traj_group5
731
732 label var H_traj_group5 "Traj 5 groups of depressive symptoms"
733
734 * labelling values
735 lab def traj_depres 1 "minimal" 2 "mild" 3 "increasing" 4 "decreasing" 5 "high"
736
737 * attach category labels to the variable through label value
738
739 lab val H_traj_group5 traj_depres
740
741 ta H_traj_group5
742
743
744
745
746
747 * Frequencies of covariates
748
749 tabulate H_age
750 summarize H_age
751
752
753 tabulate H_sex
754 summarize H_sex
755
756
757 tabulate H_eduaction
758 summarize H_eduaction
759
760
761 tabulate H_maritalstatus_4cat
762 summarize H_maritalstatus_4cat
763
764
765 tabulate H_wealthquintiles
766 summarize H_wealthquintiles
767
768
769 tabulate Hwv8_smoking_3cat
770 summarize Hwv8_smoking_3cat
771
772
773 tabulate Hwv8_physicalactivity
774 summarize Hwv8_physicalactivity
775
776
777 tabulate Hwv8_alcohol_status
778 summarize Hwv8_alcohol_status
779
780
781 tabulate Hwv8_cvd_comorbidity
782 summarize Hwv8_cvd_comorbidity
783
784
785 tabulate Hwv8_memory
786 summarize Hwv8_memory
787

```

```

788
789
790
791
792
793
794  /* MULTIPLE IMPUTATION (MI)
795
796  To handle with missing values of covariates
797
798
799  useful sources for MI and MICE:
800
801  https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/
802  https://www.stata.com/manuals/mi.pdf - see page 139
803  https://www.stata.com/meeting/switzerland16/slides/medeiros-switzerland16.pdf
804  https://www.youtube.com/watch?v=i6S0lq0mjuc&ab_channel=StataCorpLLC
805  https://dss.princeton.edu/training/MIStata.pdf
806
807
808
809  Preparing to conduct MI
810  1. examine the number and proportion of missing values among the variables of interest
811      use the mdesc command
812  2. examine missing data patterns
813      use commands mi set and mi misstable patterns
814  3. identify potential auxiliary variables
815
816
817  Run MI using chained equations (MICE)
818  using the commands
819  1. how (in what style) to store the imputations
820      mi set wide
821  2. which variables will be imputed
822      mi register imputed
823  3. optionally, which variables will not be imputed
824      mi register regular
825  4. what imputation method is implemented to impute each of var - MICE
826      mi impute chained
827
828  */
829
830
831
832
833
834  /*
835
836  1. examining missing values
837      install packages:
838      * install mdesc
839      * install tabmiss
840      * insatll dm31
841      * insall mvpatterna
842
843  */
844
845  search mdesc
846  search rmiss2
847  search mvpatterns
848
849
850
851
852
853
854  * examining number of missing values vs non-missing in each variable
855

```

```

856 mdesc H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
857 Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
858 Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl Hwv12_crp ///
859 Hwv12_memory
860
861
862
863
864 * examining missing data patterns
865
866 mi set wide
867
868 mi misstable summarize H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
869 Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
870 Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl Hwv12_crp ///
871 Hwv12_memory
872
873
874 mi misstable patterns H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
875 Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
876 Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl Hwv12_crp ///
877 Hwv12_memory
878
879
880
881
882 /*
883     identifying potential auxiliary var
884 * Auxiliary variables are either correlated with a missing variable(s)
885 (the recommendation is  $r > 0.4$ ) or are believed to be associated with missingness
886 - a priori knowledge of var that would make good auxiliary var
887 - identify potential candidates by examining associations between missing var and other var in
the dataset
888     running correlation using the command: pwcorr v1 v2 v3, obs
889     the recommendation for good correlation is  $r > 0.4$ 
890
891
892 Missing var to be imputed:
893
894     Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status
895     Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl
896     Hwv12_memory
897
898
899
900 Potential auxiliary var:
901 DV: Hwv12to14_dementia_event
902 IV: H_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3
903 other var:
904     H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles Hwv8_depressive_symptoms
905
906 */
907
908
909 * correlation
910
911 pwcorr Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_status ///
912     Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl ///
913     Hwv12_memory ///
914     Hwv12to14_dementia_event H_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3 ///
915     H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
916     Hwv8_depressive_symptoms, obs
917
918
919 * The correlation showed that all the above potential var are good auxiliary
920 * A good auxiliary does not have to be correlated with every variable to be useful
921 * And it's not problematic if it has missing info of it's own
922

```

