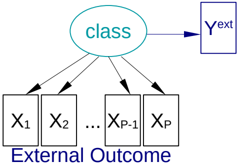
Methods in the literature



- The Naive modal method:
  Classical regression as if there was no measurement error

$$\mathscr{L}(Y_i^{ext}|(\hat{c}_i) = \sum_{i=1}^{N} \log\left(f(Y_i^{ext} \mid \hat{c}_i)\right)$$

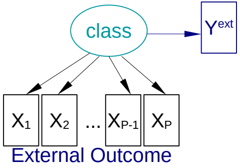
- The Naive proportional method:
  Classical regression weighted by the posterior probability

$$\mathscr{L}(Y_i^{ext}|(\hat{c}_i) = \sum_{i=1}^{N} \sum_{g=1}^{G} \hat{\pi}_{ig} \log\left(f(Y_i^{ext} \mid g)\right)$$

- The Weighting correction method (Bolck 2004, Bakk 2013):
  Classical regression weighted by the misclassification due to the assignment: P(assignment | true class)

$$\mathscr{L}(Y_i^{ext}|(\hat{c}_i) = \sum_{i=1}^{N} \sum_{g=1}^{G} w(\hat{c}_i, \hat{\pi}_{ig}) \log\left(f(Y_i^{ext} \mid g)\right)$$

Methods in the literature (cont'd)

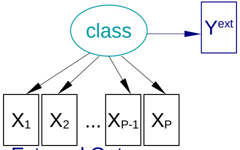


- The conditional regression on the truth (Vermunt 2010, Bakk 2013):

  The regression based on the assignment is rewritten according to our target classes

$$\mathscr{L}(Y_i^{ext}|\hat{c}_i) = \sum_{i=1}^{N} \log\big(f(Y_i^{ext} \mid \hat{c}_i)\big) = \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} f(Y_i^{ext} \mid c_i = g) \times \underbrace{P(c_i = g \mid \hat{c}_i)}_{w_{ig}}\right)$$

Methods in the literature (cont'd)



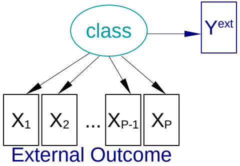
- The conditional regression on the truth (Vermunt 2010, Bakk 2013):
  The regression based on the assignment is rewritten according to our target classes

$$\mathcal{L}(Y_i^{ext}|\hat{c}_i) = \sum_{i=1}^{N} \log\left(f(Y_i^{ext} \mid \hat{c}_i)\right) = \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} f(Y_i^{ext} \mid c_i = g) \times \underbrace{P(c_i = g \mid \hat{c}_i)}_{w_{ig}}\right)$$

- The two-stage method (Xue et Bandeen-Roche 2002, Bakk et Kuha 2018, Proust-Lima 2023):
  We consider the generating model for the total information

$$\mathcal{L}(X_i, Y_i^{ext} \mid \theta_G^X, \theta_G^Y) = \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} P(c_i = g \,;\, \theta_G^X) \times f(X_i \mid c_i = g \,;\, \theta_G^X) \times f(Y_i^{ext} \mid c_i = g \,;\, \theta_G^Y)\right)$$

Methods in the literature (cont'd)



- The conditional regression on the truth (Vermunt 2010, Bakk 2013):
  The regression based on the assignment is rewritten according to our target classes

$$\mathcal{L}(Y_i^{ext}|\hat{c}_i) = \sum_{i=1}^{N} \log\big(f(Y_i^{ext} \mid \hat{c}_i)\big) = \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} f(Y_i^{ext} \mid c_i = g) \times \underbrace{P(c_i = g \mid \hat{c}_i)}_{w_{ig}}\right)$$

