**Applying a Three-Step BCH correction to GBTM and GBMTM in Stata**

Group-based trajectory and multitrajectory models (GBTM, GBMTM) (D. Nagin, 2005; D. S. Nagin et al., 2018) use repeated observations of one or more variables over time identify groups of individuals following similar developmental trajectories.

When predicting individuals' membership in group-based multi-trajectories or group-based trajectories, or when using individuals' trajectory group membership to predict a distal (later-life) outcome, the *traj* command in Stata offers a one-step approach that simultaneously estimates the group-based trajectories, their risk factors (predictors or precursors), and/or a distal outcome. This "one-step method" produces biased estimates (Bolck et al., 2004), leading scholars to adopt either two-step (Bakk & Kuha, 2018; Di Mari et al., 2023) or three-step (Bakk et al., 2013; Bakk & Kuha, 2021; Vermunt, 2010) corrections when relating external variables to latent class membership. Currently, the three-step correction is not available using the *traj* command in Stata. We apply the three-step correction as follows:

1. **Use the Stata command "*traj*" (Jones & Nagin, 2013) to identify a model of best fit using age[[1]](#footnote-1) and one (GBTM) or more (GBMTM) repeated observations of a variable measured over time.[[2]](#footnote-2)** In our *Socius* paper (Frech et al., 2023), we used only respondent age and our three GBMTM variables to identify a model of best fit. *traj* automatically assigns individuals to their single group of best-fit (the group with the highest posterior class membership probability).[[3]](#footnote-3) This assignment can be found in the created variable "\_traj\_Group". The identification of the trajectories using only age and the time-varying variable(s) and the use of modal assignment to assign individuals to a single trajectory group serves as steps 1 and 2 of the three-step correction (Bakk & Kuha, 2021; Vermunt, 2010).

2. **Identify the posterior probabilities of group membership *for each individual, for each group*.** *traj* calculates probabilities of group membership and creates these variables automatically. Variables are named \_traj\_ProbG1, \_traj\_ProbG2, etc. Each individual has a probability of belonging to *each* group (not only their assigned group); in cases where model fit is good, the assigned group will have a much higher probability than the non-assigned groups. In our *Socius* paper, respondents averaged a greater than .90 probability of class membership to their assigned group (see Table 2).[[4]](#footnote-4)

3. **Use the modified Bolck-Croon-Hagenaars (BCH) approach as described in Vermunt (2010) to calculate the D matrix; use the D inverse matrix to weight the data for analyses. We include an Excel file to aid in these calculations.** To calculate the D matrix,

a. first calculate matrix P(X|W), where X=trajectory group and W=assigned group. Enter these values into the Excel file:

Row 1 of matrix P(X|W) is the probability of trajectory group 1 assignment, given assigned group W:

sum \_traj\_ProbG1 if \_traj\_Group==1

sum \_traj\_ProbG1 if \_traj\_Group==2

sum \_traj\_ProbG1 if \_traj\_Group==3

Row 2 is the probability of trajectory group 2 assignment, given assigned group W:

sum \_traj\_ProbG2 if \_traj\_Group==1

sum \_traj\_ProbG2 if \_traj\_Group==2

sum \_traj\_ProbG2 if \_traj\_Group==3

and so on for each group X.

b. Calculate P(W) for each group X; each group X will have one value for P(W).

use the probabilities produced using "tab \_traj\_Group"

c. Use P(X|W) and P(W) to Calculate P(X) for each group X; each group X will have one value

for P(X). *Excel cells B15:B17*

calculate P(X|W1)\*P(W1)+P(X|W2)\*P(W2)+P(X|W3)\*P(W3) in

d. Calculate D matrix P(W|X) using Bayes’ Rule. *Excel cells F7:H9*

Begin with matrix P(X|W), multiply each value by P(W)/P(X) in Excel.

e. Take the inverse of the D matrix in Excel *Excel cells K7:M9*

4. **Beginning with the wide data file from Step 1, and without using the reshape command, create a long data file in Stata with X observations per individual to add weights. *Each individual is assigned a weight for each group X, regardless of their assigned group*.**

gen weight=.

gen x\_group=.

replace x\_group==1

**Using first row of the inverse D matrix:**

replace weight= [R1, C1 value] if \_traj\_Group==1

replace weight= [R1, C2 value] if \_traj\_Group ==2

replace weight= [R1, C3 value] if \_traj\_Group ==3

and so on across Row 1

append using "filename.dta"

**Using second row of the inverse D matrix:**

replace x\_group=2 if x\_group==.

replace weight= [R2, C1 value] if \_traj\_Group==1 &x\_group==2

replace weight= [R2, C2 value] if \_traj\_Group ==2&x\_group==2

replace weight= [R2, C3 value] if \_traj\_Group ==3&x\_group==2

append using "filename.dta"

and so on across Row 2

Continue for each group X.

The resulting file will have X observations for each individual, requiring the use of vce(cluster idvar) in analyses.

5. **Use the guidance provided in Vermunt (2010) to estimate relationships between group membership and an external (predictor or distal) variable. Weight the regression using your weight variable and iweights, use vce(cluster) to adjust standard errors for individuals' repeated observations.**

for our project, it looked something like this:

mlogit \_traj\_Group i.female i.edref i.raceref i.sixmos\_22\_24 [iweight = weight], base(4) vce(cluster r0000100)

1. any other time metric may be used, see Nagin (2005) [↑](#footnote-ref-1)
2. The process is identical whether using GBTM or GBMTM. traj produces variable \_traj\_Group, reflecting each individual's modal assignment. [↑](#footnote-ref-2)
3. For a discussion of modal versus proportional assignment, see Bakk et al. (2013). [↑](#footnote-ref-3)
4. Probabilities are often lower when the outcome is logistic, see for example (Frech et al., 2022) [↑](#footnote-ref-4)