```

923
924
925
926  /*
927  MI by chained equations (MICE)
928      see: https://stats.idre.ucla.edu/stata/seminars/mi\_in\_stata\_pt1\_new/
929
930  MICE is known as the fully conditional specification or sequential generalized regression
931  does not assume a joint MVN distribution
932  but instead uses a separate conditional distribution for each imputed variable.
933
934  The multivariate normal (MVN) model - mi imputed mvn -
935  assumes multivariate normality of all var
936
937  The multivariate imputation by chained equations (MICE) - mi imputed chained -
938  offers flexibility in how each var is modeled
939
940  mi impute chained allows to specify models for a
941  variety of variable types, including
942  continuous, binary, ordinal, nominal, truncated, and count variables
943
944
945  The MICE distributions available in Stata are:
946  binary, ordered and multinomial logistic regression for categorical variables,
947  linear regression and predictive mean matching (PMM)* for continuous variables,
948  and Poisson and negative binomial regression for count variables.
949
950
951
952  IMPUTATION PHASES
953
954  1. mi set wide
955      style to store imputations
956
957  2. mi register imputed
958      identifies which variables in the imputation model have missing information.
959
960  3. mi register regular (! optional)
961      which variables will not be imputed
962
963  4. mi impute chained
964      where the user specifies the imputation model to be used
965      and the number of imputed datasets to be created.
966      Example:
967          mi impute chained (regress) bmi age (logit) female ///
968          (mlogit) race = bpdiastr i.region, add(20)
969
970  5. mi estimate
971      is used as a prefix to the standard regress command.
972      This executes the specified estimation model within each of the 20 imputed datasets
973      to obtain 20 sets of coefficients and standard errors.
974      Stata then combines these estimates to obtain one set of inferential statistics.
975      In the output from mi estimate you will see some metrics: Imputation Diagnostics
976      information for RVI (Relative Increase in Variance),
977      FMI (Fraction of Missing Information),
978      DF (Degrees of Freedom) ,
979      RE (Relative Efficiency),
980      and the between imputation and the within imputation variance estimates
981      to examine how the standard errors (SEs) are calculated.
982
983
984  -----
985
986  SELECTING MY IMPUTATION MODEL
987
988  - MICE -> mi impute chained
989
990  - var to be imputed:

```

```

991
992     linear regression for continuous var (regress) ->
993     Hwv8_memory
994
995     logistic for the binary var (logit) ->
996     H_cvd_comorbidity
997
998
999     multinomial logistic for our nominal categorical var (mlogit) ->
1000     H_smoking_3cat H_physicalactivity H_alcohol_status
1001
1002
1003
1004 - auxiliary var:
1005
1006     DV -> Hwv12to14_dementia_event
1007     IV -> H_traj_group4 cesd_0 cesd_1 cesd_2 cesd_3
1008     other covariates -> H_age H_sex H_education H_maritalstatus_4cat
1009                       H_wealthquintiles Hwv8_depressive_symptoms
1010
1011
1012
1013
1014 - imputation numbers (m) -> 20
1015
1016     ELSA data were imputed 20 numbers
1017
1018     White et al. (2010) recommendation: use the rule that m should equal the percentage of
1019     incomplete cases
1020
1021 - rseed (53421) for reproducibility reasons
1022
1023
1024 - (! OPTIONAL) advance impute options -> force
1025
1026     proceed with imputation, even when missing imputed values (e.g. auxiliary have missing data)
1027     are encountered
1028
1029 - impute options -> savetrace (trace1)
1030
1031     specifies Stata to save the means and standard deviations of imputed values from each
1032     iteration to a Stata dataset named "trace1
1033 */
1034
1035 mi set wide
1036
1037 mi register imputed H_smoking_3cat H_physicalactivity H_alcohol_status H_cvd_comorbidity
1038 Hwv8_memory
1039
1040
1041 mi impute chained (logit) H_cvd_comorbidity ///
1042 (mlogit) H_smoking_3cat H_physicalactivity H_alcohol_status ///
1043 (regress) Hwv8_memory = Hwv12to14_dementia_event H_traj_group4 Hwv8_depressive_symptoms ///
1044 H_age H_sex H_education H_maritalstatus_4cat H_wealthquintiles, add(20) rseed(53421) savetrace(
1045 trace1)
1046
1047 * save imputed data
1048
1049
1050 * plot imputations
1051
1052 *it will open a file named trace1
1053 use trace1,clear

```



```

1054 describe
1055
1056
1057 reshape wide *mean *sd, i(iter) j(m)
1058 tsset iter
1059
1060
1061
1062 /*
1063 The trace plot below graphs the predicted means value produced during the first imputation chain.
1064 As before, the expectations is that the values would vary randomly to incorporate variation into
1065 the predicted values for read.
1066 */
1067 tsline H_smoking_3cat_mean1, name(mice1,replace)legend(off) ytitle("Mean of smoking")
1068 tsline H_physicalactivity_mean1, name(mice1,replace)legend(off) ytitle("Mean of physical activity")
1069 tsline H_alcohol_status_mean1, name(mice1,replace)legend(off) ytitle("Mean of alcohol status")
1070 tsline H_cvd_comorbidity_mean1, name(mice1,replace)legend(off) ytitle("Mean of cvd")
1071 tsline Hwv8_memory_mean1, name(mice1,replace)legend(off) ytitle("Mean of memory")
1072
1073
1074 /*
1075
1076 All imputation chains can also be graphed simultaneously to make sure that nothing unexpected
1077 occurred in a single chain.
1078 Every chain is obtained using a different set of initial values and this should be unique.
1079 Each colored line represents a different imputation.
1080 So all 10 imputation chains are overlaid on top of one another.
1081 */
1082
1083
1084 tsline H_alcohol_status_mean*, name(mice1,replace)legend(off) ytitle("Mean of alcohol status")
1085 tsline H_alcohol_status_sd*, name(mice2, replace) legend(off) ytitle("SD of alcohol status")
1086 graph combine mice1 mice2, xcommon cols(1) title(Trace plots of summaries of imputed values)
1087
1088 * repeat for each imputed var
1089
1090
1091
1092
1093
1094
1095 /*
1096 ---- DESCRIPTIVE STATISTICS ----
1097
1098 General characteristics of participants
1099
1100 General characteristics of participnats stratified for study inclusion
1101
1102 General characteristics of participants stratified for dementia occurence
1103
1104 Participant characteristics by depressive symptom trajectory group
1105
1106
1107 1. CHI-SQUARE (chi2) for categorical var (crosstabulation)
1108     Frequency tables -> two-way tables
1109         using the command tabulate, chi2
1110         reporting observations, column percentage (N, %) and p-value of Pearson's r
1111
1112
1113 2. one-way ANOVA for continuous var
1114     check box plot
1115     using the command oneway
1116     reporting mean, sd (summary tables) and p-value of F
1117 */
1118
1119

```