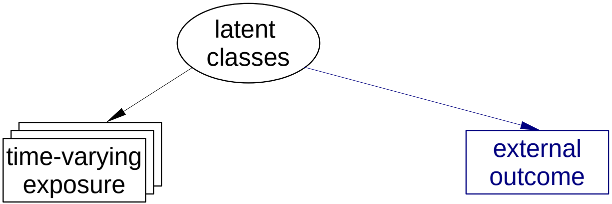
- The two-stage method (Xue et Bandeen-Roche 2002, Bakk et Kuha 2018, Proust-Lima 2023):
  We consider the generating model for the total information but we estimate it in two steps:

$$\mathcal{L}(X_i, Y_i^{ext} \mid \hat{\theta}_G^X, \theta_G^Y) = \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} P(c_i = g\ ;\ \hat{\theta}_G^X) \times f(X_i \mid c_i = g\ ;\ \hat{\theta}_G^X) \times f(Y_i^{ext} \mid c_i = g\ ;\ \theta_G^Y)\right)$$

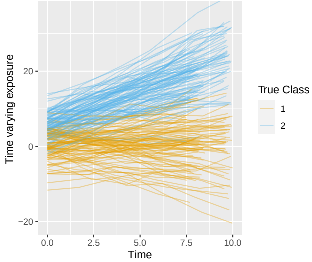
  1. estimate parameters θ_G^X concerning X_i
  2. estimate parameters θ_G^Y concerning Y^{ext} based on those of step 1

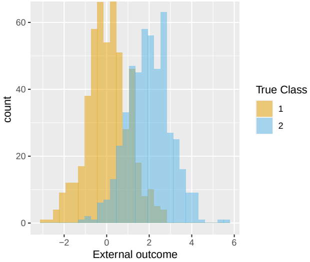
Evaluation of the methods with simulations

Simultaneous generation of the total information (Exposure and External outcome)