```

1120
1121
1122
1123 * General characteristics of HRS participants
1124
1125
1126 * Demographics
1127 sum H_age
1128 ta H_sex
1129 ta H_education
1130 ta H_maritalstatus_4cat
1131 ta H_wealthquintiles
1132 * Lifestyle and health factors
1133 ta H_smoking_3cat
1134 ta H_physicalactivity
1135 ta H_alcohol_status
1136 ta H_cvd_comorbidity
1137 ta Hwv8_diabetes_reportevr
1138 ta Hwv8_HbA1c
1139 ta Hwv8_crp
1140 ta Hwv12_hdl
1141 ta Hwv8_obesity_waist
1142 ta Hwv8_systolic_bp
1143 ta Hwv8_diastolic_bp
1144 * Depressive symptoms t1-t3 (cont and categ)
1145 sum cesd_0
1146 sum cesd_1
1147 sum cesd_2
1148 sum cesd_3
1149 ta depress_0
1150 ta depress_1
1151 ta depress_2
1152 ta depress_3
1153 * Memory score
1154 sum Hwv8_memory
1155
1156
1157
1158
1159
1160
1161 * Sample characteristics by depressive symptom trajectories
1162 * crosstabs categ var (frequencies and chi2) !report column percentage!
1163 * oneway ANOVA cont var (mean, sd)
1164
1165 * Demographics
1166 oneway H_age H_traj_group5, tabulate
1167 ta H_sex H_traj_group5, chi2 column row
1168 ta H_education H_traj_group5, chi2 column row
1169 ta H_maritalstatus_4cat H_traj_group5, chi2 column row
1170 ta H_wealthquintiles H_traj_group5, chi2 column row
1171 * Lifestyle and health factors
1172 ta H_smoking_3cat H_traj_group5, chi2 column row
1173 ta H_physicalactivity H_traj_group5, chi2 column row
1174 ta H_alcohol_status H_traj_group5, chi2 column row
1175 ta H_cvd_comorbidity H_traj_group5, chi2 column row
1176 ta Hwv8_diabetes_reportevr H_traj_group5, chi2 column row
1177 ta Hwv8_HbA1c H_traj_group5, chi2 column row
1178 ta Hwv8_crp H_traj_group5, chi2 column row
1179 ta Hwv12_hdl H_traj_group5, chi2 column row
1180 ta Hwv8_obesity_waist H_traj_group5, chi2 column row
1181 ta Hwv8_systolic_bp H_traj_group5, chi2 column row
1182 ta Hwv8_diastolic_bp H_traj_group5, chi2 column row
1183 * Memory score
1184 oneway Hwv8_memory H_traj_group5, tabulate
1185
1186
1187

```

```

1188
1189  /*
1190  ---- BINOMIAL LOGISTIC REGRESSION ON COMPLETE DATA ----
1191
1192  Command is:
1193  logistic DV IVs
1194           OR
1195  logit DV IVs, or
1196
1197
1198  --- Building the model using baseline covariates ---
1199
1200  Model 1: unadjusted - single predictor of depressive symptom trajectories C_traj_group5
1201  Model 2: model 1 + sociodemographics: age sex education marital status and wealth
1202  Model 3: model 2 + health behaviours: smoking, alcohol consumption
1203
1204
1205  */
1206
1207
1208
1209  * Unadjusted model - model 1 - single predictor
1210
1211  logistic Hwv11_anydementia_report H_traj_group5
1212
1213  *OR
1214
1215  logit Hwv11_anydementia_report H_traj_group5, or
1216
1217
1218
1219  * define design var by using i.
1220
1221  logistic Hwv11_anydementia_report i.H_traj_group5
1222
1223  *OR
1224
1225  logit Hwv11_anydementia_report i.H_traj_group5, or
1226
1227
1228  * Adjusted models - multivariable logistic regression
1229  * controlling for covariates
1230
1231  * model 2: model 1 + adjust for demographics: age sex education marital status and wealth
1232
1233  logistic Hwv11_anydementia_report i.H_traj_group5 ///
1234  H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles
1235
1236  * model 3: model 2 + adjust for lifestyle factors
1237
1238  logistic Hwv11_anydementia_report i.H_traj_group5 ///
1239  H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
1240  H_smoking_3cat H_alcohol_status H_cvd_comorbidity
1241
1242
1243
1244
1245
1246
1247
1248  /*
1249  ---- SURVIVAL ANALYSIS ----
1250
1251  Tests of proportional-hazards assumption
1252  Kaplan Meier survival curves
1253  Person-time
1254  Cox proportional regression - Hazard ratios - stcox
1255  Postestimation tools for stcox

```

```

1256 Test of Goodness of Fit
1257
1258 *** Cox regression in full data, complete data (listwise deletion of missing data) and imputed data
1259 Cox PH regression in complete data
1260 Cox PH regression model in imputed dataset - mi estimate
1261
1262
1263 */
1264
1265
1266
1267 * check dataset variables of interest only
1268
1269 codebook Hwv12to14_time_of_event_dementia Hwv12to14_dementia_event H_traj_group4 ///
1270 H_age H_sex H_education H_maritalstatus_4cat H_wealthquintiles ///
1271 Hwv12_smoking_3cat Hwv12_physicalactivity Hwv12_alcohol_freq ///
1272 Hwv12_cvd_comorbidity Hwv12_glycemia Hwv12_bp Hwv12_obesity Hwv12_hdl ///
1273 Hwv12_loneliness_quintiles Hwv12_memory_compact
1274
1275
1276
1277 * Declare Data to be Survival Data
1278 * Time to event: Hwv12to14_time_of_event_dementia (months)
1279 * Censoring: Hwv12to14_dementia_event (1=dementia, 0=censored)
1280 * Command is stset TIMETOEVENT, failure(CENSORVARIABLE)
1281
1282
1283 stset Hwv12to14_time_of_event_dementia, failure (Hwv12to14_dementia_event==1) id(RAHHIDPN)
1284
1285
1286
1287 *describe survival data using commnad stsum
1288
1289 stsum
1290
1291 stsum, by(H_traj_group)
1292
1293
1294
1295
1296 * Kaplan Meier Curve estimation
1297
1298 sts list
1299
1300 sts list, by(H_traj_group5)
1301
1302
1303
1304 * Kaplan Meier Curve Plot
1305
1306 * no frills plot
1307
1308 sts graph
1309
1310 * with frills
1311
1312 sts graph, xtitle("Time in Months") ytitle("Survival Prob") ///
1313 title("Kaplan Meier Curve") subtitle("n=4475, # events=233") ///
1314 caption("graph02.png", size(vsmall))
1315
1316
1317 * With Greenwood CI limits
1318
1319 sts graph, gwood legend(off) xtitle("Time in Months") ytitle("Survival Prob") ///
1320 title("Kaplan Meier Curve") subtitle("n=4475, # events=233") caption("graph03.png", size(vsmall))
1321
1322
1323

```

```

1324
1325 * Group Kaplan-Meier Curve Estimation
1326 * Command is sts graph, by(GROUPVAR) OPTION OPTION OPTION Note: Must have sorted by GROUPVAR first
1327
1328 sort H_traj_group5
1329
1330 sts list, by(H_traj_group5)
1331
1332 * graph with frills
1333
1334 sts graph, by(H_traj_group5) xlabel(0(20)120) ylabel(0.80(.05)1) xtitle("Time in Months") ///
1335 ytitle("Survival Prob") title("Kaplan Meier Curve") subtitle("n=4475, # events=233") ///
1336 caption("graph04.png", size(vsmall))
1337
1338
1339
1340
1341 * calculate person-time and incidence rates using command stptime
1342
1343 stptime, title(Person-years)
1344
1345 stptime, title(Person-years) per(1000)
1346
1347
1348 * calculate person-time by category of H_traj_group5
1349
1350 stptime, by(H_traj_group5)
1351
1352 stptime, by(H_traj_group5) per(1000)
1353
1354
1355 * calculate the median of the follow-up
1356
1357 sum Hwv12to14_time_of_event_dementia, detail
1358
1359
1360 /* Log Rank Test of equality of survival distributions
1361 (NULL: equality of survival distributions among H_traj_group5 groups)
1362 We will consider including the predictor if the test has a p-value of 0.2 - 0.25 or less.
1363 If the predictor has a p-value greater than 0.25 in a univariate analysis
1364 it is highly unlikely that it will contribute anything to a model which includes other
1365 predictors.
1366 Command is sts test GROUPVAR
1367 */
1368
1369 sts test H_traj_group5, logrank
1370
1371 sts test H_age, logrank
1372
1373 sts test H_sex, logrank
1374
1375 sts test H_eduaction, logrank
1376
1377 sts test H_maritalstatus_4cat, logrank
1378
1379 sts test H_wealthquintiles, logrank
1380
1381 sts test H_smoking_3cat, logrank
1382
1383 sts test H_cvd_comorbidity, logrank
1384
1385
1386
1387
1388
1389 /* Cox PH regression model
1390

```