e.g., BMI trajectory, Physical Activity in young adulthood

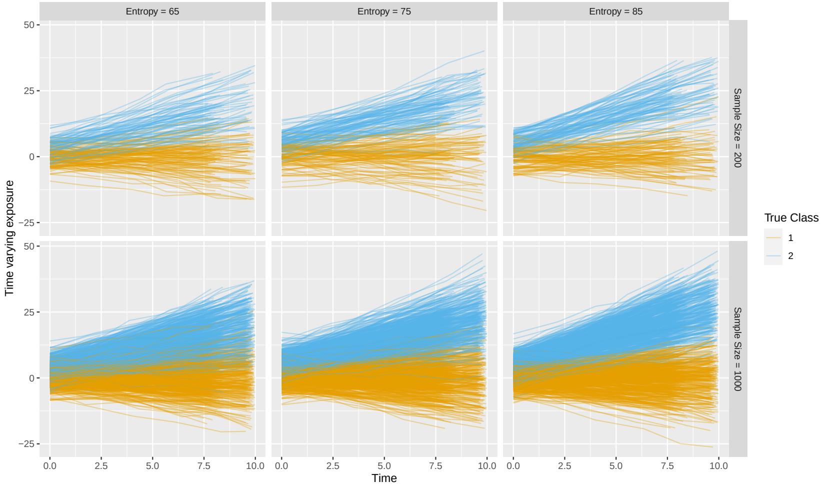




e.g., late-life cognition, BMI

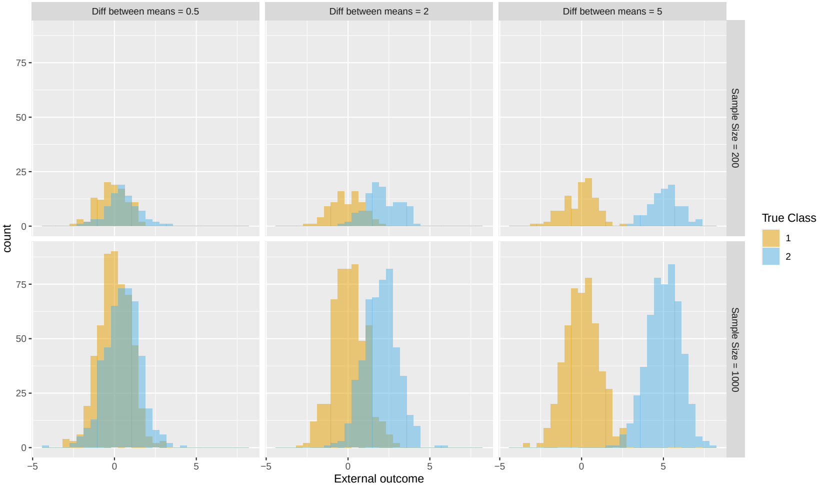
Scenarios of Time-Varying Exposures

- 2 classes (probability 0.5); 2 sample sizes (N=200, 1000); 3 levels of separation (entropy=65%, 75%, 85%)



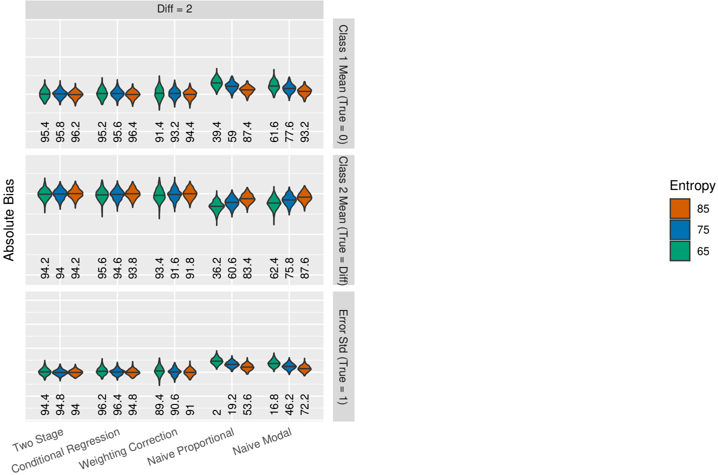
Scenarios of continuous cross-sectional external outcome

- 3 levels of distance between classes (mean difference = 0.5, 2 or 5)



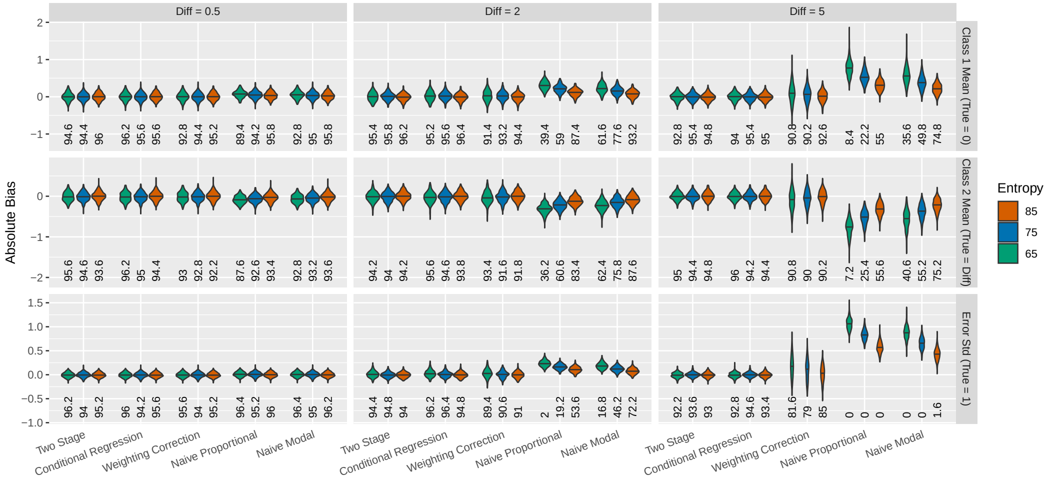
Performances: bias in the external outcome model? N=200

- 3 parameters to examine: mean in each class + variance of the error



Performances: bias in the external outcome model? N=200

- 3 parameters to examine: mean in each class + variance of the error



Performances: bias in the external outcome model? N=1000

- 3 parameters to examine: mean in each class + variance of the error