```

1391 using the command stcox
1392
1393 --- Building the model ---
1394
1395 Model 1: unadjusted - single predictor of depressive symptom trajectories E_traj_group4
1396 Model 2: model 1 + sociodemographics: age sex education marital status and wealth
1397 Model 3: model 2 + lifestyle and health : smoking, alcohol consumption, cvd
1398
1399 */
1400
1401
1402
1403 * Unadjusted model - model 1 - single predictor
1404
1405 stcox H_traj_group5
1406
1407 * define design var by using i.(low, moderate, high, ref: minimal)
1408
1409 stcox i.H_traj_group5
1410
1411
1412 * Adjusted models - multivariable Cox model
1413 * controlling for covariates
1414
1415 * model 2: model 1 + adjust for demographics: age sex education marital status and wealth
1416
1417 stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles
1418
1419
1420 * model 3: model 2 + adjust for lifestyle and health behaviours
1421
1422 stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
1423 i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity
1424
1425
1426
1427
1428
1429
1430 * Coefficients instead of hazard ratios by specifying the option nohr
1431
1432 stcox i.H_traj_group5, nohr
1433
1434
1435 stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
1436 i.H_smoking_3cat i.H_alcohol_status ///
1437 i.H_cvd_comorbidity, nohr
1438
1439
1440
1441
1442
1443
1444 * Multivariable model development
1445 * Likelihood-ratio tests
1446
1447
1448
1449 *install eststo
1450 findit eststo
1451
1452
1453 * ---- rx controlling for age and sex ----*
1454 quietly: stcox H_age i.H_sex
1455 eststo modelagesex
1456
1457 quietly: stcox H_age i.H_sex i.H_traj_group5
1458 eststo modelagesex_4group

```

```

1459
1460 lrtest modelagesex modelagesex_4group
1461
1462
1463
1464 * ---- rx controlling for sociodemographics ----*
1465 quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles
1466 eststo modelsociodemo
1467
1468 quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles i.
H_traj_group5
1469 eststo modelsociodemo_4group
1470
1471 lrtest modelsociodemo modelsociodemo_4group
1472
1473
1474 * ---- rx controlling for lifestyle and health behaviours-----*
1475 quietly: stcox i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity
1476 eststo modellifestyle
1477
1478 quietly: stcox i.H_smoking_3cat i.H_alcohol_status i.H_traj_group5
1479 eststo modellifestyle_4group
1480
1481 lrtest modellifestyle modellifestyle_4group
1482
1483
1484
1485
1486
1487 * side-by-side comparison of models
1488
1489
1490 quietly: stcox i.H_traj_group5
1491 eststo model1
1492
1493
1494 quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles i.
H_traj_group5
1495 eststo model2
1496
1497
1498
1499 quietly: stcox H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
1500 i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity i.H_traj_group4
1501 eststo model3
1502
1503
1504
1505
1506 * Display Betas and Summary Statistics
1507 estout model1 model2 model3, stats(n chi2 bic, star(chi2)) prehead("Betas")
1508
1509 /* Key Interpretation
1510 Chi2 = Value of LR test comparing the model fit ("full") to intercept only ("reduced")
1511 bic = Schwarz' Bayesian Information Criterion = It is a function of the log-likelihood.
1512 Smaller values indicate a better fit.
1513 */
1514
1515 * Display Hazard Ratios and Model Fit Statistics. Option eform produces hazard ratios
1516 estout model1 model2 model3, eform stats(n chi2 bic, star(chi2)) prehead("Hazard Ratios")
1517
1518
1519
1520
1521 * Postestimation tools for stcox
1522
1523 /* Test of proportional hazards
1524

```

```

1525 If the tests in the table are not significance (p-values over 0.05)
1526 then we can not reject proportionality and we assume
1527 that we do not have a violation of the proportional assumption.
1528 */
1529
1530 estat phtest, detail
1531
1532
1533 /* Proportionality Assumption - method 1
1534 We will check proportionality by including time-dependent covariates in the model
1535 by using the tvc and the texp options in the stcox command.
1536 Time dependent covariates are interactions of the predictors and time.
1537 In this analysis we choose to use the interactions with log(time)
1538 because this is the most common function of time used in time-dependent covariates
1539 but any function of time could be used.
1540 If a time-dependent covariate is significant this indicates
1541 a violation of the proportionality assumption for that specific predictor.
1542 The conclusion is that all of the time-dependent variables are not significant
1543 either collectively or individually thus supporting the assumption of proportional hazard.
1544 */
1545
1546
1547
1548 stcox i.H_traj_group5 H_age i.H_sex i.H_eduaction i.H_maritalstatus_4cat i.H_wealthquintiles ///
1549 i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity, nohr ///
1550 tvc(H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
1551 H_smoking_3cat H_alcohol_status H_cvd_comorbidity) texp(ln(Hwv12to14_time_of_event_dementia))
1552
1553
1554
1555 /* Proportionality Assumption - method 2
1556 by using the Schoenfeld and scaled Schoenfeld residuals
1557 In the stphtest command we test the proportionality of the model as a whole
1558 and by using the detail option we get a test of proportionality for each predictor.
1559 By using the plot option we can also obtain a graph of the scaled Schoenfeld assumption.
1560 If the tests in the table are not significance (p-values over 0.05)
1561 then we can not reject proportionality and we assume
1562 that we do not have a violation of the proportional assumption.
1563 The stphplot command uses log-log plots to test proportionality
1564 and if the lines in these plots are parallel then we have further indication
1565 that the predictors do not violate the proportionality assumption.
1566 */
1567
1568 quietly stcox H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
1569 H_smoking_3cat H_alcohol_status H_cvd_comorbidity, schoenfeld(sch*) scaledsch(sca*)
1570 stphtest, detail
1571 stphtest, plot(H_traj_group5) msym(oh)
1572 stphtest, plot(H_age) msym(oh)
1573 stphtest, plot(H_sex) msym(oh)
1574 stphtest, plot(H_eduaction) msym(oh)
1575 stphtest, plot(H_maritalstatus_4cat) msym(oh)
1576 stphtest, plot(H_wealthquintiles) msym(oh)
1577 stphtest, plot(H_smoking_3cat) msym(oh)
1578 stphtest, plot(H_alcohol_status) msym(oh)
1579 stphtest, plot(H_cvd_comorbidity) msym(oh)
1580
1581
1582
1583
1584
1585 stphplot, by(H_traj_group5) plot1(msym(oh)) plot2(msym(th))
1586 stphplot, by(H_age) plot1(msym(oh)) plot2(msym(th))
1587 stphplot, by(H_sex) plot1(msym(oh)) plot2(msym(th))
1588 stphplot, by(H_eduaction) plot1(msym(oh)) plot2(msym(th))
1589 stphplot, by(H_maritalstatus_4cat) plot1(msym(oh)) plot2(msym(th))
1590 stphplot, by(H_wealthquintiles) plot1(msym(oh)) plot2(msym(th))
1591 stphplot, by(H_smoking_3cat) plot1(msym(oh)) plot2(msym(th))
1592 stphplot, by(H_alcohol_freq) plot1(msym(oh)) plot2(msym(th))

```



```

1593 stphplot, by(H_cvd_comorbidity) plot1(msym(oh)) plot2(msym(th))
1594
1595
1596 * Assessment of PH Assumption: adjust for age and sex
1597 stphplot, by(H_traj_group5) adjust(H_age H_sex) nolntime plot1opts(symbol(none) color(black)
1598 lpattern(dash)) ///
1599 plot2opts(symbol(none) color(navy)) plot3opts(symbol(none) color(green)) plot4opts(symbol(none)
1600 color(red)) ///
1601 title("Assessment of PH Assumption") subtitle(" Predictor is H_tarj_group5") xtitle("months")
1602
1603 * Assessment of PH Assumption: adjust for model 2
1604 stphplot, by(H_traj_group5) adjust(H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles)
1605 ///
1606 nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
1607 plot2opts(symbol(none) color(navy)) plot3opts(symbol(none) color(green)) plot4opts(symbol(none)
1608 color(red)) ///
1609 title("Assessment of PH Assumption") subtitle(" Predictor is H_tarj_group5") xtitle("months")
1610
1611 * Assessment of PH Assumption: adjust for model 3
1612 stphplot, by(H_traj_group4) adjust(H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles
1613 ///
1614 H_smoking_3cat H_alcohol_status H_cvd_comorbidity) ///
1615 nolntime plot1opts(symbol(none) color(black) lpattern(dash)) ///
1616 plot2opts(symbol(none) color(navy)) plot3opts(symbol(none) color(green)) plot4opts(symbol(none)
1617 color(red)) ///
1618 title("Assessment of PH Assumption") subtitle(" Predictor is H_tarj_group5") xtitle("months")
1619
1620
1621
1622
1623 /* Test of overall goodness of fit
1624 Goodness of fit of the final model
1625 2 methods:
1626 - by using the commnad stcoxgof (good fit = non sig p-value)
1627 - by using the Cox-Snell residuals
1628 to create the Nelson-Aalen cumulative hazard function
1629 If the hazard function follows the 45 degree line then we know that it approximately
1630 has an exponential distribution with a hazard rate of one and that the model fits the data
1631 well.
1632 If the model fits the data, the plot of the cumulative hazard versus cs
1633 should approximate a straight line with slope 1.
1634 */
1635
1636 * by using the commnad stcoxgof
1637
1638 * install stcoxgof
1639 findit stcoxgof
1640
1641
1642 stcox H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
1643 H_smoking_3cat H_alcohol_status H_cvd_comorbidity, mgale(mgale)
1644
1645
1646 stcoxgof
1647
1648
1649
1650
1651 * by using the Cox-Snell residuals
1652
1653 quietly stcox H_traj_group5 H_age H_sex H_eduaction H_maritalstatus_4cat H_wealthquintiles ///
```

```

1654 H_smoking_3cat H_alcohol_status H_cvd_comorbidity
1655 predict cs, csnell
1656
1657 * or
1658
1659 quietly stcox H_traj_group5
1660 predict cs, csnell
1661
1662
1663 stset cs, failure(Hwv12to14_dementia_event)
1664 sts generate km = s
1665 generate H = -ln(km)
1666 line H cs cs, sort ytitle("") clstyle(. refline)
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676 /* ----- BINOMIAL LOGISTIC REGRESSION IN IMPUTED DATASET using baseline covariates ----- */
1677
1678 Command is
1679
1680 mi estimate : logit DV IV, or
1681
1682     OR
1683
1684 mi estimate: logistic DV IV
1685
1686 */
1687
1688
1689
1690 * Unadjusted model - model 1 - single predictor
1691
1692 mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report H_traj_group5
1693
1694 *OR
1695
1696 mi estimate, eform("Odds Ratio"): logit Hwv11_anydementia_report H_traj_group5, or
1697
1698
1699
1700 * define design var by using i.
1701
1702 mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report i.H_traj_group5
1703
1704 *OR
1705
1706 mi estimate, eform("Odds Ratio"): logit Hwv11_anydementia_report i.H_traj_group5, or
1707
1708
1709 * Adjusted models - multivariable logistic regression
1710 * controlling for covariates
1711
1712 * model 2: model 1 + adjust for demographics: age sex education marital status and wealth
1713
1714 mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report i.H_traj_group5 ///
1715 H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles
1716
1717 * model 3: model 2 + adjust for lifestyle factors
1718
1719 mi estimate, eform("Odds Ratio"): logistic Hwv11_anydementia_report i.H_traj_group5 ///
1720 H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles ///
1721 H_smoking_3cat H_alcohol_status H_cvd_comorbidity

```

```

1722
1723
1724
1725
1726
1727
1728
1729
1730 * ----- COX PH REGRESSION MODEL IN IMPUTED DATASET ----- *
1731
1732
1733 * Declare Data to be Survival Data by using mi
1734
1735 mi stset Hwv12to14_time_of_event_dementia, failure (Hwv12to14_dementia_event==1) id(RAHHIDPN)
1736
1737
1738 * Run Cox regression analysis in imputed dataset by using "mi estimate:"
1739 * Building the Model: Model 1 (unadjusted), Model 2, Model 3, Model 4
1740 * Interactions
1741
1742
1743 * Model 1 (default coefficients)
1744 mi estimate: stcox H_traj_group5
1745
1746 * Model 1: define design var by using i.(low, moderate, high, ref: minimal)
1747 mi estimate: stcox i.H_traj_group5
1748
1749
1750 * Model 1 ask for hazard ratio by using the option eform("Haz.Ratio")
1751
1752 mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5
1753
1754
1755 * Model 2: sociodemographics
1756
1757 mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
1758 H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles
1759
1760
1761 * Model 3: lifestyle factors
1762
1763 mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
1764 H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles ///
1765 i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity
1766
1767
1768
1769
1770
1771
1772 /*
1773
1774 *** SENSITIVITY ANALYSES ***
1775
1776 1) single assessment of depressive symptoms and dementia risk at t0 and t3
1777 continuous var of CES-D 8 items (0-8)
1778 model 3 was further adjusted for cesd_0 and cesd_3
1779
1780 2) LCGA logit trajectories with dichotomous variable
1781
1782
1783 3) Complete data
1784
1785 */
1786
1787
1788
1789 * 1) Single assessment of cesd_0 and cesd_3 at model 3

```

```

1790
1791
1792 * IMPUTED DATA
1793
1794
1795 * single assessment model 3 adjust for cesd_0 and cesd_3
1796
1797
1798 mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
1799 H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles ///
1800 i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity ///
1801 cesd_0
1802
1803
1804 mi estimate, eform("Haz. Ratio"): stcox i.H_traj_group5 ///
1805 H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles ///
1806 i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity ///
1807 cesd_3
1808
1809
1810
1811
1812
1813
1814
1815
1816 /*
1817 2) Logistic model LCGA
1818
1819 use Hwv8_depressive_symptoms dichotomous variables (0-1)
1820
1821 Logistic (logit) model
1822
1823 use http://www.andrew.cmu.edu/user/bjones/traj/data/cambridge.dta,
1824 clear
1825
1826 traj, var(p1-p23) indep(tt1-tt23) model(logit) order(0 3 3)
1827
1828 trajplot, xtitle(Scaled Age) ytitle(Prevalence)
1829
1830 /* Assigned group and probabilities of group membership */
1831 list _traj_Group - _traj_ProbG3 if _n > 400, ab(12)
1832
1833 */
1834
1835
1836
1837
1838
1839
1840 *rename Ewv2_depressive_symptoms score across the waves
1841
1842 rename Hwv8_depressive_symptoms depress_0
1843 rename Hwv9_depressive_symptoms depress_1
1844 rename Hwv10_depressive_symptoms depress_2
1845 rename Hwv11_depressive_symptoms depress_3
1846
1847
1848 net from http://www.andrew.cmu.edu/user/bjones/traj
1849 net install traj, force
1850 help traj
1851
1852
1853
1854
1855 *** LOGIT MODEL
1856
1857

```

```

1858 * 1 class - logit model - quadratic polynomial (2)
1859 traj, var(depress_*) indep(t_*) model(logit) order(2)
1860 trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci
1861
1862 /* Assigned group and probabilities of group membership */
1863 list _traj_Group - _traj_ProbG1 if _n < 3, ab(12)
1864
1865
1866
1867
1868 * 2 class - logit model - quadratic polynomial (2 2)
1869 traj, var(depress_*) indep(t_*) model(logit) order(2 2)
1870 trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci
1871
1872 /* Assigned group and probabilities of group membership */
1873 list _traj_Group - _traj_ProbG2 if _n < 3, ab(12)
1874
1875
1876
1877
1878
1879 * 3 class - logit model - quadratic polynomial (2 2 2)
1880 traj, var(depress_*) indep(t_*) model(logit) order(2 2 2)
1881 trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci
1882
1883 /* Assigned group and probabilities of group membership */
1884 list _traj_Group - _traj_ProbG3 if _n < 3, ab(12)
1885
1886
1887
1888
1889
1890 * 4 class - logit model - quadratic polynomial (2 2 2 2)
1891 traj, var(depress_*) indep(t_*) model(logit) order(2 2 2 2)
1892 trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci
1893
1894 /* Assigned group and probabilities of group membership */
1895 list _traj_Group - _traj_ProbG4 if _n < 3, ab(12)
1896
1897
1898
1899
1900
1901
1902 * 5 class - logit model - quadratic polynomial (2 2 2 2 2)
1903 traj, var(depress_*) indep(t_*) model(logit) order(2 2 2 2 2)
1904 trajplot, xtitle(Time in Months) ytitle(Depressive symptom caseness) ci
1905
1906 /* Assigned group and probabilities of group membership */
1907 list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)
1908
1909
1910
1911
1912
1913 * The 4-model depressive traj is selected to be tested in different shapes.
1914
1915 * 4 class - logit model - quadratic polynomial (3 3 3 3)
1916 traj, var(depress_*) indep(t_*) model(logit) order(3 3 3 3)
1917 trajplot, xtitle(Time in Months) ylabel(0(.20)1) ytitle(Depressive symptom caseness)
1918
1919 /* Assigned group and probabilities of group membership */
1920 list _traj_Group - _traj_ProbG4 if _n < 3, ab(12)
1921
1922
1923
1924
1925

```

```

1926
1927 program summary_table_procTraj
1928     preserve
1929     *look at the average posterior probability
1930     gen Mp = 0
1931     foreach i of varlist _traj_ProbG* {
1932         replace Mp = `i' if `i' > Mp
1933     }
1934     sort _traj_Group
1935     *and the odds of correct classification
1936     by _traj_Group: gen cesdG = _N
1937     by _traj_Group: egen groupAPP = mean(Mp)
1938     by _traj_Group: gen counter = _n
1939     gen n = groupAPP/(1 - groupAPP)
1940     gen p = cesdG/_N
1941     gen d = p/(1-p)
1942     gen occ = n/d
1943     *Estimated proportion for each group
1944     scalar c = 0
1945     gen TotProb = 0
1946     foreach i of varlist _traj_ProbG* {
1947         scalar c = c + 1
1948         quietly summarize `i'
1949         replace TotProb = r(sum)/ _N if _traj_Group == c
1950     }
1951     gen d_pp = TotProb/(1 - TotProb)
1952     gen occ_pp = n/d_pp
1953     *This displays the group number [_traj_~p],
1954     *the cesd per group (based on the max post prob), [countG]
1955     *the average posterior probability for each group, [groupAPP]
1956     *the odds of correct classification (based on the max post prob group assignment), [occ]
1957     *the odds of correct classification (based on the weighted post. prob), [occ_pp]
1958     *and the observed probability of groups versus the probability [p]
1959     *based on the posterior probabilities [TotProb]
1960     list _traj_Group cesdG groupAPP occ_pp p TotProb counter == 1
1961     restore
1962 end
1963
1964 summary_table_procTraj
1965
1966
1967
1968 ta _traj_Group
1969
1970 recast _traj_Group 1234
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980 * IMPUTED DATA: Logistic regression (Odds Ratio)
1981
1982 * Unadjusted model (model 1)
1983
1984 mlogit (logit) _traj_Group _report
1985
1986 * model 2: model 1 + adjust for demographics: age sex education marital status and wealth
1987
1988 mlogit (logit) _traj_Group _report
1989 H _traj_Group _report
1990
1991 * model 3: model 2 + adjust for lifestyle and health factors
1992
1993 mlogit (logit) _traj_Group _report

```

```

1994   H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles ///
1995   i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity
1996
1997
1998
1999
2000
2001
2002   * IMPUTED DATA: Cox regression (Hazard Ratio)
2003
2004
2005   * Declare Data to be Survival Data by using mi
2006
2007   mi stset Hwv12to14_time_of_event_dementia, failure (Hwv12to14_dementia_event==1) id(RAHHIDPN)
2008
2009
2010
2011   * Unadjusted model (model 1)
2012
2013   mi estimate, eform("Haz. Ratio"): stcox i._traj_Group
2014
2015   * Model 2: sociodemographics
2016
2017   mi estimate, eform("Haz. Ratio"): stcox i._traj_Group ///
2018   H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles
2019
2020
2021   * Model 3: lifestyle and health factors
2022
2023   mi estimate, eform("Haz. Ratio"): stcox i._traj_Group ///
2024   H_age i.H_sex i.H_education i.H_maritalstatus_4cat i.H_wealthquintiles ///
2025   i.H_smoking_3cat i.H_alcohol_status i.H_cvd_comorbidity
2026
2027
2028
2029
2030   * 3) complete data analysis (see above)
2031
2032
2033
2034   * ----- *
2035
2036
2037
2038

```